

UNIVERSITY OF KWAZULU-NATAL

Genetic algorithm based prediction of students' course performance using learning analytics

**By
Rushil Raghavjee
201295456**

**A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy**

**School of Management, IT and Governance
College of Law and Management Studies**

Supervisor: Prof. Prabhakar Rontala Subramaniam

Co-supervisor: Prof. Irene Govender

2024

Declaration

I, Rushil Raghavjee, declare that

- (i) The research reported in this dissertation/thesis, except where otherwise indicated, is my original research.
- (ii) This dissertation/thesis has not been submitted for any degree or examination at any other university.
- (iii) This dissertation/thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
- (iv) This dissertation/thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
 - a) their words have been re-written but the general information attributed to them has been referenced;
 - b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.
- (v) Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.
- (vi) This dissertation/thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.

Signature:



Date: 31 January 2024

Declaration 2 – Derived Publications

Raghavjee, R., Subramaniam, P. & Govender, I. (2021). Learning Analytics in Higher Education. In Perspectives on ICT4D and Socio-Economic Growth Opportunities in Developing Countries (pp. 398-431): IGI Global.

Raghavjee, R., Subramaniam, P. & Govender, I. (2023). Using Design Science Research to Enable Performance Prediction for IS&T Students. In Proceedings of SACLA 2023, Pretoria, South Africa, 19-21 July 2023.

Signed (Student – Rushil Raghavjee)



Signed (Main supervisor – Prof. Rontala Prabhakar Subramaniam)



Signed (Co-supervisor – Prof. Irene Govender)



Acknowledgements

Firstly, to my family. Thank you for your support and patience over the many years that it took to complete this PhD.

I would like to thank my supervisors, Prof. Rontala Prabhakar Subramaniam and Prof. Irene Govender, for their support and guidance during the course of the PhD study.

I am grateful for the assistance of the Insurance Sectors Education and Training Authority (INSETA) for their assistance in covering my fees during the 2022 year. In addition, I am grateful for the University Capacity Development Programme (UCDP) for funding received for lecturer replacement and editing of the thesis.

I acknowledge my colleagues in the discipline of IS&T, especially (in no particular order) Sue Price, Rosemary Quilling, Nurudeen Ajayi, Ntabeni (Joseph) Jere, Surika Civilcharran, Sanjay Ranjeeth, Anyetei (Sonny) Ako-Nai, Patrick Ndayizigamiye and Mudaray (Ashley) Marimuthu, who provided valuable advice and/or support during the course of this PhD process.

I acknowledge the work of those provided in the reference list as well as others that have contributed in the learning analytics, artificial intelligence and associated fields. Without these individuals, my contribution would not exist.

List of abbreviations

Abbreviation	Meaning
AA	Academic Analytics
.arff	Attribute-Relation File Format
.csv	Comma Separated Values
CHE	Council on Higher Education
DSRM	Design Science Research Framework
DS	DataSet
DT	Decision Tree
EDM	Educational Data Mining
HEI	Higher Education Institution
ICT	Information and Communication Technology
II	Institutional Intelligence
IS&T/ISTN	Information Systems and Technology
LA	Learning Analytics
LMS	Learning Management System
Moodle	Modular object-oriented dynamic learning environment
POPIA	Protection of Personal Information Act
RF	Random Forest
SMOTE	Synthetic Minority Oversampling Technique
SMS	Student Management System
UKZN	University of KwaZulu-Natal
WEKA	Waikato Environment for Knowledge Analysis

Abstract

Learning Analytics (LA) can play a key role in understanding students' learning and academic performance. By identifying poorly performing students early, LA can also be used to identify students who are at risk of dropping out of programmes. This enables academic advisors to intervene early and provide help to ensure students stay on track and succeed in their studies.

Hence, LA is becoming a common trend in education particularly in higher education. Previous studies of LA have not dealt with specific courses in information systems and information technology. Therefore, the aim of this study was to develop a model for the application of LA to different courses with the discipline of Information Systems and Technology using various data sources. This study used the design science research approach to help towards solving the problem of understanding students' learning and performance in Higher Education Institutions (HEIs). Multiple data sources were used. The data that was obtained was pre-processed using MS Excel. Thereafter, the WEKA tool was used in the analysis of the data and prediction of performance. Decision tree, Random Forest and genetic-based algorithms were used to develop prediction models for each of the courses in the discipline of Information Systems and Technology at the University of KwaZulu-Natal.

The study also resulted in the development of an integrated dataset for the discipline of Information Systems and Technology in higher education and a process model for the implementation of LA in a specific discipline. The involvedness of the data allows future researchers to continuously improve/evolve the area of LA. This study should, therefore, be of value to LA practitioners wishing to implement LA to courses within other disciplines as well.

Table of contents

Declaration.....	i
Declaration 2 – Derived Publications	ii
Acknowledgements	iii
List of abbreviations	iv
Abstract.....	v
Table of contents	vi
List of figures.....	xiv
List of tables.....	xvi
Chapter 1 – Introduction.....	1
1.1. Introduction.....	1
1.2. Background and Motivation	3
1.3. Research problem	5
1.4. Research questions.....	6
1.5. Research objectives.....	7
1.6. Research methodology	7
1.7. Research contribution	8
1.8. Structure of the thesis.....	8
1.9. Chapter summary	10
Chapter 2 – Literature review	11
2.1. Introduction.....	11
2.2. Learning Analytics.....	13
2.2.1. Learning Analytics and Big Data	15
2.2.2. Comparing Learning Analytics with Educational Data Mining and Academic Analytics	16
2.2.3. Applications of Learning Analytics	18
2.2.3.1. Feedback systems	18
2.2.3.2. Early warning systems.....	20
2.2.3.3. Explanatory Learning Analytics.....	22
2.2.3.4. Learning Analytics for teaching	25
2.2.4. Benefits of Learning Analytics.....	26
2.2.5. Challenges of Learning Analytics	28
2.3. Data sources and feature (factor) identification for success in Learning Analytics	30

2.4. Ethics regarding data use in Learning Analytics	34
2.5. Preparing data for Learning Analytics	35
2.5.1. Handling missing or inconsistent data	36
2.5.2. Data discretization	36
2.5.3. Noise and outlier detection	37
2.5.4. Feature selection.....	37
2.5.5. Normalization and derivation	37
2.5.6. Dealing with imbalanced datasets	38
2.5.7. Data formatting.....	39
2.6. Data analysis and prediction in Learning Analytics	40
2.6.1. Clustering.....	41
2.6.2. Neural networks	42
2.6.3. Naïve Bayes.....	43
2.6.4. Support Vector Machine	44
2.6.5. Random Forest.....	44
2.6.6. Decision Tree algorithms.....	45
2.7. Common tools or applications used for Learning Analytics	47
2.7.1. WEKA.....	47
2.7.2. KNIME.....	48
2.7.3. R.....	48
2.7.4. Python and Jupyter notebook.....	49
2.7.5. RapidMiner	49
2.8. Identification of potential gaps in the literature.....	49
2.9 Chapter Summary	53
Chapter 3 – Research methodology.....	54
3.1. Introduction.....	54
3.2. Research philosophies.....	56
3.3. Information Systems research	57
3.4. Design Science research approach.....	58
3.4.1. Problem identification and motivation.....	61
3.4.2. Defining the objectives for a solution	62
3.4.3. Design and development.....	63
3.4.4. Demonstration	65

3.4.5. Evaluation	65
3.4.6. Communication	65
3.5. Research model	66
3.5.1. An overview of learning analytics models and frameworks.....	66
3.5.1.1. Five stage LA model (Campbell, DeBlois & Oblinger, 2007).....	66
3.5.1.2. Sequence model for learning analytics (Mahzoon et al., 2018)	67
3.5.1.3. Learning analytics cycle	68
3.5.1.4. Learning analytics model by Siemens (2013).....	69
3.5.1.5. Learning analytics models and frameworks that focus on conceptual and physical implementation.....	70
3.5.1.6. Learning analytics model adopted for this study	71
3.5.2. Process model used in this study.....	72
3.6. Data collection	75
3.6.1. Ethical clearance	76
3.6.2. Data used for the study.....	76
3.6.3. Reliability and validity.....	76
3.7. Tools used for data analysis	77
3.8. Chapter summary	78
Chapter 4 – Dataset preparation	79
4.1. Introduction.....	79
4.2. From data acquisition to preparation	80
4.2.1. Description of datasets.....	81
4.2.1.1. Biographical and registration dataset (DS1)	81
4.2.1.2. Dataset consisting of high school marks (DS2).....	86
4.2.1.3. Moodle LMS course data and activity completion datasets (DS3 and DS4).....	86
4.2.2. Data anonymization	88
4.2.3. Cleaning and preparation of datasets	89
4.2.3.1. Handling missing data	90
4.2.3.2. Additional attributes for analysis and prediction purposes	90
4.2.3.3. Data discretization	91
4.2.3.4. Removal of unnecessary attributes and instances.....	92
4.2.3.5. Data integration.....	93
4.3. Chapter summary	94

Chapter 5 – LA model prediction development	96
5.1. Introduction.....	96
5.2. Course dataset description	98
5.2.1. Testing variations of the course datasets	98
5.2.2. Establishment of the training and validation sets	98
5.2.3. Dealing with imbalanced dataset	99
5.2.4. Course descriptions, characteristics and imbalance levels.....	100
5.3. Processing of datasets for prediction.....	104
5.3.1. Feature selection.....	105
5.3.2. Decision Tree algorithm	105
5.3.3. Random Forest algorithm	107
5.3.4. Tools and techniques.....	109
5.3.5. Assessment metrics	111
5.3.5.1. Accuracy	111
5.3.5.2. Kappa statistic	112
5.3.5.3. Receiver operator characteristics (ROC).....	112
5.3.5.4. Precision, recall and F-measure.....	112
5.3.5.5. Assessment metrics used for this study	113
5.4. Results of experiments conducted	114
5.4.1. Experiments for the ISTN100 dataset.....	116
5.4.1.1. Experiment-100-Sampling [None].....	116
5.4.1.2. Experiment-100-Sampling [US].....	117
5.4.1.3. Experiment-100-Sampling [OS].....	118
5.4.1.4. Experiment-100-Sampling [SMOTE].....	119
5.4.1.5. Analysis of experiments conducted.....	119
5.4.2. Experiments for the ISTN101 dataset.....	120
5.4.2.1. Experiment-101-Sampling [None]	120
5.4.2.2. Experiment-101-Sampling [US].....	122
5.4.2.3. Experiment-101-Sampling [OS].....	124
5.4.2.4. Experiment-101-Sampling [SMOTE].....	126
5.4.2.5. Analysis of experiments conducted.....	128
5.4.3. Experiments for the ISTN103 dataset.....	130
5.4.3.1. Experiment-103-Sampling [None].....	130

5.4.3.2. Experiment-103-Sampling [US].....	131
5.4.3.3. Experiment-103-Sampling [OS].....	132
5.4.3.4. Experiment-103-Sampling [SMOTE].....	133
5.4.3.5. Analysis of experiments conducted.....	135
5.4.4. Experiments for the ISTN2IP dataset.....	136
5.4.4.1. Experiment-2IP-Sampling [None].....	136
5.4.4.2. Experiment-2IP-Sampling [US].....	138
5.4.4.3. Experiment-2IP-Sampling [OS].....	138
5.4.4.4. Experiment-2IP-Sampling [SMOTE].....	140
5.4.4.5. Analysis of experiments conducted.....	141
5.4.5. Experiments for the ISTN211 dataset.....	142
5.4.5.1. Experiment-211-Sampling [None].....	142
5.4.5.2. Experiment-211-Sampling [US].....	143
5.4.5.3. Experiment-211-Sampling [OS].....	143
5.4.5.4. Experiment-211-Sampling [SMOTE].....	144
5.4.5.5. Analysis of experiments conducted.....	146
5.4.6. Experiments for the ISTN212 dataset.....	147
5.4.6.1. Experiment-212-Sampling [None].....	147
5.4.6.2. Experiment-212-Sampling [US].....	149
5.4.6.3. Experiment-212-Sampling [OS].....	149
5.4.6.4. Experiment-212-Sampling [SMOTE].....	151
5.4.6.5. Analysis of experiments conducted.....	153
5.4.7. Experiments for the ISTN3SA dataset.....	154
5.4.7.1. Experiment-3SA-Sampling [None].....	154
5.4.7.2. Experiment-3SA-Sampling [US].....	155
5.4.7.3. Experiment-3SA-Sampling [OS].....	155
5.4.7.4. Experiment-3SA-Sampling [SMOTE].....	156
5.4.7.5. Analysis of experiments conducted.....	158
5.4.8. Experiments for the ISTN3AS dataset.....	159
5.4.8.1. Experiment-3AS-Sampling [None].....	159
5.4.8.2. Experiment-3AS-Sampling [US].....	159
5.4.8.3. Experiment-3AS-Sampling [OS] and Experiment-3AS-Sampling [SMOTE].....	159
5.4.8.4. Analysis of experiments conducted.....	160

5.4.9. Experiments for the ISTN3SI dataset	160
5.4.9.1. Experiment-3SI-Sampling [None]	161
5.4.9.2. Experiment-3SI-Sampling [US]	162
5.4.9.3. Experiment-3SI-Sampling [OS]	162
5.4.9.4. Experiment-3SI-Sampling [SMOTE]	162
5.4.9.5. Analysis of experiments conducted	164
5.4.10. Experiments for the ISTN3ND dataset	166
5.4.10.1. Experiment-3ND-Sampling [None]	166
5.4.10.2. Experiment-3ND-Sampling [US]	166
5.4.10.3. Experiment-3ND-Sampling [OS]	167
5.4.10.4. Experiment-3ND-Sampling [SMOTE]	169
5.4.10.5. Analysis of experiments conducted	170
5.5. Chapter summary	172
Chapter 6 – Prediction using genetic algorithms	174
6.1. Introduction	174
6.2. Course datasets to be applied to genetic algorithms	176
6.3. An overview of genetic algorithms	178
6.3.1. Genetic algorithm used for feature selection	178
6.3.2. Optimized forest (OF) algorithm	180
6.3.3. Performance measures and parameters	181
6.4. Results of genetic based experiments conducted	181
6.4.1. Experiments for the ISTN100 dataset	182
6.4.1.1. Experiment-100-FS [Genetic]	182
6.4.1.2. Experiment-100-Algorithm [OF]	182
6.4.2. Experiments for the ISTN2IP dataset	183
6.4.2.1. Experiment-2IP-FS [Genetic]	183
6.4.2.2. Experiment-2IP-Algorithm [OF]	185
6.4.3. Experiments for the ISTN3AS dataset	188
6.4.3.1. Experiment-3AS-FS [Genetic]	188
6.3.3.2. Experiment-3AS-Algorithm [OF]	189
6.5. Chapter summary	191
Chapter 7 – Performance measure comparison with other studies	193
7.1. Introduction	193

7.2. Studies for comparison	194
7.3. Comparison of performances related to first year courses	200
7.4. Comparison of performance related to second year courses	202
7.5. Comparison of performance related to third year courses	203
7.6. Comparison of studies dealing with technology-related courses	204
7.7. Comparison with other LA studies based on technique used	207
7.7.1. Decision tree algorithm.....	207
7.7.2. Random forest algorithm	210
7.7.3. Comparison with other techniques.....	211
7.8. Performance comparison with studies that show training and testing accuracy	212
7.9. Chapter summary	214
Chapter 8 – Conclusion	216
8.1. Introduction.....	216
8.2. Major findings of the literature review chapter	218
8.3. Discussion of the research questions and objectives	218
8.3.1. How can the data from the relevant data sources (SMS, Moodle logs, registers etc.) be integrated?	218
8.3.2. How can the integrated data be organized in preparation for data analysis?	219
8.3.3. How can the data be used for training towards identifying learning patterns?	220
8.3.4. How can the trained data be used to predict student academic performance?	221
8.3.5. How can the resultant information of student academic performance predictions be evaluated?	222
8.4. Discussion: How effective are the predictions in influencing or enabling monitoring of student academic behaviour?.....	223
8.5. Recommendations	227
8.5.1. Effective data acquisition and management	227
8.5.2. Better use of learning management systems	228
8.5.3. Improved communication of analytical findings between student and lecturer	229
8.6. Contributions of the study.....	229
8.6.1. Development of a dataset.....	229
8.6.2. Learning Analytics process artefact	230
8.6.3. Addition of learning analytics research within South Africa and the African continent	231
8.7. Limitations.....	232

8.7.1. Scope of the study	232
8.7.2. Working with Moodle data	232
8.7.3. Availability of data sources for ISTN100 and ISTN3AS courses	233
8.8. Directions for future work	233
8.9. Chapter summary	234
References	236
Appendix A – Experimental results not listed in Chapter 5	249
Appendix B – Experimental results not listed in Chapter 6	267
Appendix C – Letter from professional editor	269
Appendix D – Ethical clearance letter	270

List of figures

Figure 1.1: Thesis structure	1
Figure 2.1: Thesis structure	11
Figure 2.2: Map overview of Chapter 2	12
Figure 2.3: Overview of Section 2.2	15
Figure 2.4: Benefits of LA identified for different beneficiaries	28
Figure 3.1: Thesis structure	54
Figure 3.2: Map for Chapter 3 coverage	55
Figure 3.3: Complementary relationship between design science and behavioural science research areas (Hevner & Chatterjee, 2010)	58
Figure 3.4: DSRM Framework (Peffer et al., 2007)	61
Figure 3.5: LA model suggested by Campbell et al. (2007)	66
Figure 3.6: Sequence model for student representation in LA project proposed by (Mahzoon et al., 2018)	68
Figure 3.7: LA Cycle adapted from Chatti and Muslim (2019)	68
Figure 3.8: Learning analytics model by Siemens (2013)	69
Figure 3.9: Proposed LA model for the study	72
Figure 3.10: Symbolic notation for process model	73
Figure 3.11: Process model developed in this study	74
Figure 4.1: Thesis structure	79
Figure 4.2: Map for Section 4.2 coverage	80
Figure 4.3: Initial dataset hierarchy structure	81
Figure 4.4: Data preparation to integration	95
Figure 5.1: Thesis structure	96
Figure 5.2: Map for Chapter 5 coverage	97
Figure 5.3: Decision tree example and concepts	106
Figure 5.4: Bagging process followed in Random Forest ensemble algorithms	108
Figure 5.5: Experiment notation	115
Figure 5.6: Table format for presenting performance analysis	116
Figure 5.7: Accuracy comparison of best three models	129
Figure 5.8: Performance measure comparison of best three models	130
Figure 5.9: Accuracy comparison for best three models	136
Figure 5.10: Accuracy comparison for best three models	141
Figure 5.11: Performance measure comparison of three best models	142
Figure 5.12: Accuracy comparison for best three models	146
Figure 5.13: Performance measures of best three models	147
Figure 5.14: Accuracy comparison of four best models	153
Figure 5.15: Performance measures of best four models	154
Figure 5.16: Accuracy comparison for best three models	158
Figure 5.17: Assessment measure comparison of three best models	159

Figure 5.18: Accuracy comparison for four best models.....	165
Figure 5.19: Assessment measure comparison for four best models.....	165
Figure 5.20: Accuracy comparison of four models	171
Figure 5.21: Assessment measure comparison for four models.....	171
Figure 6.1: Thesis structure	174
Figure 6.2: Map for Chapter 6 coverage.....	176
Figure 6.3: Crossover example and resultant offspring	180
Figure 6.4: Mutation Example.....	180
Figure 6.5: Accuracy comparison for two OF generated models with best model from Chapter 5 (VAR3-None-RF)	187
Figure 6.6: Comparison of performance measure of two OF generated models against best model from Chapter 5 (VAR3-None-RF).....	188
Figure 7.1: Thesis structure	193
Figure 7.2: Bar chart showing distribution of studies using different algorithms/techniques.....	197
Figure 7.3: Histogram showing distribution of studies based on student numbers	199
Figure 8.1: Thesis structure	216
Figure 8.2: Map of Chapter 8 content.....	217
Figure 8.3: Data integration and preparation	220
Figure 8.4: Data training and accuracy prediction.....	222
Figure 8.5: LA Process model developed from this study	230

List of tables

Table 2.1: Differences between LA, EDM and AA (Adejo & Connolly, 2017b)	18
Table 2.2: Format types and descriptions	30
Table 2.3: Categories of data sources with examples	32
Table 2.4: Studies that used clustering with study objectives	42
Table 2.5: Studies that used neural networks with objectives and accuracy achieved	43
Table 2.6: List of recent studies using Naïve Bayes algorithm with objective and accuracy	43
Table 2.7: Accuracy achieved for prediction studies using SVM	44
Table 2.8: Objectives of studies using Random Forest algorithm with accuracy achieved	45
Table 2.9: Objectives for studies using Decision Tree algorithms	46
Table 2.10: Recent LA/EDM application studies conducted in Africa	50
Table 2.11: LA/EDM Africa-based application studies with problem characteristics	51
Table 3.1: Philosophies, research methods and suggested instruments	56
Table 3.2: Design Science guidelines by Hevner et al. (2004)	59
Table 3.3: List of potential types of artefacts (Vaishnavi et al., 2004)	64
Table 4.1: Attribute description for dataset DS1	82
Table 4.2: Attribute description for dataset DS2	86
Table 4.3: Attributes and descriptions for dataset DS3	87
Table 4.4: Attributes and descriptions for dataset DS4	88
Table 4.5: Addressing missing values in different attributes within the dataset	90
Table 4.6: Ranges for marks	92
Table 4.7: Description of UKZN ISTN dataset	95
Table 5.1: Details of each course in the UKZN ISTN dataset	101
Table 5.2: Characteristics of course datasets	103
Table 5.3: WEKA Filter functions and parameters used for sampling techniques	110
Table 5.4: WEKA <code>WrapperSubsetEval</code> parameters	111
Table 5.5: Assessment metrics to be used for this study with acceptance criteria	114
Table 5.6: Summary analysis for RF generated model – Experiment-100-Sampling [None]-VAR1	116
Table 5.7: Summary analysis for RF generated model – Experiment-100-Sampling [US]-VAR1 .	117
Table 5.8: Summary analysis for RF generated model – Experiment-100-Sampling [OS]-VAR1 .	118
Table 5.9: Summary analysis for RF generated model – Experiment-100-Sampling [SMOTE]- VAR1	119
Table 5.10: Summary analysis for RF generated model – Experiment-101-Sampling [None]-VAR1	121
Table 5.11: Summary analysis for RF and DT generated models – Experiment-101-Sampling [None]-VAR2	121
Table 5.12: Summary Analysis for RF and DT generated models – Experiment101-Sampling [None]-VAR3	122
Table 5.13: Summary analysis for RF and DT generated models – Experiment-101-Sampling [US]- VAR1	123

Table 5.14: Summary analysis for RF and DT generated models – Experiment-101-Sampling [US]-VAR2	123
Table 5.15: Summary analysis for RF and DT generated models – Experiment-101-Sampling [US]-VAR3	124
Table 5.16: Summary analysis for RF and DT generated models – Experiment-101-Sampling [OS]-VAR1	124
Table 5.17: Summary analysis for RF and DT generated models – Experiment-101-Sampling [OS]-VAR2	125
Table 5.18: Summary analysis for RF and DT generated models – Experiment-101-Sampling [OS]-VAR3	126
Table 5.19: Summary analysis for RF and DT generated models – Experiment-101-Sampling [SMOTE]-VAR1	127
Table 5.20: Summary analysis for RF and DT generated models – Experiment-101-Sampling [SMOTE]-VAR2 and Experiment-101-Sampling [SMOTE]-VAR3	128
Table 5.21: Best three models generated from Experiment-101	129
Table 5.22: Summary analysis for RF and DT generated models – Experiment-103-Sampling [None]-VAR2	131
Table 5.23: Summary analysis for RF and DT generated models – Experiment-103-Sampling [None]-VAR3	131
Table 5.24: Summary analysis for RF and DT generated models – Experiment-103-Sampling [OS]-VAR1	132
Table 5.25: Summary analysis for RF and DT generated models – Experiment-103-Sampling [OS]-VAR2	133
Table 5.26: Summary analysis for RF and DT generated models – Experiment-103-Sampling [OS]-VAR3	133
Table 5.27: Summary analysis for RF and DT generated models – Experiment-103-Sampling [SMOTE]-VAR1	134
Table 5.28: Summary analysis for RF and DT generated models – Experiment-103-Sampling [SMOTE]-VAR2	134
Table 5.29: Summary analysis for RF and DT generated models – Experiment-103-Sampling [SMOTE]-VAR3	135
Table 5.30: Best three models generated from Experiment-103	135
Table 5.31: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [None]-VAR1	137
Table 5.32: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [None]-VAR2	137
Table 5.33: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [None]-VAR3	138
Table 5.34: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [OS]-VAR1	139
Table 5.35: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [OS]-VAR2	139
Table 5.36: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [SMOTE]-VAR1	140
Table 5.37: Best three models generated from Experiment-2IP	141

Table 5.38: Summary analysis for RF generated models – Experiment-211-Sampling [None]-VAR1 and Experiment-211-Sampling [None]-VAR2 and Experiment-211-Sampling [None]-VAR3.....	143
Table 5.39: Summary analysis for RF and DT generated models – Experiment-211-Sampling [SMOTE]-VAR1	144
Table 5.40: Summary analysis for RF and DT generated models – Experiment-211-Sampling [SMOTE]-VAR2	145
Table 5.41: Summary analysis for RF and DT generated models – Experiment-211-Sampling [SMOTE]-VAR3	145
Table 5.42: Best three models generated from Experiment-211.....	146
Table 5.43: Summary analysis for RF and DT generated models – Experiment-212-Sampling [None]-VAR1	148
Table 5.44: Summary analysis for RF and DT generated models – Experiment-212-Sampling [None]-VAR2.....	148
Table 5.45: Summary analysis for RF and DT generated models – Experiment-212-Sampling [None]-VAR3.....	149
Table 5.46: Summary analysis for RF and DT generated models – Experiment-212-Sampling [US]-VAR2.....	149
Table 5.47: Summary analysis for RF and DT generated models – Experiment-212-Sampling [OS]-VAR1.....	150
Table 5.48: Summary analysis for RF and DT generated models – Experiment-212-Sampling [OS]-VAR2.....	151
Table 5.49: Summary analysis for RF and DT generated models – Experiment-212-Sampling [SMOTE]-VAR1	152
Table 5.50: Summary analysis for RF and DT generated models – Experiment-212-Sampling [SMOTE]-VAR2	152
Table 5.51: Best four models generated from Experiment-212	153
Table 5.52: Summary analysis for RF generated models – Experiment-3SA-Sampling [None]-VAR1.....	154
Table 5.53: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [OS]-VAR1	155
Table 5.54: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [OS]-VAR2	156
Table 5.55: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [SMOTE]-VAR1	157
Table 5.56: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [SMOTE]-VAR2	157
Table 5.57: Best three models from Experiment-3SA	158
Table 5.58: Summary analysis for RF generated model – Experiment-3SI-Sampling [None]-VAR1	161
Table 5.59: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [None]-VAR2.....	161
Table 5.60: Summary analysis for RF generated models – Experiment-3SI-Sampling [None]-VAR3	162
Table 5.61: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [SMOTE]-VAR1	163

Table 5.62: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [SMOTE]-VAR2	163
Table 5.63: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [SMOTE]-VAR3	164
Table 5.64: Best four models for Experiment-3SI	164
Table 5.65: Summary analysis for RF generated model – Experiment-3ND-Sampling [None]-VAR1	166
Table 5.66: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [US]-VAR2	167
Table 5.67: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [US]-VAR3	167
Table 5.68: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [OS]-VAR1	168
Table 5.69: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [OS]-VAR2	168
Table 5.70: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [SMOTE]-VAR1	169
Table 5.71: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [SMOTE]-VAR2	170
Table 5.72: Performance measures for best four models for Experiment-3ND	170
Table 5.73: Best performing algorithms for each course based on accuracy	172
Table 6.1: Prediction performance for DT algorithm extracted from Table 5.9 for ISTN100 course	177
Table 6.2: Prediction performance for RF algorithm extracted from Table 5.33 for ISTN2IP course	177
Table 6.3: Summary analysis for RF and DT algorithms – Experiment-100-FS [Genetic]	182
Table 6.4: Summary analysis for OF algorithm – Experiment-100-Algorithm [OF]	183
Table 6.5: Summary analysis for RF and DT algorithms – Experiment-2IP-FS [Genetic] – VAR1	184
Table 6.6: Summary analysis for RF and DT algorithms – Experiment-2IP-FS [Genetic] – VAR2	184
Table 6.7: Summary analysis for RF and DT algorithms – Experiment-2IP-FS [Genetic] – VAR3	185
Table 6.8: Summary analysis for OF algorithm – Experiment-2IP-Algorithm [OF] – VAR1	185
Table 6.9: Summary analysis for OF algorithm – Experiment-2IP-Algorithm [OF] – VAR2	186
Table 6.10: Summary analysis for OF algorithm – Experiment-2IP-Algorithm [OF] – VAR3	186
Table 6.11: Summary analysis for RF and DT algorithms – Experiment-3AS-FS [Genetic] – VAR1	189
Table 6.12: Summary analysis for RF and DT algorithms – Experiment-3AS-FS [Genetic] – VAR2	189
Table 6.13: Summary analysis for RF and DT algorithms – Experiment-3AS-FS [Genetic] – VAR3	189
Table 6.14: Summary analysis for OF algorithm – Experiment-3AS-Algorithm [OF] – VAR2	190
Table 6.15: Summary analysis for OF algorithm – Experiment-3AS-Algorithm [OF] – VAR3	190

Table 6.16: Best models from Chapter 5 and Chapter 6 experiments	192
Table 7.1: Comparison strategy.....	194
Table 7.2: List of classification studies based on algorithms used.....	195
Table 7.3: Summary of studies and student numbers	198
Table 7.4: Colour coding for comparison performances.....	199
Table 7.5: Performance measures for studies on 1st year courses	200
Table 7.6: Performance measures for studies on 2nd year courses	202
Table 7.7: Performance measures for studies on 3rd year courses or students.....	203
Table 7.8: Comparison of studies relating to programming-based education	205
Table 7.9: Comparison of studies that used decision tree algorithms	207
Table 7.10: Comparison of studies that used Random Forest algorithms	210
Table 7.11: LA or EDM Studies that have used other techniques.....	211
Table 7.12: Studies that included both training and test/validation accuracy	213
Table 8.1: Identified attributes per course	225
Table 8.2: Number of occurrences of parameters in best prediction models from Table 6.16	226

Chapter 1 – Introduction

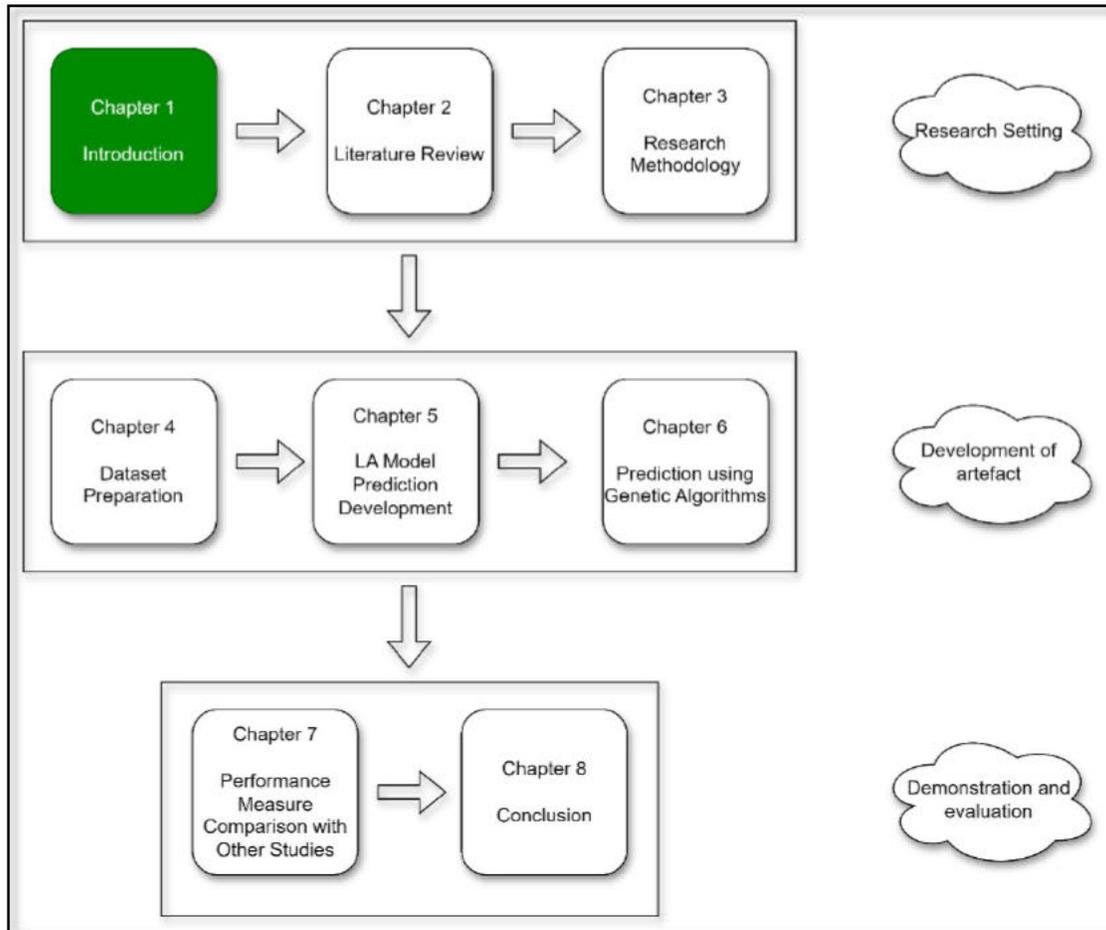


Figure 1.1: Thesis structure

1.1. Introduction

Higher education institutions (HEI) have the responsibility of ensuring that high quality students or graduates are produced that are competent in their knowledge area, thereby allowing them to perform and succeed in the working world (CHE, 2013). To accomplish this, the integration of technology, not only in the learning of subject matter but also as a tool to making learning effective and efficient, is needed. The technology is capable of storing and/or generating large amounts of data related to students and the academic activities that they are involved in (Avella, Kebritchi, Nunn & Kanai, 2016). Some of the data stored include attendance registers, learning management system (LMS) logs, student biographical data and student academic performance data, to name a few. These data are in a variety of formats, both digitized and non-digitized, and structured,

unstructured and semi-structured (Romero & Ventura, 2020). Furthermore, these data have been collected and hoarded for a long period of time, and will continue to be collected, probably at a faster pace due to the evolution of mobile networks, cloud computing and reduced cost of resources (Romero & Ventura, 2020). Despite the advances, both Daniel (2015) and Prinsloo and Kaliisa (2022b) note that this data is not being used optimally by decision makers within the higher education environment.

With the evolution of technology and advanced data analysis techniques (Bollier & Firestone, 2010), many institutions around the world are now attempting to integrate these data sources and use them intelligently in order to make decisions regarding improving teaching and learning. Thus, the concept of Learning Analytics (LA) was introduced, described by Siemens et al. (2011, p. 4) as “The measurement, collection, analysis and reporting of data about learners and their contexts, for the purpose of understanding and optimizing learning and the environments in which it occurs. Learning analytics are largely concerned with improving learner success.”

The advent of LA has resulted in researchers studying a number of areas surrounding LA. Firstly, before LA can be implemented, the ethical and privacy implications of LA implementation need to be debated. This is an area of importance identified by Slade and Prinsloo (2013), Olivier (2020) and Prinsloo and Kaliisa (2022a), amongst others. Students need to be made aware that digital data about them and their academic activities are being recorded and for what purposes, while data anonymization ensures that student data is kept private (Prinsloo & Kaliisa, 2022a).

Besides the ethical and privacy issues, the actual acquisition, cleaning and preparation of the data must be properly understood, with Romero and Ventura (2020) stating that this is an often neglected area of LA research and can make up more than half the time required to solve the LA problem. Strategies such as dealing with missing data and inconsistent data, data discretization, outlier detection and feature selection, amongst others, must be dealt with in order for the analytics to produce reliable information (Romero & Ventura, 2020).

The most commonly researched aspect of LA is the application of learning algorithms to datasets. Commonly used learning algorithms include Decision Trees, Neural Networks, Naïve Bayes, and

Clustering (Aggarwal, 2020). The output of the learning algorithms, either in the form of prediction models or cluster groups, can be used to predict student performance, identify strengths and weaknesses in student learning as well as determine or predict students that may need assistance with improving their academic performance.

There are other aspects of LA such as data visualization, which focuses on presentation of data so that it may be interpreted by relevant decision makers (Romero & Ventura, 2020), and prescriptive analytics, where the objective is to strategize and be pro-active based on predictions made (Bonnin & Boyer, 2017). However, with the lack of research related to LA on the African continent, it is first necessary to focus on the initial aspects of LA. Thus, the main objective of this research is the development of an artefact to guide the process of LA. This will involve the collection of data resulting in the development of a dataset based on the discipline of Information Systems and Technology (IS&T) at the University of KwaZulu-Natal (UKZN). Learning algorithms will then be applied to this dataset, resulting in the creation of models that can be used to predict whether a student will pass or fail a course, based on features such as individual demographics and registration data, past academic performance and course interaction data. Finally, the entire LA process will be documented and presented in the form of a process model that can be used to guide future researchers in conducting LA for their particular discipline.

1.2. Background and Motivation

In the current environment, higher education is regarded as a critical component for increasing the possibility of employment for individuals as well as to improve economic performance of a country (Chiramba & Ndofirepi, 2023; Pinheiro, Wangenge-Ouma, Balbachevsky & Cai, 2015). As a result, student performance is probably one of the most important aspects of higher education institutions and is seen as a key objective of HEIs in South Africa (CHE, 2013). Two factors that measure student performance is observing assessments as well as yearly graduation rates (Shahiri & Husain, 2015). In the case of South Africa, Ngqulu (2018) mentions that HEIs in the country are struggling with dealing with poor student success rates and throughput.

Within the South African higher educational context, there is a continuous increase in enrollment while government cannot support this increase, resulting in inadequate funding (Badat, 2016;

Chiramba & Ndofirepi, 2023; Mlambo, Mlambo & Adetiba, 2021). The resultant lack of funding has been the main motivator for the *#FeesMustFall* movement where students are pushing for government to ensure that higher education is free, livable study accommodations are provided, and there is improved access to the technology and infrastructure required (Raghavjee, Subramaniam & Govender, 2021). The consequences of this push have been constant protest actions resulting in numerous interruptions to teaching and learning.

In addition to the protests, in 2020, the world experienced the COVID-19 pandemic resulting in all South African HEIs moving from a face-to-face teaching and learning model to an online learning model. To assist with online learning, students were assisted by universities and government by being supplied with data, laptops and tablets while some students were allowed to return to campus to access technological infrastructure when restrictions were lifted (Raghavjee et al., 2021). From a teaching perspective, the online learning model was one that most academic and administrative staff were not familiar with (Hedding, Greve, Breetzke, Nel & Jansen van Vuuren, 2020).

To maintain the quality of education in this new online environment, it has become necessary to find a way to monitor and understand student progress in terms of online learning and their academic performance. In order to accomplish this, universities are striving to make better use of the data that is being continuously collected and stored, including past and current academic performance data, student interactions in academic activities as well as student biographical background. All this data is stored and available via different university systems.

Since the movement of paper-based records to digital records, the majority of universities make use of a Student Management System (SMS) that stores all student applications (past and present) and includes their biographical data (names, addresses and contact details, nationalities, high school education, qualifications etc.), course registrations, degree registrations and academic results (tests and exams).

Many universities also make use of a Learning Management System (LMS) which allows lecturers to upload course material and activities that students may interact with on or off campus. Common

features of an LMS include calendars with important due dates and course relevant events, personalized dashboards, file management, activity tracking and online assessments, amongst others (Foreman, 2017). Administrative tasks that are found in an LMS include secure login authentication, mass enrollment, continuous updating of security measures, high interoperability with external applications and plug-ins, detailed logs about student interaction and mark management (Foreman, 2017).

With the continued advancements and reliability of technology and the data that it generates, the South African HEIs have been slow to take advantage of Big Data analytics when compared to other areas of industry (Prinsloo & Kaliisa, 2022b). Furthermore, another inherent weakness is that these data sources are separate and isolated (that is, not related or linked) from each other. Most universities store the data for a set number of years, mainly for record and/or legal purposes. In addition, the data is often incomplete with a number of errors and inconsistencies, and stored in a variety of different formats. With the advances in computer processing and networks, the current trend is to integrate these data sources and, using data analysis techniques, find interesting trends and/or patterns and predictions within student academic activities (Daniel, 2015).

Thus, with this important aspect of student monitoring not being fully utilized, it is necessary to better study the use of LA to predict student performance.

1.3. Research problem

The societal issues affecting South African universities currently (protest actions, reduced government support) have affected the quality of education, yielding low pass rates and increased drop-out rates, thus resulting in poor graduation throughput (Marongwe, Mbodila & Kariyana, 2020; Moodley & Singh, 2015). One strategy to address this problem is to continuously monitor and regulate student academic activities and the progress that they are making in their coursework. However, due to the ever-increasing student numbers and limited resources and assistance provided by government, this is becoming logistically difficult for academic staff to achieve.

Additionally, students and lecturers are continuously using technology as part of the teaching and learning process. All registration data is stored electronically and course content is distributed to

students via learning management systems. Communication is not only conducted through face-to-face meetings but via e-mail, online discussion forums and social media applications. While academic staff and institutions collect and store data regarding their students, the data sources are used in isolation (that is, without consideration of other possible data sources) or just for recording purposes.

Although the LA concept was introduced in 2010 (Prinsloo & Kaliisa, 2022b), it is still seen by many as a fairly new area of research (Axelsen, Redmond, Heinrich & Henderson, 2020; Viberg, Hatakka, Bälter & Mavroudi, 2018). The majority of studies are focused on data acquisition and the application of a variety of algorithms to determine accuracy. While other aspects of LA studies such as intervention strategies and its impact on academic performance are important, this is still an area of LA to become proficient at, that is just as important.

Thus, in summary, the research problem is stated as follows:

In order to predict and understand student academic performance in higher education institutions, the use of technology in learning analytics has become increasingly important due to limited resources and an ever-increasing number of students.

1.4. Research questions

An avenue that has only recently been looked at in the last five to ten years is the analysis of various data sources to regulate and monitor student progress. To effectively use the various data sources, the combined use of these data sources is proposed in this study by means of learning analytics. This research, using data mining techniques, will involve analyzing and predicting students' academic progress based on the various data sources available, including the university LMS interactions, SMS data with registrations, demographics and previous academic performance.

The main research question is thus as follows:

How can the different sources of Information Systems and Technology (IS&T) student data be used effectively in the learning analytics process?

From the main research question, the following sub questions have been established:

1. How can the data from the relevant data sources (SMS, Moodle logs etc.) be integrated?
2. How can the integrated data be organized in preparation for data analysis?
3. How can the data be used for identifying learning patterns (training)?
4. How can the trained data be used to predict student academic performance?
5. How can the resultant information of student academic performance predictions be evaluated?

1.5. Research objectives

Based on the research questions listed in Section 1.4, the main research objective is as follows:

To develop and implement a Learning Analytic model in order to effectively use IS&T student data sources for predicting academic performance.

The following are the sub-objectives of this research study:

1. To integrate the relevant university data sources in preparation for classification.
2. To extract, clean and classify the integrated data.
3. To train the data in order to determine patterns and useful information for student performance prediction.
4. To determine the effectiveness of the training techniques by evaluating their accuracy in terms of how they predict student performance.
5. To evaluate the results generated by the artefact against other similar artefacts.

1.6. Research methodology

In order to accomplish the objectives (and thus answer the research questions) listed above, a Design Science Research Methodology will be adopted. Design Science research is a commonly used research methodology in Information Systems that results in the development of an artefact. In the case of this study, a process model will be proposed, designed, demonstrated and evaluated. This process model will guide researchers through the learning analytics process from data acquisition to the application of prediction algorithms.

1.7. Research contribution

Tertiary institutions are continuously collecting large amounts of digital data, but are mostly not using it effectively, if at all. Thus, this study aims to contribute to the growing literature within the LA field. In addition, as the focus is on a university within South Africa, this study will address how other universities may better take advantage of the digital data stored in order to improve student learning outcomes, both from the perspective of meeting course objectives as well as improving undergraduate throughput. This is an area where higher education appears to be lagging when compared to other areas of industry (Joksimović, Kovanović & Dawson, 2019; Ştefan, 2017). A more detailed discussion on the contribution of the research is provided in section 8.4.

1.8. Structure of the thesis

Figure 1.1 in Section 1.1 depicts the overall structure of this thesis with the current chapter (Chapter 1) highlighted. As shown, the study is divided into three segments. Segment one (1) consists of the first three (3) chapters, where the research setting is established by outlining the problem, objectives as well as establish the position of the research in the current literature and finally, describe the methodology used to conduct the research. Segment two (2) covers chapters four (4) to six (6) which is the application of the research in order to address the research problem. The final segment, chapter seven (7) and chapter eight (8), demonstrates and evaluates the application conducted and summarizes the entire study.

This chapter (one) introduces the study and its justification in the current higher education context. It covers the research in terms of its background and motivation, the research questions and subsequent objectives, the methodology used, the research contribution and finally, the thesis structure.

Chapter 2 is a literature review chapter that focuses on the domain of LA, its position in the overall area of Big Data analytics, as well as the concepts, terminology and past studies related to LA. The chapter then focuses further on the area of LA by expanding upon the different processes involved and tools available to conduct LA research.

The research methodology of the study is covered in Chapter 3. This chapter covers the methodology of the design science research used in this study. The chapter also covers the design of the research artefact and how it is used to meet the objectives of the study. The chapter concludes by covering data collection and ethical clearance and the tools that were intended to be used for the study.

Chapter 4 covers the first two questions of this research project. From an LA project perspective, this is the steps of the initial data acquisition and preparation phases. As will be discussed, this often-overlooked area of LA will cover the requirements for acquisition, cleaning and integration of the data sources in preparation for data analysis and prediction.

Chapter 5 addresses research questions 3 and 4 regarding how data is analyzed with the objective of predicting student academic performance using established learning algorithms; these being the Decision Tree algorithm and the Random Forest algorithm. Using feature selection, these algorithms were applied to the different course datasets with the objective of generating models that can accurately predict student performance. The experiments conducted for the two algorithms are also presented in this chapter.

Chapter 6 covers the use of genetic algorithms to find or improve prediction models for any courses identified in Chapter 5 where the models were not acceptable or if better models could be found. In this case, experiments are presented where genetic algorithms were used, either as part of feature selection or incorporated into the classification process. In the case of the latter, an optimized forest algorithm was used, where genetic algorithms are used for determining the best Decision Tree within the forest.

Chapter 7 answers research question 5 by presenting a comparison between the performances of the prediction models obtained in Chapter 5 and Chapter 6 with the performance measures of other LA or Educational Data Mining (EDM) studies from the literature. Various comparisons are made, where the performances of the experiments from chapter 5 and 6 are compared to similar experiments conducted in other studies. This includes comparison with studies involving other 1st, 2nd and 3rd year courses, other computer based courses amongst other comparison types.

Finally, Chapter 8 provides a discussion of how this LA study and the prediction models generated can contribute positively towards student academic performance at the UKZN institution. Further conclusions of the study are provided, as well as suggestions for improvement of LA implementation and areas of study for future work related to this study.

1.9. Chapter summary

The chapter introduces the importance of LA in the current higher education climate. The advances in networks, storage and other technologies necessitate the need for higher education institutions to make better use of the different types of academic data that is being collected, in order to better understand student academic performance.

Section 1.2 provides a background of the current scenario facing higher education in South Africa and motivates on how LA can assist in dealing with the lack of resources and improving throughput at South African universities.

The research questions and objectives are outlined (Sections 1.4 and 1.5). A brief description of the methodology used is provided in Section 1.6. The research contribution is provided in Section 1.7. With LA being fairly new to the African continent, there is a need for research related to the application of learning algorithms to African-based datasets, as well as making these datasets available for future studies. Finally, the outline of the thesis is provided.

Before looking at the application of LA in a South African university context, it is important to understand and appreciate what has already been covered in LA. Thus, the next chapter covers an overview of LA and previous related research.

Chapter 2 – Literature review

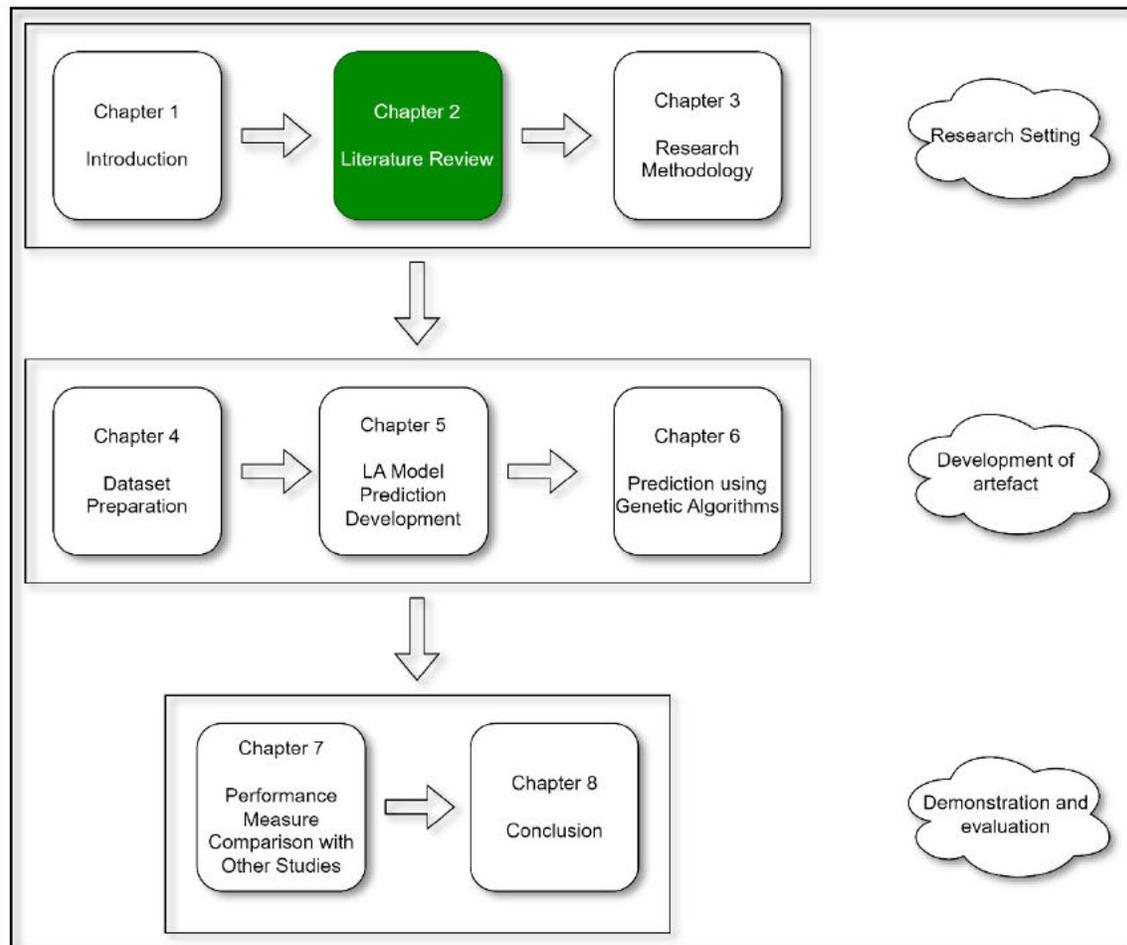


Figure 2.1: Thesis structure

2.1. Introduction

This chapter is a literature review covering learning analytics (LA) and its related concepts. A map depicting the content of this literature review is shown in Figure 2.2.

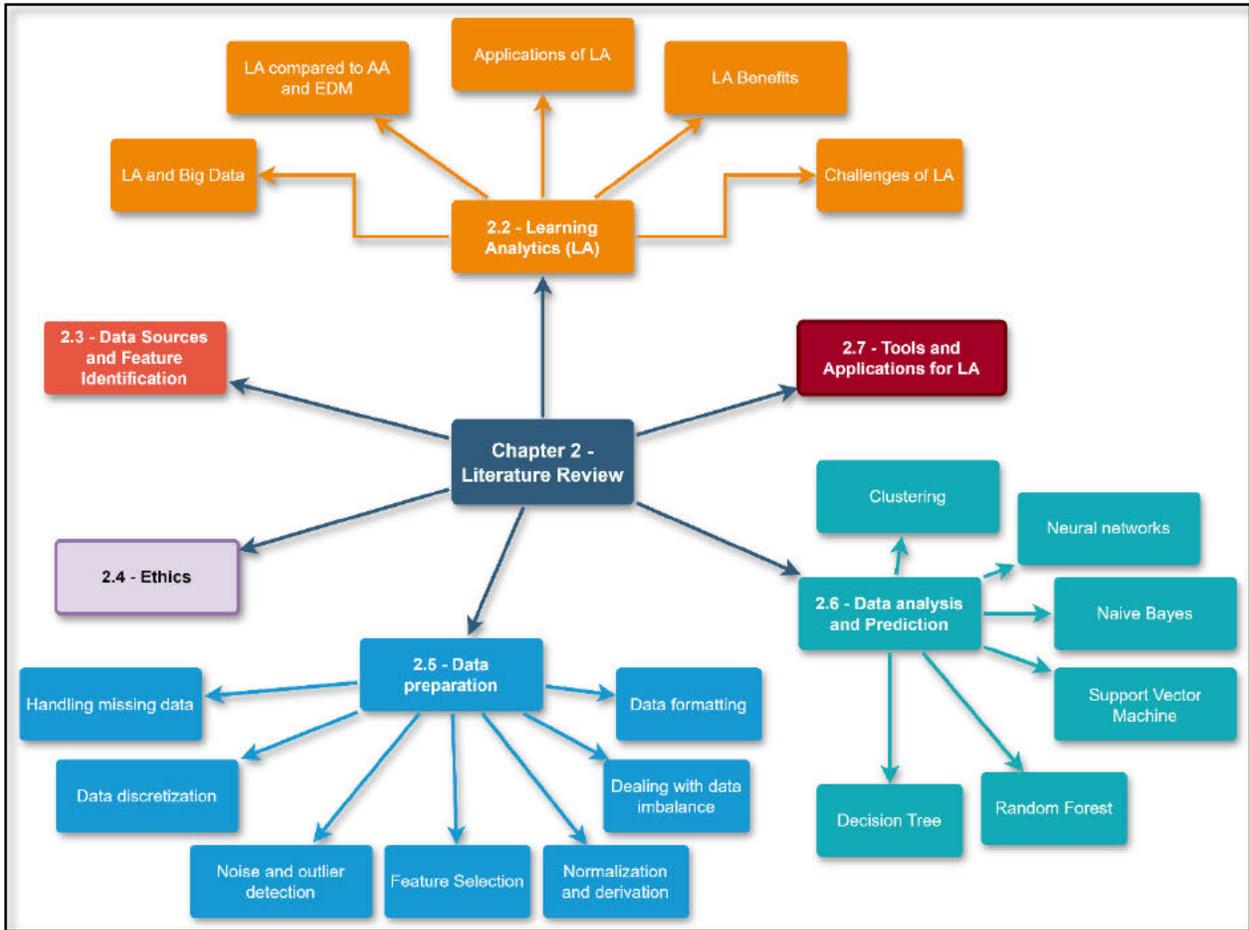


Figure 2.2: Map overview of Chapter 2

Section 2.2 provides an overview of LA, including its relationship with Big Data and how it relates to Academic Analytics (AA) and Electronic Data Mining (EDM). This section also covers how LA has been applied in higher education institutions along with the benefits and challenges of LA. Section 2.3 covers one of the initial aspects of an LA undertaking, that being identification of relevant data sources while section 2.4 covers the ethical and privacy aspects of dealing with digital data sources, specifically from a South African perspective. Section 2.5 provides a literature overview of aspects related to preparing the acquired data for analysis and prediction. Section 2.6 describes the most common techniques and algorithms used in LA while section 2.7 provides an overview of commonly identified tools used in LA.

This literature review chapter, along with Chapter 1 and Chapter 3 forms part of the chapters that establish the setting of the research (See Figure 2.1). From this literature review, the knowledge

gap can be identified and so too, the justification for the research. The literature review also contributes towards the development of the research artefact by identifying what techniques and tools work well when implementing LA.

2.2. Learning Analytics

In the current academic environment, there is a great reliance on the use of technology for teaching and learning, resulting in the generation of an enormous amount of data (Avella et al., 2016). The challenge is now to use this collection of data from a variety of sources in order to better understand student academic performance and plan a way forward to improve the quality of teaching and/or inform students about where they can improve their learning processes (Avella et al., 2016), thus resulting in the concept of LA. As stated in Section 1.1, LA is commonly defined as the application of data for measurement, collection, analysis and reporting purposes, with the objective being to better understand and improve the quality of the learning environment (Siemens et al., 2011).

Learning Analytics is said to be a bricolage field (Dawson, Joksimovic, Poquet & Siemens, 2019), meaning that the area of study emerged from multiple combinations of different disciplines. According to Haggag, Latif and Helal (2018), some of these disciplines include data mining, psychology, statistics, information science, machine learning as well as sociology. Ferguson (2012) states that LA has a strong connection to web analytics, business intelligence, educational data mining and decision support systems.

Boyer and Bonnin (2016) describe four avenues of LA that can be followed by HEIs, these being descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics.

According to Boyer and Bonnin (2016), descriptive analytics answer the question of “what happened?”. This question is answered using general computational and statistical techniques and visualizations that are applied to teaching and learning related data. Examples of visualization in descriptive analytics include pie charts, bar charts or line graphs. This kind of LA is mainly used by students and teachers to evaluate their academic performance and teaching pedagogy, respectively (Boyer & Bonnin, 2016).

The objective of diagnostic analytics is to answer the question “Why did it happen?”. In this case, the data is analyzed to better understand the root cause of the teaching and learning problem, an example being the identification of events that contribute to a student failing (Xin & Singh, 2021). This form of LA requires the use of data discovery or pattern identification as well as statistical correlation (Boyer & Bonnin, 2016).

In the case of predictive analytics, the question of “What will happen?” is dealt with. The objective of this form of LA is to provide insight and anticipate what may happen given a specific situation, based on past and present data. This form of LA may inform a student of whether or not they are achieving their learning objective(s) based on their actions (interactions in class or online). A teacher may use predictive analytics to determine students that are at-risk of failing, thus allowing for interventions to prevent failure (Boyer & Bonnin, 2016).

The final avenue of prescriptive analytics answers the question “How can we make it happen?”. In this case, data or digital content is analyzed in order to determine an efficient and effective strategy to achieve the required goal(s) (Boyer & Bonnin, 2016). Similar to predictive analytics, prescriptive analytics allows relevant stakeholders to discover trends of student drop-out and allows them to be pro-active in their academic activities. Prescriptive analytics also allows course lecturers and assistants to develop personalized learning plans for students.

The following subsections provides a description of the concept of LA, including how it relates to Big Data (Section 2.2.1) as well as a comparison to two overlapping areas of study, namely Academic Analytics (AA) and Educational Data Mining (EDM) (Section 2.2.2). An overview of previous studies and how LA was applied is covered in Section 2.2.3. Finally, the benefits and challenges of LA are covered in section 2.2.4 and 2.2.5 respectively. An overview of this section is shown below (Figure 2.3):

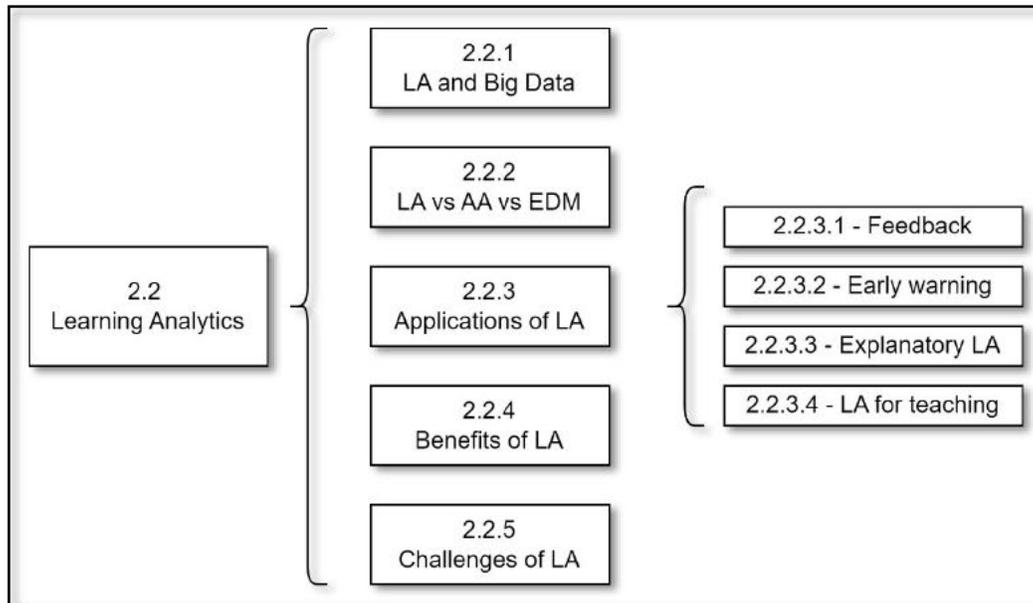


Figure 2.3: Overview of Section 2.2

2.2.1. Learning Analytics and Big Data

Over the past few years, there has been an increased demand in all industries to better capture and use data. This demand has been supported by an increased ability to store large amounts of data as well as improved computational power (Hershkovitz & Alexandron, 2020). The ability to store and intelligently process large amounts of data with the use of networked, online digital systems has been termed Big Data (Hershkovitz & Alexandron, 2020).

Most definitions of Big Data include the characteristics of the three V's, these being Volume, Velocity and Variety (Oussous, Benjelloun, Lahcen & Belfkih, 2018). The Volume characteristic covers the amount of data available to an organization (Kaisler, Armour, Espinosa & Money, 2013). Velocity relates to the speed of data creation, streaming and aggregation (Kaisler et al., 2013). Finally, data Variety refers to the richness of the data representation, i.e. the different data item types and formats contained within the data (Kaisler et al., 2013). Additional characteristics of Big Data that have been noted by Oussous et al. (2018) include Vision (purpose of the Big Data application), Verification (processed data that conforms to specifications), Validation (purpose is fulfilled), Value (organizations find the data useful for different scenarios), Complexity (data is complex and requires organization and proper relationships) and Immutability (Stored and well managed Big Data can be permanent).

In terms of Big Data and higher education, Fischer et al. (2020) identified two major trends appearing in most institutions. The first major trend is the continuous digitization and storing of student profile information and academic records in the form of a student information system. This data is usually heterogeneous and multimodal (Fischer et al., 2020) and thus fulfils the Volume and Variety characteristics of Big Data. The second major trend is the continuous capturing of student activities through LMSs (also referred to as clickstream or log data). This trend occurs on a daily basis (Fischer et al., 2020) and thus fulfils the Velocity requirement.

According to Joksimović et al. (2019), most industries such as health, finance, insurance and aviation have seen the importance of the analysis of large amounts of data. Higher education, however, has been very slow in realizing its importance in implementing effective systems that analyze learning related data for better decision making (Joksimović et al., 2019; Ştefan, 2017). Ştefan (2017) states that Big Data in higher education is still in the incipient stage, meaning universities are still researching and experimenting in this area. According to Dawkins (2018), LA research falls under the area of Big Data, where the assumption is that larger-sized datasets have the capability of providing better intelligence, thus allowing for the potential of better decision making.

However, it should be noted that LA is not the only field that falls under Big Data, specifically in the educational field. Two other commonly researched fields that overlap with LA are that of Educational Data Mining and Academic Analytics. These three fields of research are distinguished in Section 2.2.2.

2.2.2. Comparing Learning Analytics with Educational Data Mining and Academic Analytics

Big Data research in higher education covers a variety of areas including student academic performance, student teaching evaluation, university throughput, and resource usage evaluation, amongst other areas. From the review of the literature, Big Data at HEIs are covered through three fields of research, these being LA, Academic Analytics (AA) and Educational Data Mining (EDM).

The objective of LA is to provide information using analytical tools, statistical and predictive methods and models that will allow decision makers, usually teachers and/or students, to take action to improve teaching and learning (Avella et al., 2016). According to Adejo and Connolly (2017b), LA is focused on improving teaching and learning and providing useful information to learners, teachers and course administrators.

While LA is focused on the improvement of teaching and learning, AA is aimed at improving and/or making better decisions at an educational management or operational level (Boyer & Bonnin, 2016). Academic Analytics is defined as the application of business intelligence and associated tools with the goal of improving decision making and academic performance for educational institutions (Avella et al., 2016). The use of AA is mainly aimed at allowing for better decision making for university administrators, governments and funding agencies (Adejo & Connolly, 2017b; Viberg et al., 2018).

From a higher education perspective, the application of Big Data has been performed at different levels within HEIs (Mendez, Ochoa, Chiluzia & De Wever, 2014). According to Siemens and Long (2011), these five levels are course, departmental, institutional, regional and national/international. Learning analytics is said to fall under the first two levels (course and departmental) while the remaining three levels (institutional, regional and national/international) are said to fall under the concept of AA (Mendez et al., 2014).

Electronic Data Mining is a concept closely related to both LA and AA (Avella et al., 2016) but is more technically oriented (Baek & Doleck, 2023) and focuses specifically on the development of methods or techniques that are able to find patterns, discoveries and/or make predictions within educational data (Avella et al., 2016). Electronic data mining is defined as being concerned with developing methods that explore educational data with the objective of better understanding students and the environment in which they learn (Adejo & Connolly, 2017b).

There are several similarities between LA, EDM and AA, such as all focusing on a data intensive approach to education. The differences between LA, EDM and AA, adapted from Adejo and Connolly (2017b), are listed in Table 2.1.

Table 2.1: Differences between LA, EDM and AA (Adejo & Connolly, 2017b)

	EDM	AA	LA
Target audience	Teachers and administrators	Educational institutions	Teachers, students and educational institutions
Implementation benefit	Automated adaptation and method of interpretation	Automatic iterative processes	Support human interventions and interpretation of data
Application focus	Software and student modelling	Administrative concerns	Systematic intervention
Research focus	Techniques and methodology		Application of analysis, techniques and methodology
Data application	Makes use of data mining techniques	Makes use of statistical techniques and predictive modelling	Makes use of quantitative methods, data mining techniques, visualization tools

2.2.3. Applications of Learning Analytics

Learning analytics projects have been implemented at different institutions with various objectives. According to Hooda and Rana (2020), this is dependent on the target stakeholders that the implementation is aimed at as well as the framework that was being used. This section describes some of the main applications of LA from the literature as well as recent studies of LA applications, what was done and the outcome of the research conducted.

2.2.3.1. Feedback systems

According to both Evans (2013) and Wise (2019), feedback is a critical component of improving student learning outcomes and should ideally allow the student to evaluate their own progress and regulate or adjust their learning styles based on the feedback provided. Further to this, it is also important to better understand the feedback being provided so as to improve this (feedback) process as well (Evans, 2013). This area of application is rarely researched as there is a greater focus of research on using LA to predict student performance and improve graduation throughput (Gašević, Jovanović, Pardo & Dawson, 2017). It is particularly important in the higher educational context due to the increase in class sizes as well as the increase in the socio-economic diversity of the student population (Iraq, Fudge, Faulkner, Pardo & Kovanović, 2020).

In research by Gašević et al. (2017), the aim was to use LA to analyze student learning strategies using log data, as well as study how these strategies influence learning outcomes. In the context of this study, a learning strategy contains the thoughts, behaviours, beliefs, or emotions that allow for the accumulation of new knowledge and skills (Gašević et al., 2017). A questionnaire, log data and assessment results were used as data sources. Statistical techniques were used to detect patterns in learning behaviour. Four specific learning sequences were found and the researchers were able to map these sequences to different learning styles, these being deep and surface level learning. The research concluded by stating that students that followed a deep learning style were found to have higher exam scores overall than students with a surface learning style; and the research had the potential to provide students with information on what type of studying they fall under and the potential consequences of their studying style.

In a study by Saucerman, Ruis and Shaffer (2017), the focus was on automating the detection of reflection on action, which is defined as the ability to remember past problem solutions and in turn apply them to solve a current problem being experienced (Saucerman et al., 2017). The study focused on high school and college students from the USA doing a Land Science internship course. Using statistical techniques and automated algorithms, the authors were able to identify comments made by students and relate these comments to whether students were reflecting on their actions.

A study by Kovanović et al. (2018) also used LA to better understand student reflections. However, in this case, the aim was to improve a student's self-regulated learning ability. Using reflection recordings from student groups as well as individually, the data was quantitatively analyzed with the results indicating that their system was able to classify student reflections. This allowed for better understanding of student reflections, which opens the possibility of automated feedback to students on improving their learning as well as academic performances in the courses.

Student feedback and the use of personalized feedback messages was the focus of a study by Iraj et al. (2020). The intention was to understand student feedback and its effect on academic performance, how feedback relates to student demographics and how students react to feedback. Using statistical methods, the authors found that there was a relationship between students that reacted to test feedback and improved student performance in subsequent tests. The results of the

study also indicated that females and non-English speaking students were more likely to interact with feedback compared to males and English-speaking students respectively (Iraj et al., 2020). Similarly, LA based feedback was seen as a benefit in a study conducted by Ustun, Zhang, Karaođlan-Yilmaz and Yilmaz (2023). The study found that in a 10-week course with 62 students, LA based interventions that included visual and written feedback was found to improve academic performance (Ustun et al., 2023).

2.2.3.2. Early warning systems

Early warning systems involve the identification of risk factors that are used to predict whether or not a student will pass a course or end up failing or even dropping out of a course (Jokhan, Sharma & Singh, 2019). Macfadyen and Dawson (2010) highlight that the foundations of early warning systems have already been established for HEIs, i.e. the integration of ICT into teaching and learning, improved detail and availability of LMS tracking data, the emergence of analytics in the educational sector and increased attention of the social nature of education. Many studies relating to early warning systems produce some form of prediction accuracy, indicating how well the system is able to predict a student's performance based on the factors provided. Prediction is a form of supervised learning that occurs via a training set containing known data. The application of learning algorithms to the training set leads to trends and patterns emerging, eventually resulting in a model that is able to predict a target value based on a set of supplied input values or predictor variables (Wise, 2019).

A study by Jayaprakash, Moody, Lauría, Regan and Baron (2014) was based on the development of an LA system to detect at-risk students. Using four (4) data mining techniques, the authors attempted to develop a model to predict student drop-out rate. A further objective was to determine how well this prediction model can be used in other institutions. The model was developed and applied to four related institutions and the authors determined that while drop-out prediction could be well determined (drop-out prediction accuracy ranging between 70% to 82% for three out of the four institutions), greater care must be taken for institutions with greater demographic diversity.

The purpose of a study by Oloruntoba and Akinode (2017) used the support vector machine algorithm to create a model for academic performance prediction using student high school results as well as initial 1st year results. The developed prediction model achieved a 98% accuracy and

could be used for identifying potentially excellent students for scholarships and to assist in enrollments and identifying students that are unlikely to graduate (Oloruntoba & Akinode, 2017).

A study was conducted at the University of Cape Town to predict academic performance of first year Computer Science students (Nudelman, Moodley & Berman, 2019). Using Bayesian networks and Decision Trees, the authors predicted student academic performance based on matric results, type of high school attended and university registration details. The techniques used in the study were able to produce 91% accuracy in detecting students that would be unsuccessful in passing first year Computer Science courses.

Dorodchi et al. (2018) conducted a study to determine the impact of how student self-reflection can determine whether the student is at-risk of failing or not. The study was conducted with a group of ninety-one (91) Computer Science students. The data sources acquired include demographics, performance scores, student self-reflections and self-assessments. Using sequence analysis and linguistics analysis, the authors stated that results were promising (no prediction accuracy was provided) and that further research was warranted.

Another early warning system was developed as a Moodle plugin by Jokhan et al. (2019) and applied to an online Information Literacy course consisting of 1523 students from the University of South Pacific, Fiji. For this study, the focus was on the student's interaction with the Moodle LMS for the course and how it affected their final mark. Predictor variables that were identified and used to make predictions were activity completion rate, login frequency and coursework interaction.

Hasan et al. (2020) used LA and data mining to predict student academic performance for 772 students registered for e-commerce technology courses. The data sources included student academic information (such as GPA, number of attempts for the course, at-risk status and previous coursework performance), student activity on the Moodle LMS and interactions with coursework video files (number of times video was played, paused, liked, and rewound). Several classification techniques (such as Random Forest, Decision Trees, Naïve Bayes and Neural Networks) were applied and the results were compared. The accuracy for each of the algorithms ranged from 82%

to 87%, with the Decision Tree models producing the best accuracy. The study also used feature selection algorithms to identify the most effective predictor variables. When feature selection algorithms were applied, the Random Forest classifier model produced the best accuracy (88%). The authors also preferred the results from their rule inducer algorithm as it provided information that was easy for non-expert users to understand. The use of an information dashboard was identified as important and formed part of future research.

A study by Renò et al. (2022) focused on the development of a prediction model to assist in predicting whether or not students will pass or fail automated online assessments. The dataset is a benchmark dataset containing course information, student information and LMS log data. The dataset consisted of 32 593 students with 97 attributes. The Random Forest algorithm was chosen and the resultant prediction model achieved a 95% accuracy. Future research includes consistently capturing student data from the LMS and the development of a feedback system to assist struggling students.

A study by Silva, Rupasingha and Kumara (2022) involved the implementation of four machine learning algorithms to a dataset of 200 graduates from a Sri Lankan university. The data source was in the form of a questionnaire requesting student demographics, study habits, hobbies and academic activities. The Random Forest algorithm produced the best accuracy (97.5%) followed by the multilayer perceptron algorithm with the Naïve Bayes algorithm producing the lowest accuracy (70%). The researchers concluded that the study would benefit the institution in identifying weak students that require assistance to pass and intended, as future research, to increase the number of instances and consider more attributes to include in the questionnaire.

2.2.3.3. Explanatory Learning Analytics

Another common application of LA is to better understand different factors that play a role or affect the academic performance of a student, referred to by Wise (2019) as Explanatory LA. These factors could relate to a student's background, academic performance, financial status or interaction with course content amongst others.

A recent study by Preetha (2021) attempted to understand the impact of student health on academic performance. A health-based questionnaire focusing on student disabilities, sport participation, health and nutrition activities was distributed to 113 students at an Indian university. The current mark percentage for the semester was also requested from each respondent. With each question representing an attribute, a K-means algorithm was implemented and the students were placed into one of two clusters based on whether or not the student will pass or fail. A genetic search algorithm was also used to identify the best set of attributes for best predicting academic performance. The authors concluded that the clustering algorithm performed well for prediction but more instances were required in the future.

Asif, Merceron, Ali and Haider (2017) conducted a study to predict student academic performance at a university in Pakistan. Using Decision Trees and clustering techniques, the objective of the study was to determine the role of high school marks and previous academic marks in predicting a student's final mark of a four-year Information Technology degree. Using the Decision Tree classifier, the authors were able to achieve accuracy prediction in the range of 55% to 83%. The study also looked at better understanding a student's progression through the degree and used clustering to divide students into high performing and low performing groups of students.

A study by Daud et al. (2017) attempted to determine the effect on family expenditure and personal characteristics (such as marital and employment statuses) on student academic performance. Five classification techniques were applied to a dataset of 776 Pakistani students from years 2004 to 2011 and the results were studied. An accuracy of 86% was reported using the support vector machine classifier. A further conclusion from the study was that family expenditure and the identified personal information attributes had a great impact on student academic performance.

Mwalumbwe and Mtebe (2017) conducted a study to determine the relationship between student academic performance and student interaction with the LMS for two courses at a Tanzanian university. An application was developed to keep track of LMS log data in terms of number of logins, time logged in and types of interactions (such as forum posts, downloads, exercises performed). Using regression analysis, the authors found that forum interactions, peer interactions and exercises were significant activities that had an impact on student academic performance

(Mwalumbwe & Mtebe, 2017). Koç (2017) also conducted a study looking at user interactions via LMS and its impact on academic achievement. In the case of this study, structured equation modelling was used to determine if the relationship exists between student interaction using LMS and academic achievement. The author specifically focused on discussion forums, online lecture attendance and assignment submissions. The results indicated that there was a positive impact on academic performance when students were more involved via discussion forums and lecture attendance. Learning Management System interaction was also the focus of a study by Avcı and Ergün (2019) where multivariate analysis of variance (MANOVA) was applied to the log data of 65 undergraduate students. For this study, it was determined that student online interactions positively influence student engagement and academic performance.

The study by Al luhaybi, Tucker and Yousefi (2018) looked at prediction of academic performance based on admission data, module related data and 1st year final grades for a 2nd year Computer Science course at Brunel University (London). The predictive model would classify students as either high, medium or low risk of failure. Using clustering and classification techniques, the study identified that the model generated using the Naïve Bayes classification technique provided a better accuracy than when developing a model using the Decision Tree algorithm. Another outcome of the study was that the student qualification upon registration and 1st year marks had a significant impact on academic performance.

The objective of the research by Fincham et al. (2019) was to better understand the concept of engagement in Massive Open Online Courses (MOOCs) by developing a framework. The study used two data sources for each of the three (3) courses being analyzed, that being log data from the LMS as well as tone and linguistic analysis from student discussion forums and other posts. To better understand the role that the data source attributes played in academic performance, exploratory factor analysis was applied first, with the objective of better understanding learning in non-formal educational settings. Secondly, structural equation modelling was used to understand the relationship between the data source attributes and learning outcomes. The authors concluded that their developed framework could allow researchers to view content engagement from the perspective of the individual (the effect of their background and motivation) as well as the course (how course design influences student engagement).

Kumar and Singh (2017) used a collection of academic and personal data from post-graduate students to predict their academic performance for the year. Some of these attributes include past academic performance, parent's qualification and current financial status of the student's family. To predict the performance of the student, a number of classifiers were applied to the dataset, including Decision Trees, Naïve Bayes, Random Forest and Bayes network. The authors identified that the Random Forest classifier performed the best and stated that family and academic attributes could be important factors in student academic performance prediction.

2.2.3.4. Learning Analytics for teaching

Learning analytics can also be used by teachers to identify strengths and weaknesses of content available in courses (Nguyen, Gardner & Sheridan, 2017; Nguyen, Tuunanen, Gardner & Sheridan, 2021). Once identified, teachers can strategize ways to improve understanding or restructure the course to ensure better engagement between the students and the available content.

A recent study by Nguyen et al. (2021) addressing the improvement of teaching and learning focused on the development of an information system that provided lecturers with information related to students' interactions with live lecture recordings. The objective of the study was to design an LA system using LA design principles that were proposed by the author. These design principles were found to be useful guidelines for the development of an LA information system that supports teaching and learning. Three design principles were used and evaluated, that being the principle of actionable information (reporting of information about learners and their learning), the principle of information timeliness, and the principle of availability and interoperability (Nguyen et al., 2021).

A study by Balbay and Kilis (2018) collected student log data as well as student questionnaires with the aim of enhancing the quality and efficiency of a course relating to improving English language skills. Using descriptive statistics and deductive content analysis, the authors identify the study as an important starting point that provides useful, practical information for course improvement.

From an assessment perspective, Amigud, Arnedo-Moreno, Daradoumis and Guerrero-Roldan (2017) integrated LA into the assessment process. By collecting student assignments and applying them to machine learning techniques and language and writing analysis, the application was able to associate each student with their writing and language styles. The authors stated that this application had the potential to improve academic integrity, especially in an online environment.

A study by Mendez et al. (2014) focused on the use of past student grades to better understand and assist in curriculum development. Using student past grades from approximately 2500 Computer Science students, data analysis was conducted from past academic performance. Questionnaires were also used to better understand course difficulty of the various Computer Science courses. The conclusions of the study indicated that these data sources could potentially inform relevant stakeholders regarding the quality of courses which could assist in curriculum redevelopment.

2.2.4. Benefits of Learning Analytics

Several benefits of LA have been identified in the literature. These benefits have arisen as a result of a number of small or large LA applications in higher education. The most commonly identified beneficiaries of LA are the learners (or students), teachers (or lecturers/educators), administrators and the research community (Romero & Ventura, 2020).

The use of statistical and prediction techniques allow academics to define and uncover student problems and needs with regard to the academic courses that they are undertaking. This allows for not only detecting whether students are at risk of failing (Gašević et al., 2017; Nguyen et al., 2021; Patwa, Seetharaman, Sreekumar & Phani, 2018; Sclater, Peasgood & Mullan, 2016) but to also better understand how students perceive the learning process (Muljana & Placencia, 2018).

The understanding of how students learn and the ability to predict their performance allows for the development of early intervention and improvement strategies (Chatti & Muslim, 2019; Gašević et al., 2017; Mahrooian, Daniel & Butson, 2017). This will allow for better student guidance towards passing their courses, thus improving student retention and/or graduation throughput (Patwa et al., 2018; Sclater et al., 2016).

A further benefit of LA implementation linked to performance prediction is that of personalized learning, which was identified as a key benefit of LA (Bonnin & Boyer, 2017; Chatti & Muslim, 2019; Ellaway, Pusic, Galbraith & Cameron, 2014; Muljana & Placencia, 2018; Patwa et al., 2018). This can be accomplished through better understanding of student demographics and academic related behaviours (Mahroeian et al., 2017). Muljana and Placencia (2018) further state that the realization of a “one size does not fit all” approach is important, i.e. student knowledge acquisition and assessments should be tailored individually to meet the diverse learning abilities of each individual student.

Teachers or lecturers are also seen as potential beneficiaries of LA. This is accomplished in terms of curriculum development and analysis of teaching performance. The predictions resulting from LA provides the opportunity for lecturers to reconsider or revise their learning activities with the objective of improving the quality of the course activities and resources such as notes, slides, videos and tutorials (Bonnin & Boyer, 2017; Leitner, Khalil & Ebner, 2017; Nguyen et al., 2021). In addition, LA could improve the use and allocation of resources based on prediction of student enrollment and requirements, to maximize graduation throughput (Avella et al., 2016; Mahroeian et al., 2017).

Both Avella et al. (2016) and Leitner et al. (2017) stated that LA will also benefit the research community that engages in furthering knowledge on using Big Data in all forms of education. Even after more than a decade of research, LA is seen as a relatively new area of research, and further implementation and publication of LA research will allow for the identification of gaps between academia and industry so that research problems can be further studied and addressed (Avella et al., 2016). This benefit was also identified by Gašević et al. (2017) as well as Chatti and Muslim (2019). Figure 2.4 summarizes the LA benefits with the intended beneficiaries:

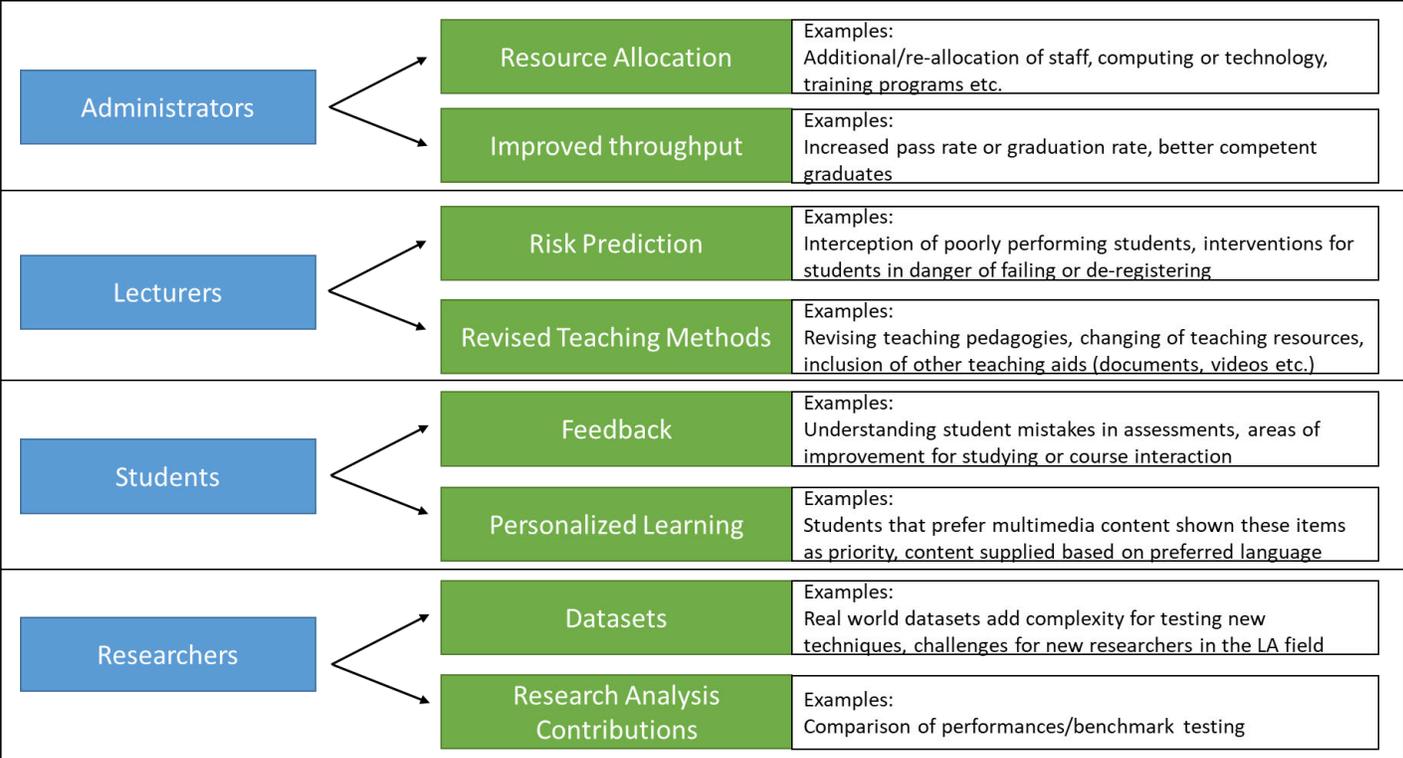


Figure 2.4: Benefits of LA identified for different beneficiaries

2.2.5. Challenges of Learning Analytics

Despite more than a decade of research revolving around LA, institutions are still struggling with taking advantage of learner and organizational data to address educational challenges (Axelsen et al., 2020).

The challenge of technical infrastructure refers to the cost and complexity of data integration as well as hardware and software acquisition (Mahroeian et al., 2017; Ngqulu, 2018). Data collection is usually the initial challenge, requiring the consideration of several aspects including data availability, categories of data to consider (demographic data, learner interaction data, financial records etc.) and data ownership (Avella et al., 2016). As technology evolves, the ability to capture data also evolves, such as the use of readily available datasets offered by LMSs, mobile data, biometric data and mood data (Avella et al., 2016). From its early beginnings, the fact that data stored by institutions are isolated (separated) within different departments is an obstacle to effectively analyze the large amount of student data being captured daily (Ngqulu, 2018).

Another obstacle to overcome is the evolving challenge of the ethical, legal and risk considerations as identified by Alzahrani et al. (2023) as well as Guzmán-Valenzuela, Gómez-González, Rojas-Murphy Tagle and Lorca-Vyhmeister (2021). This is due to the way that data and applications are stored on cloud services (Avella et al., 2016). Adejo and Connolly (2017b) and Guzmán-Valenzuela et al. (2021) both state that the question of data ownership for all collected data must be determined and students should be made aware that data is being collected and used for academic analysis (Patwa et al., 2018). Related to ethical considerations is the issue of data privacy concerns. Administrators must ensure that guidelines are in place to monitor the access and usage of student data (Adejo & Connolly, 2017b). Leitner, Ebner and Ebner (2019) state that while regulations are in place regarding data and ethics, codes of practice regarding LA implementation are lacking and must be addressed.

In terms of implementation challenges, there is a lack of standardization and frameworks for data modelling in LA. This includes dealing with structured and unstructured data, data types and working with missing data (Adejo & Connolly, 2017b; Daniel, 2015; Mahroeian et al., 2017). Even after more than a decade of research, Gašević et al. (2017) as well as Nguyen et al. (2021) found that there is a lack of LA implementation guidelines. There is also a lack of LA adoption policies (Leitner et al., 2019).

With regard to stakeholder challenges, the majority of teachers, students and administrators are unfamiliar with LA and its related concepts. This makes collaboration between these stakeholders and LA developers difficult (Guzmán-Valenzuela et al., 2021; Leitner et al., 2019; West, Heath & Huijser, 2016). Ngqulu (2018) states that it is imperative that for a LA initiative to be successful, staff training of LA practices should be mandatory and staff recruitment must take Big Data and analytics competency into consideration. This can be difficult due to high teaching staff workload resulting in lack of time or motivation for further training (Kaliisa, Kluge & Mørch, 2022). Leitner et al. (2019) also identified lack of leadership as an obstacle for LA implementation.

Related to the challenges above, Ngqulu (2018) identified funding as a challenge for LA implementation. This is especially the case for the acquisition of specialized technology, software, and personnel. In the case of poorer and developing countries, funding may be prioritized for other

initiatives such as infrastructure maintenance and other resources (Guzmán-Valenzuela et al., 2021).

Institutions must weigh the benefits and challenges of learning analytics before attempting to begin an LA initiative. When the benefits are understood and the challenges addressed, data sources must be identified within the institution to leverage the benefits of LA.

2.3. Data sources and feature (factor) identification for success in Learning Analytics

Since LA follows a data driven approach, it is important to identify educational data sources that can be used. This is usually the first step in the LA process followed by preprocessing, feature selection and finally, analysis and/or prediction (Gao, Xie & Tao, 2016). According to Chatti, Dyckhoff, Schroeder and Thiis (2012), LA data sources fall into two (2) categories, i.e. centralized educational systems and distributed learning environments. Examples of centralized education systems include LMSs such as Moodle and Blackboard. These are large, multipurpose applications that accumulate large amounts of data, including student activities and interaction (log) data (for example viewing and updating learning material, interacting with test questions and viewing summaries or reports). On the other hand, distributed learning environments involve the acquisition of data sources beyond the LMS. These data sources can be formal or informal as well as available in a number of different formats (Chatti et al., 2012). Examples of these different formats are summarized in Table 2.2.

Table 2.2: Format types and descriptions

Format Type	Description
File format	Files presented and opened using different applications such as MS-Excel (.xls, .xlsx), Adobe (.pdf), notepad (.txt), MS-Word (.doc, .docx) etc.
Structure format	Data within files presented differently using columns, rows, comma separated files (CSF). Files may also contain text, images, embedded multimedia files etc.
Data format	Data items within files use different formatting such as Dates (Date/Month/Year, Month, Date, Year), telephone numbers (including country codes, brackets, extensions)

Once the identification of data sources is complete, the contents of the data sources must be understood in order to effectively extract the data required. From a research perspective, the objectives of the study will guide the researcher in determining what data is required to be extracted from the datasets for analysis and prediction purposes. This section looks at the different types of data sources and what features (factors) can be found within these data sources. From an LA perspective, features are the variables or attributes in a dataset that are used to analyze the dataset with the objective of meeting the LA goal, such as the prediction or understanding of the learning outcome (Fong, Biuk-Aghai & Millham, 2018). The term feature is identified in many LA studies but attributes, variables and factors have also been synonymously used. The term factor is commonly used in research where the impact of a specific attribute is being studied.

For decades, researchers have been conducting studies with the objective of determining critical success factors for student academic performance. Understanding the importance of these factors can play a critical role from an academic standpoint as teaching staff can identify students that may potentially struggle, and assist in improving their marks (Yusuf & Lawan, 2018). From an administrative perspective, understanding the factors that play a role in academic success can help in identifying potentially good students to admit into their academic institution. This is important as, according to Chen, Hsieh and Do (2014), the levels of research and training improves when there are a better caliber of students registered at the academic institution.

There have been numerous studies related to student academic success factors at HEIs. The methodologies for these studies varied by studying data from various data sources, including attendance registers, assessment marks, questionnaires, interviews and secondary data, amongst other items. Thus, by identifying factors that determine academic success, the researcher can identify what data is important within a data source that can be used to predict student performance. The factors identified can fall into different categories depending on the data sources from which they are obtained. The most commonly identified data sources are listed in Table 2.3 (adapted from Adejo and Connolly (2017a)), along with examples of factors that relate to these categories.

Table 2.3: Categories of data sources with examples

Data source category	Examples of factors (features)
Demographic	Age, gender, place of birth, location, education, employment status
Academic background	GPA, high school marks, coursework assessments
Financial	Sources of funding including fee status, funding, tuition update, financial clearance status
Historical progression	Student graduation data, degree types, employment status, forwarding addresses, degree changes, de-registrations etc.
Behavioural (academic)	Interaction data that may include interaction logs (also referred to as clickstream data), discussion forums, course metadata, study methods, teaching and learning styles, reflection, click-through rates
Behavioural (human)	Stress, alcohol consumption, support structures, self-esteem, motivation and resiliency

Student demographic data is always captured by HEIs when students apply for admission to any program of HEIs. Date of birth (with age calculated where required), gender, race, nationality, employment status, and location are the most common features but other identified features from the literature include parents' occupation and/or qualifications (Pal, 2012; Werner, McDowell & Denner, 2013), disability status (Algur, Bhat & Ayachit, 2016), relationship status (Ghorbani & Ghousi, 2020), language of education (Gulati, 2015) and whether the student has siblings or children (Gulati, 2015).

In terms of academic background, many studies have looked at past academic performance as indicators to predict future academic performance. In this case, past academic performance includes high school marks (Hamoud, Humadi, Awadh & Hashim, 2017; Jayaprakash et al., 2014; Nudelman et al., 2019; Olaniyi, Kayode, Abiola, Tosin & Babatunde, 2017; Oloruntoba & Akinode, 2017) as well as marks obtained in previous and current courses registered for at the institution (Al luhaybi et al., 2018; Dorodchi et al., 2018; Hasan, Palaniappan, Raziff, Mahmood & Sarker, 2018; Mahzoon, Maher, Eltayeb, Dou & Grace, 2018; Salal, Abdullaev & Kumar, 2019).

Financial data sources relate to the student's current financial status, amount owing to the institution or for specific courses, student bursary/loan information and family (mother and father) financial status (Anuradha & Velmurugan, 2015; Ribot, Ribot, Perez & Cayabyab, 2020).

Historical progression is a data source with features relating to student graduation information, past qualifications, employment status and updated address details. These features can also form part of the demographic data as seen by Jayaprakash et al. (2014).

Behavioural factors may include student behaviour with regard to academic coursework. Before the advent of online learning and LMSs, this area was mostly limited to attendance to lecture, tutorial and/or practical sessions (Devadoss & Foltz, 1996; Fraser & Killen, 2005; Thatcher, Fridjhon & Cockcroft, 2007; Wadesango & Machingambi, 2011), insight into teaching quality (Wadesango & Machingambi, 2011) and study approaches (Ali, Haider, Munir, Khan & Ahmed, 2013). With the introduction of making course content and teaching content available via online learning applications such as an LMS, behavioural factors can now include online activities. According to Sclater et al. (2016), variables relating to how a student interacts with the content is far more effective in determining/predicting academic performance than past historical data or demographic data. This was particularly the case when comparing user clicks (total hits) and assessment clicks against the individual's characteristics and past academic performance.

Human behavioural factors have also been researched with regard to predicting student academic performance. Examples of these factors include motivation (Dennis, Phinney & Chuateco, 2005), stress (Pritchard & Wilson, 2003), self-esteem, fatigue (Pritchard & Wilson, 2003), peer support (Dennis et al., 2005), health status (Preetha, 2021) resiliency (McMillan & Reed, 1994), alcohol consumption (Pritchard & Wilson, 2003), and student self-reflection (Dorodchi et al., 2018). Psychological factors have also been identified as playing a role in predicting academic achievement, with studies such as Kappe and Van der Flier (2012) highlighting the critical role of personality traits and motivation.

Along with the identification of data sources, it is also important to understand that the ethical implications of using these data sources, the majority of which contain personal student information and actions performed by these students.

2.4. Ethics regarding data use in Learning Analytics

According to Leitner et al. (2019), ethics can be defined as various concerns regarding the understanding and defending of values such as life, security, happiness, health, knowledge, resources, freedom, etc. It is further described as the important decision of ascertaining what is right, wrong, good and bad before action can be taken (Adejo & Connolly, 2017a). From an LA perspective, ethics relates to how data used and generated in LA applications are interpreted by users and how they have an impact on students and their happiness (Adejo & Connolly, 2017a). According to Greller and Drachsler (2012), dealing with data in LA applications may result in stakeholders feeling that their privacy is at risk, resulting in resistance in LA and its further development.

While there have been several advances around LA development through EDM, visualization and other practical aspects, there continues to be debate related to the ethical uses of LA (Gupta & Saxena, 2021). The initial and current challenge relates to the lack of legal clarity with respect to data ownership (Guzmán-Valenzuela et al., 2021). In most research projects currently, data collected in a study belongs to the owner of the data collection tool as well as the institution conducting the research. The data collection tools are usually in the form of questionnaires and interview schedules that include attached ethical clearance and individual consent information. In the current environment, with the increase of new technologies such as GPS tracking and/or biometric sensors etc., there is an increase in the digital capturing of individual actions without the individual's awareness or even consent (Liu & Khalil, 2023).

Thus, it is important that LA initiatives be conducted in a manner such that the use of academic data not be abused. According to Greenleaf and Cottier (2020), at the end of 2020, a total of 142 countries had implemented data privacy laws that operate at both private and public levels. From a South African perspective, the Protection of Personal Information Act (POPIA) was approved on the 13th of November 2013 with the objective of protecting the personal information of both

public and private bodies (USAf, 2020). The main objectives of POPIA are to ensure that every South African's constitutional right to privacy is safeguarded, to balance the rights of privacy to that of other rights such as the access to information, to regulate how personal information is processed while ensuring that an individual's rights is/are protected, promoted and enforced, and to provide an individual with rights and guidance should privacy protection be broken (USAf, 2020). According to POPIA, personal information relates to any piece of information related to an individual that is living and can be identified. This includes (but is not limited to) information relating to a person's age, gender, marital status, physical or mental status, qualifications, medical history, financial status and history, religion, culture, personal opinions etc.

To assist researchers, a set of principles were outlined when dealing with data and POPIA (USAf, 2020). The four rules state that a researcher must de-identify the data as soon as possible, only collect data that is relevant to the study, ensure that the participants are aware of the study and how the data will be used, and finally, the data must be kept safe.

Once the ethical and privacy issues have been addressed along with the acquisition of the data sources, these data sources need to be prepared for LA applications. The issue of data preparation is covered in the next section.

2.5. Preparing data for Learning Analytics

An important issue identified before performing data analysis and prediction is the stage of data preparation or data pre-processing. The data preparation or pre-processing stages are areas of LA that has not received a lot of analysis and research (Munk, Drlík, Benko & Reichel, 2017; Romero, Romero & Ventura, 2014). Further to this, this stage of LA is seen as requiring a lot of effort and can form a large portion of the LA overall process (Romero & Ventura, 2020). Data preprocessing and preparation involves the detection, cleaning and filtering of any incomplete, missing, inconsistent and unnecessary data items (Tsai, Lai, Chao & Vasilakos, 2015).

Some of the more common challenges that must be addressed when preparing data for analysis and prediction are discussed in the following sections.

2.5.1. Handling missing or inconsistent data

The majority of datasets in the real world contain incomplete or partially complete data (Alexandropoulos, Kotsiantis & Vrahatis, 2019). There are several reasons why datasets may be incomplete, such as data items were lost, unavailable or not recorded at the time, or just forgotten by the data-capturer (Alexandropoulos et al., 2019).

Addressing the issue of missing data can be handled in a number of ways. Missing values can be replaced either by the most commonly found value in the dataset or an average can be calculated from existing values and be used. The value can also be replaced with a predicted value using a regression model (Alexandropoulos et al., 2019). The missing values can also be ignored when processing, i.e. only the values presented are used for analysis and/or prediction (Alasadi & Bhaya, 2017). Instances with missing data have also been known to be removed from the dataset altogether. Minaei-Bidgoli, Kashy, Kortemeyer and Punch (2003) reduced the number of students in their study from 261 to 227 as some students did not complete sufficient assignments to qualify for a final mark. In the cases of Kovanovic, Gašević, Dawson, Joksimovic and Baker (2016), as well as Waddington, Nam, Lonn and Teasley (2016), students that did not complete the course, for whatever reason, were removed from the dataset as these records did not have all assessment marks associated with them. Gudivada, Apon and Ding (2017) expressed caution when removing records with incomplete data as removing a large number of records will have an adverse effect on statistical results. An alternative to row deletion would be to remove only the attribute if that attribute has many values that are missing (Gudivada et al., 2017).

Inconsistent data refers to data item(s) that is/are different from other data items in the same attribute within the dataset or when compared to other datasets (Romero et al., 2014). Examples of this may include duplicate records from different periods of time. In this example, an age from one record may be different from the age of the associated duplicate record. Another example of inconsistency is when incorrect display formats are used, such as displaying dates as yyyy/mm/dd and mm/dd/yyyy (Romero et al., 2014).

2.5.2. Data discretization

This process involves the categorization of data ranges to improve comprehension and interpretation (Romero, Ventura & García, 2008). Knowles (2015) refers to this process as

recoding, where values of certain attributes are changed in order to be consistent across the datasets. A common example in educational data analysis is to categorize assessment scores into a specified number of categories. In the case of Naser, Zaqout, Ghosh, Atallah and Alajrami (2015), high school scores were categorized into one of a number of domains (for example, 1: Above 80%, 2: 75% to 79%, 3: 70% to 74%, etc.). A balance needs to be found with simplifying the data while at the same time not generalizing the data and losing valuable information. For example, an attribute for disability can be generalized to a yes/no value but this would result in a loss of information, such as the type of disability. On the other hand, having different types of disabilities included may result in excess information or data values that overlap (Knowles, 2015).

2.5.3. Noise and outlier detection

According to Romero et al. (2014), there are a select number of instances found in large datasets that do not match the behaviour of the other instances in the dataset. These instances are referred to as outliers. These outliers could be a natural occurrence or occur due to a mistake in data capture (Alasadi & Bhaya, 2017; Romero et al., 2014). Many algorithms such as binning have the ability to minimize or remove the influence of outliers (Alasadi & Bhaya, 2017). Romero et al. (2014) state that knowledge of the domain area is important to ascertain whether the outlier is a real possibility (e.g., an excellent student standing out from the rest of the class) or whether the outlier is a typographical error that needs to be corrected/removed.

2.5.4. Feature selection

Feature selection is the process of selecting relevant attributes from all available attributes. This task is necessary to remove attributes that are redundant or do not contribute to analysis or prediction techniques (Romero et al., 2014). Feature selection is said to be an important task in data preparation as it can improve accuracy (by reducing overfitting of a model, i.e. where the generated model works only for the data that was used to generate that model) and computation time (by removing unnecessary attributes) (Alexandropoulos et al., 2019; Romero et al., 2014).

2.5.5. Normalization and derivation

Normalization is a data transformation process where a data value is scaled to within a defined range. This range is usually from -1.0 to 1.0 or from 0.0 to 1.0 depending on the context of the problem being addressed. Normalization is said to potentially improve prediction accuracy and

efficiency of data mining algorithms by reducing the distance between maximum and minimum values (Romero et al., 2014).

Derivation is the process of creating new attributes from existing attributes. The attributes are usually achieved by applying some formula to other attributes that results in the new attribute value. This could be in the form of conversions, summations or count of values in other attributes (Romero et al., 2014).

2.5.6. Dealing with imbalanced datasets

Data that is continuously generated in real time is often prone to suffering from data imbalance (Madasamy & Ramaswami, 2017). This is a potential issue in the case of educational data. A dataset can be described to be imbalanced when the quantity of one of the classes (attribute values) is much greater than that of another class within the same attribute. The class that has a high representation is known as the majority class while the converse is referred to as the minority class (Madasamy & Ramaswami, 2017). From an educational perspective, an example of an imbalanced dataset would be one that has an extremely high proportion of students that have passed a course (majority class) compared to the proportion that failed (minority class). As stated by Kaur, Pannu and Malhi (2019), the process of classification of imbalanced datasets is a major problem in all domain areas (such as fraud and intrusion detection, image processing and medical science) and will result in reduced predictive performance of the generated model. This is because the model tends to display a stronger bias toward the class that has the majority instances (Bekkar & Alitouche, 2013; Madasamy & Ramaswami, 2017).

Common ways to address imbalanced datasets are from a data level or an algorithmic level (Bekkar & Alitouche, 2013). From a data level perspective, the sampling-based techniques of oversampling and undersampling are described as an effective way of dealing with data imbalance (Ghorbani & Ghousei, 2020). Oversampling is the process of increasing the number of instances of the minority class with the objective of reducing the imbalance. This is accomplished by duplicating minority class instances (Ghorbani & Ghousei, 2020). The disadvantages identified by oversampling, besides the increase in computational time, is a resultant bias or increased weighting towards minority class instances (Fernández, Garcia, Herrera & Chawla, 2018). Undersampling,

on the other hand, is the process of removing the number of instances in the majority class so that the imbalance is reduced. The advantage of this technique is that of reduced computational cost. However, the removal of instances to improve balance also results in reduced variance in the dataset. Furthermore, instances that may be useful for classification could also be removed (Fernández et al., 2018).

An additional sampling technique used is the synthetic minority oversampling technique or SMOTE. Unlike oversampling, SMOTE creates new instances of the minority class rather than duplicating minority class instances (Thai-Nghe, Busche & Schmidt-Thieme, 2009). The new instances are created based on variations of instances within the minority class (Fernández et al., 2018). While seen as a more effective sampling-based technique than oversampling and undersampling (Ghorbani & Ghousi, 2020), it has been noted that SMOTE can result in the generation of more unhelpful instances as well as results in the generation of noisy data (Jiang, Pan, Zhang & Yang, 2021).

From an algorithmic level, ensemble classifier algorithms or weight allocation have also been used to address the data imbalance problem (Bekkar & Alitouche, 2013). Ensemble classifiers are described as a combination of multiple learning algorithms (Madasamy & Ramaswami, 2017). While being a fairly new learning technique, ensemble classifiers have been seen to perform better in prediction accuracy than individual learning classifiers (Madasamy & Ramaswami, 2017). With regard to the weight allocation approach, a cost or weighting is allocated to individual instances or groups of instances. With this approach, the importance of specific instances (most likely in the minority class) are considered during the learning process. Identification of these instances and the exact weighting or cost requires understanding of the dataset and its context (Krawczyk, 2016).

2.5.7. Data formatting

Data formatting relates to the process of transforming data from its original form to a format that allows for it to be processed or analyzed by another application (Gao et al., 2016; Romero et al., 2014). Since data comes from various sources with different formats (most likely), it is necessary to ensure that a mechanism is in place to ensure that different representations of the same data are identified as one and the same, e.g., different date formats need to be made consistent to allow for

correct age calculations, naming conventions of data within files, file names and file extensions etc. (HersHKovitz & Alexandron, 2020).

In addition, formatting also deals with the conversion of the data file into a format that can be accepted by the application used for analysis. For example, the WEKA application accepts files in .arff format or as comma separated files (.csv). Without formatting, the application will not be able to distinguish between data attributes, rows or values, resulting in incorrect interpretation of the data.

Once the data has been prepared, the next stage is that of using the data for analysis and prediction, which is discussed next.

2.6. Data analysis and prediction in Learning Analytics

Learning Analytics requires the processing of large amounts of academic data, thus relying on a variety of techniques. Some of these techniques include classification/prediction, clustering, relationship and text mining, outlier detection process mining, statistics and visualization (Hooda & Rana, 2020; Kumar & Salal, 2019; Romero & Ventura, 2020). The techniques require the application of individual or combinations of algorithms in order to effectively analyze the data and/or make predictions. These algorithms can be classified either as supervised or unsupervised learning algorithms.

According to Berry, Mohamed and Yap (2019), supervised learning is the ability of a technique or algorithm to generalize knowledge from the provided data with labeled (known) instances. From this knowledge, the technique or algorithm would be able to predict target values for new or unseen instances. With supervised learning, the input dataset is divided into two parts: the training dataset and the test dataset. The selected supervised learning algorithm attempts to identify patterns using the training dataset. This is followed by applying these patterns to the test dataset with the objective of predicting an attribute value (Alloghani, Al-Jumeily, Mustafina, Hussain & Aljaaf, 2020). Examples of techniques using supervised learning include Decision Trees, Naïve Bayes and support vector machines (Alloghani et al., 2020; Limbu & Sah, 2019). Limbu and Sah (2019)

also identify neural networks, K-nearest neighbor algorithms and linear regression as being supervised learning algorithms.

On the other hand, unsupervised learning is the process of receiving unlabeled cases (instances) and the algorithm must learn or identify patterns in order to generate labels for these cases (Jain, Murty & Flynn, 1999). The algorithm accomplishes this by determining relationships from the available data and groups the data with similar features or characteristics (Berry et al., 2019). With regard to unsupervised learning algorithms, Limbu and Sah (2019) identify clustering algorithms (such as K-means and Gaussian Mixture models) as the most commonly used form of unsupervised learning.

The following subsections 2.6.1 to 2.6.6 cover commonly used algorithms identified in the literature that are applied to educational data for analysis or prediction purposes.

2.6.1. Clustering

According to Jain et al. (1999), clustering is an unsupervised learning algorithm that divides a set of observations or data items into groups referred to as clusters. Similarly, Hooda and Rana (2020) describe it as the process of identifying data items that are similar to each other, allowing for better decision making with regard to understanding the similarities and differences between datasets. Unlike classification, clustering is regarded as unsupervised learning in that no training set is provided. In addition, no labels are given to the data items. Rather, the learning process of clustering places data items into different groups (clusters) depending on the characteristics identified within each of these groups (clusters). Each group (cluster) represents different labels that have been generated based on what has been learnt (Jain et al., 1999). Clustering is used for pattern-analysis, grouping, document segmentation and pattern classification amongst others (Jain et al., 1999). From an LA perspective, Leitner et al. (2017) describes clustering as a grouping of similar material or students based on their learning and interaction patterns. This technique can be used to detect early drop-out of students, to better understand student interaction and engagement in the learning process.

A number of clustering algorithms exist, such as hierarchical clustering, K-means and fuzzy clustering amongst others. Clustering is also a commonly identified technique in the LA domain, the more recent studies of which are listed in Table 2.4. Hierarchical and K-means are identified as the most commonly used clustering algorithms. With the K-means clustering algorithm, objects are divided into an unknown number (k) of groups. An iterative process is often implemented in order to determine the ideal k value (Asif, Merceron, et al., 2017). With hierarchical clustering, all objects are initially their own cluster. The algorithm then identifies two objects with similar characteristics and merges these clusters. This continues until a diagram (called a dendogram) of a hierarchical series of nested clusters are formed. These are clusters of merged or broken up objects (Jain et al., 1999).

Table 2.4: Studies that used clustering with study objectives

Clustering Objective	Clustering algorithm	References
Health effect on academic performance	Hierarchical	Preetha (2021)
MOOC session analysis	K-means	de Barba et al. (2020)
Categorizing students based on marks	Hierarchical	Limbu and Sah (2019)
	K-means	Razaque et al. (2017)
Online interaction effect on engagement, literacy and performance	Hierarchical and non-hierarchical	Avcı and Ergün (2019)
	K-means	Khalil and Ebner (2017)
Categorizing students as procrastinators or not	K-means	Hooshyar, Pedaste and Yang (2019)
	K-means	Akram et al. (2019)
Determining attributes for underperforming students	Not specified	Ekubo and Esiefarienrhe (2019)
Determining at-risk status	Progressive	Mahzoon et al. (2018)
Identify learning strategies	Not specified	Gašević et al. (2017)
Identification of high, medium and low performing students	K-means	Asif, Merceron, et al. (2017)
Student drop-out analysis	K-means	Iam-On and Boongoen (2017)

2.6.2. Neural networks

A neural network is composed of an interconnected set of elements (known as neurons). The algorithm learns by adjusting the connections (referred to as weights) between the neurons, allowing for the neural network to perform a specific task or solve a problem (Beale, Hagan & Demuth, 2010). Typically, an input value(s) is provided into the neural network and the weights

are adjusted, resulting in an output value. Neural networks are useful for pattern discovery as well for classification problems (Adejo & Connolly, 2018; Beale et al., 2010). Recent studies that used neural networks, along with the study objectives and accuracy achieved, are listed in Table 2.5.

Table 2.5: Studies that used neural networks with objectives and accuracy achieved

Objective	Reference	Accuracy
Performance prediction	Bawah and Ussiph (2018)	90.4%
	Ha, Loan, Giap and Huong (2020)	86.1%
	Olive, Huynh, Reynolds, Dougiamas and Wiese (2019)	71.1 - 81.6%
	Umar (2019)	73.6%
	Adejo and Connolly (2018)	35 - 73.1%
	Asif, Hina and Haque (2017)	70.4%
	Asif, Merceron, et al. (2017)	62.5%
	Taodzera, Twala and Carroll (2017)	60.2%
Predicting procrastination	Hooshyar et al. (2019)	88.1 % - 99.5 %

2.6.3. Naïve Bayes

The Naïve Bayes algorithm assumes that all attributes in a dataset are independent of each other given a specific value (class). Using this assumption, the algorithm attempts to assign a value to a target attribute of a given instance (Rish, 2001). Based on the exact nature of the probability model, the dataset is then trained by the Naïve Bayes algorithm in a supervised learning setting. Despite the unlikely assumption of feature independence, Naïve Bayes has been noted to work effectively in solving many complex real-world problems. The benefits of using Naïve Bayes is reduced training time as well as removal of irrelevant features to improve classification performance (Kavipriya & Karthikeyan, 2019). A list of recent studies covering the Naïve Bayes algorithm is shown in Table 2.6.

Table 2.6: List of recent studies using Naïve Bayes algorithm with objective and accuracy

Objective	Reference	Accuracy
Performance prediction	Silva et al. (2022)	88.1%
	Ha et al. (2020)	86.1%
	Ndou, Ajoodha and Jadhav (2020)	83.4 – 84.4%
	Asif, Merceron, et al. (2017)	83.6%
	Asif, Hina and Haque (2017)	75.6%
	Taodzera et al. (2017)	63.4%
Predicting procrastination	Hooshyar et al. (2019)	80.5 – 99.4 %
Algorithm comparison	Fynn and Adamiak (2018)	59.4 – 90.2 % (different faculties)
Enrollment prediction	Wanjau and Muketha (2018)	72 %

2.6.4. Support Vector Machine

Support vector machines (SVM) are learning algorithms that are commonly used for pattern recognition, prediction tasks and data analysis. Assuming a given set of labeled instances (examples) belonging to one of two classes, the algorithm develops a linear model that is capable of assigning class values to unseen instances. When learning, a model (referred to as a hyperplane) is developed that separates the instances of the two different classes. According to Adejo and Connolly (2018), SVM can learn a greater number of patterns quickly and is more accurate in generalization because of its errors minimization capacity. In addition, it has the ability to update training patterns dynamically as more data instances are made available.

Some of the studies that applied SVM with the objective to predict performance accuracy are listed in Table 2.7.

Table 2.7: Accuracy achieved for prediction studies using SVM

Reference	Accuracy
Ha et al. (2020)	85.6 %
Ndou et al. (2020)	84.4 – 89.2 %
Eddin, Khodeir and Elnemr (2018)	49.2 %
Hooshyar et al. (2019)	71.9 – 99.6 %
Taodzera et al. (2017)	64.6 %

2.6.5. Random Forest

The Random Forest algorithm falls under the category of ensemble algorithms, which can be defined as a combination of classifiers into a meta classifier. It is a process of utilizing multiple algorithms with the objective of obtaining better predictions when compared to using just a single classifier algorithm (Madasamy & Ramaswami, 2017). In the case of Random Forest, Kovanović et al. (2018) describe it as a combination of a large number of Decision Trees with the final classification model being obtained via a voting mechanism built into the algorithm. Each constructed Decision Tree is based on population sub-sample referred to as a bootstrap. These bootstraps contain random instances with some of these instances being duplicated. The resultant tree is then evaluated against a sample of instances that were not part of the bootstrap. Further to this, each Decision Tree is created using only a subset of the features (attributes) of the dataset (Kovanović et al., 2018). Ensemble algorithms have been noted to produce models with greater

accuracy and higher generalization capacity (Kovanović et al., 2018). The success of the algorithm is further justified by Batool et al. (2023) who conducted a literature survey study of 260 articles and identified Random Forest as one of the most commonly used and successful algorithms for performance prediction.

A number of studies covering the Random Forest algorithm are covered in the literature. These studies are listed in Table 2.8:

Table 2.8: Objectives of studies using Random Forest algorithm with accuracy achieved

Objective	Reference	Accuracy
Performance prediction	Silva et al. (2022)	97.5%
	Akram et al. (2019)	87.2 – 95.4%
	Ndou et al. (2020)	93 – 95%
	Sandoval, Gonzalez, Alarcon, Pichara and Montenegro (2018)	82 – 86.1%
	Adejo and Connolly (2018)	73.1 – 81.6%
	Ha et al. (2020)	80.7%
	Eddin et al. (2018)	72.8%
	Asif, Merceron, et al. (2017)	71.1%
	Asif, Hina and Haque (2017)	69.5%
Predicting procrastination	Hooshyar et al. (2019)	86.8 – 99.5%
Understand student self- reflection	Kovanović et al. (2018)	87%

2.6.6. Decision Tree algorithms

According to Nudelman et al. (2019), Decision Tree algorithms apply the concept of information entropy to divide the classification process into smaller sub-problems which are easier to solve. As the name states, the algorithm represents a tree structure made up of nodes and branches. Each node represents an attribute of the dataset and a number of branches stem from the node, where each branch represents a value that the attribute can take (Alloghani et al., 2020). A node in a Decision Tree is continuously divided into sub-nodes via its descendants. A node with zero (0) descendants indicates a prediction has been made. Nudelman et al. (2019) states that an attribute's influence is determined by its place in the Decision Tree: the higher the node (attribute), the greater the influence the attribute has in predicting a value.

Decision tree algorithms are one of the most commonly used algorithms for performing educational prediction or classification (Kumar & Salal, 2019; Wise, 2019). Table 2.9 lists recent

studies that have implemented Decision Tree algorithms grouped by the main objective of each of the studies. The majority of studies focused on accuracy as the primary measure for performance, although other performance measures were also reported upon (discussed further in chapter 7). Thus, the accuracy of each study is also included in Table 2.9.

Table 2.9: Objectives for studies using Decision Tree algorithms

Objective (number of studies)	Reference	Accuracy
Performance prediction (21)	Bawah and Ussiph (2018)	100%
	Saheed, Oladele, Akanni and Ibrahim (2018)	98.3%
	Akram et al. (2019)	94.5%
	Nudelman et al. (2019)	92%
	Ndou et al. (2020)	91.4%
	Agrawal, Vishwakarma and Sharma (2017)	90%
	Hasan et al. (2020)	87%
	Sunday et al. (2020)	87%
	Khakata, Omwenga and Msanjila (2019)	84.6%
	Abaah Jnr (2019)	82%
	Tegege and Alemu (2018)	81.4%
	Adejo and Connolly (2018)	78%
	Silva et al. (2022);	77.5%
	Asif, Hina and Haque (2017)	74.7%
	Jalota and Agrawal (2019)	73.6%
	Ha et al. (2020)	73.4%
	Taodzera et al. (2017)	65.8%
	Olaniyi et al. (2017)	65.7%
Hasan et al. (2018)	63.6%	
Hamoud, Hashim and Awadh (2018)	63.4%	
Kumar and Singh (2017)	61.4%	
At-risk prediction (2)	Ribot et al. (2020)	92.1%
	Al luhaybi et al. (2018)	84%
Procrastination prediction (1)	Hooshyar et al. (2019)	99.6%
Enrollment prediction (1)	Wanjau and Muketha (2018)	84%
Drop-out prediction (1)	Viloria et al. (2020)	79.8%
Algorithm comparisons (3)	Fynn and Adamiak (2018)	90.5%
	Eddin et al. (2018)	72.5%
	Asif, Merceron, et al. (2017)	69.2%

A commonly identified advantage of using a Decision Tree algorithm is that the model created is usually easy to understand by the analyst and end user (Yusuf & Lawan, 2018). From an implementation standpoint, Decision Tree algorithms are flexible enough to handle different input data types, namely text, numeric and nominal data types. Decision tree algorithms are also able to process erroneous or missing data values by creating branches specifically for these problematic values. Finally, Decision Tree algorithms are known to be implemented fairly quickly with minimal time to create the model (Hamoud et al., 2018).

In order to conduct the analysis described in this section, various software or tools are available. Some of the more common tools are described in the next section.

2.7. Common tools or applications used for Learning Analytics

This section describes common tools or applications used for LA tasks. These tools relate to commonly identified software applications aimed at applying different algorithms and functions to user datasets with the objective of data cleaning, preparation, analysis and prediction of student's learning interaction and performance.

2.7.1. WEKA

The Waikato Environment for Knowledge Analysis (WEKA) tool is an open-source application that can be used for various data mining tasks and is an accepted tool for performing student prediction tasks (Batoool et al., 2023). To accomplish these tasks, the application is constituted of a number of algorithms and functions that can be used for preprocessing, classification, clustering and attribute selection, amongst others (Vambe & Sibanda, 2017). WEKA is developed using the Java programming language and is available for use on most operating systems. WEKA accepts data as a single flat file specified in .arff (Attribute-Relation File Format) format as well as .csv format amongst others.

Advantages of WEKA, according to Abaah Jnr (2019) is that it is open-source and freely available, platform independent, and can be easily used by non-data mining specialists. Salihoun (2020) also stated the availability of online support via WEKA mailing lists, tutorials, wikis and bug reports.

WEKA also allows for the addition of user created algorithms or downloading of algorithms and functions created by others in the online analytics community.

2.7.2. KNIME

According to Berthold et al. (2009), the Konstanz Information Miner (KNIME) is a modular environment that allows for visual assembly and interactive execution of a data mining task. The tool is open-source and allows for both data mining and reporting tasks similar to that of WEKA (Salihoun, 2020). Also, similar to WEKA, KNIME incorporates integration of user created algorithms and tools for data mining purposes. The modular nature of KNIME allows for the ability to incorporate a number of different data sources in different formats such as database files, MS-Excel files, .csv files, .arff files etc.

A KNIME workflow is created using a combination of nodes, with each node performing a specific function such as pre-processing, analysis, colour allocation, machine learning application, graph display, etc. (Berthold et al., 2009). Connections are formed between the nodes, indicating the transport of data between two nodes. An advantage of this approach is that the workflow node stores the result permanently and can be stopped at any time to be resumed later. A user can then adjust the nodes and the entire workflow need not be started from the beginning (Berthold et al., 2009).

2.7.3. R

R is an open-source programming language as well as a data analysis environment. As with WEKA and KNIME, being open-source allows for the development of new techniques and functions that can be incorporated into the R environment for use by data scientists (Patil, 2016). The common version of R consists of an Integrated Development Environment (IDE) consisting of a console window, workspace view and data editor.

R provides a wide variety of statistical, graphical and machine learning techniques. It has several built-in functions to allow for data extraction, data preparation, statistical analysis, predictive modelling and data visualization. It is one of the more popular tools used in industry and has a growing online community support that updates and adds new functionality consistently (Prajapati, 2013).

2.7.4. Python and Jupyter notebook

Python is a programming language that allows for the manipulation of data and engineering of various techniques and functions. Salihoun (2020) states that Python with Jupyter Notebook is an interactive environment with features for the creation and sharing of documents as well as data cleaning and transformation, simulation, modeling, visualization and machine learning. Besides this, one of its main functions is to keep track of the research process (Randles, Pasquetto, Golshan & Borgman, 2017). From an academic perspective, keeping track of the steps of the research process allows for better reproduction or replication of any experiments undertaken in the research (Randles et al., 2017).

Jupyter notebook allows for storage of data within online repositories that can be easily accessed by a variety of research objects. Jupyter notebook also has the advantage of being both machine and human-readable, allowing for interoperability with other compatible applications as well as for academic communication (Randles et al., 2017).

2.7.5. RapidMiner

RapidMiner is another popular tool used in analytics due to its easy to learn user interface (Prekopcsak, Makrai, Henk & Gaspar-Papanek, 2011). It is a data science software platform that allows for data preparation, machine learning as well as predictive analytics. The tool contains a number of built-in algorithms for handling classification, clustering, rule mining, regression and others (Salihoun, 2020). Similar to WEKA and KNIME, it also allows for the addition of user created extensions that can provide additional statistical, analytical and machine learning functions (Prekopcsak et al., 2011). RapidMiner is free and open-source and a number of tutorials are available to assist new users (Salihoun, 2020).

2.8. Identification of potential gaps in the literature

This literature review chapter outlines the influence of LA within the higher education environment. The general definition of LA is given, followed by how it fits into the world of Big Data in higher education. The most common aspects of the LA process are discussed, these being data acquisition, data preparation, algorithms applied to the data, and commonly used tools for LA application studies. From an algorithm perspective, the most commonly used algorithms in recent

literature were Decision Trees, clustering, Naïve Bayes, Neural Networks and Random Forest algorithms.

After reviewing the literature found, LA or EDM research is being conducted in numerous countries, with the majority of studies emanating from the first world countries such as those in Europe, the United States and Australasia. From a developing country perspective, there is slow progress in the development of LA within the African continent (Prinsloo & Kaliisa, 2022b). In a 2023 literature survey study by Sghir, Adadi and Lahmer (2023), out of 74 studies identified between 2012 and 2022, the majority of studies emanated from the United Kingdom, USA, India and Spain with only one (1) study identified from Africa. Table 2.10 outlines studies that were found relating to LA or EDM that involved countries in Africa from 2017 to 2022 (six years).

Table 2.10: Recent LA/EDM application studies conducted in Africa

LA/EDM applications (19)	LA implementation research (2)
Olaniyi et al. (2017); Mwalumbwe and Mtebe (2017); Taodzera et al. (2017); Vambe and Sibanda (2017); Oloruntoba and Akinode (2017); Bawah and Ussiph (2018); Saheed et al. (2018); Tegege and Alemu (2018); Wanjau and Muketha (2018); Kritzinger, Lemmens and Potgieter (2018); Popoola et al. (2018); Gulint and Adam (2019); Khakata et al. (2019); Nudelman et al. (2019); Adekitan and Salau (2019); Ogunde and Ajibade (2019); Umar (2019); Ndou et al. (2020); Sunday et al. (2020)	Prinsloo and Slade (2017); Okewu and Daramola (2017);
	LA/EDM overviews (1)
	Maphosa and Maphosa (2020)
Performance comparisons (1)	Challenges for adoption (6)
Fynn and Adamiak (2018)	Prinsloo (2018); Prinsloo, Slade and Khalil (2018); Ngqulu (2018); Olivier (2020); Prinsloo and Kaliisa (2022b); Prinsloo and Kaliisa (2022a)

As can be seen in Table 2.10, only 29 articles over six (6) years were identified, indicating the lack of research related to LA within the African continent. The lack of LA or EDM research was also observed by both Prinsloo and Kaliisa (2022b) and Maphosa and Maphosa (2020) respectively. Table 2.11 outlines a summary of the characteristics of the LA/EDM application studies conducted in Africa.

Table 2.11: LA/EDM Africa-based application studies with problem characteristics

Study	Country	Data Source	No. of students in study	Technique
Mwalumbwe and Mtebe (2017)	Tanzania	LMS (2 Courses)	171	Correlation Regression
Taodzera et al. (2017)	South Africa	Demographics School marks School details	1366	SVM Neural network Decision tree Regression Naïve Bayes
Olaniyi et al. (2017)	Nigeria	Past university marks Course activities	285	BFTree CART Decision tree
Vambe and Sibanda (2017)	South Africa	Past university marks	476	Decision tree
Oloruntoba and Akinode (2017)	Nigeria	School marks Past university marks	89	SVM Neural network Decision tree Regression
Bawah and Ussiph (2018)	Ghana	School marks School details Past university marks	525	Neural network Decision tree K-nearest neighbour
Saheed et al. (2018)	Nigeria	Demographics Financial Lecture attendance	234	Decision tree Regression trees
Tegegne and Alemu (2018)	Ethiopia	School marks Entry exam marks Choice of degree 1 st year marks	5729	Decision tree
Fynn and Adamiak (2018)	South Africa	Demographic Registration School marks Past university marks	186 174 instances	ZeroR OneR Naïve Bayes Regression Decision tree
Wanjau and Muketha (2018)	Kenya	School marks Financial Opinion Demographics	209	Decision tree Naïve Bayes Regression trees
Kritzinger et al. (2018)	South Africa	Demographics School marks Past university marks Learning strategies	1084	ANOVA CHAID analysis

Continued on next page...

Table 2.11 continued				
Study	Country	Data Source	No. of students in study	Technique
Popoola et al. (2018)	Nigeria	Past university marks	1841	Descriptive statistics Frequency distribution ANOVA Post-hoc tests
Gulint and Adam (2019)	Ethiopia	Opinion	5454 instances	Association rules
Khakata et al. (2019)	Kenya	Opinion	747	Decision tree
Adekitan and Salau (2019)	Nigeria	Past university marks	1841 - Same as Popoola et al. (2018)	Neural network Random Forest Decision trees Naïve bayes Tree ensemble Regression
Ogunde and Ajibade (2019)	Nigeria	Past university marks	10601 instances	K-nearest neighbour
Nudelman et al. (2019)	South Africa	School marks Demographics Registration	783	Random Forest Decision Tree Naïve Bayes Bayesian Network
Umar (2019)	Nigeria	Demographics School marks	61	Neural Network
Ndou et al. (2020)	South Africa	School marks Demographics	2000	Decision Trees Naïve Bayes Random Forest SMO Regression Logistic Model Trees
Sunday et al. (2020)	Nigeria	Past university marks Course activities	239	Decision Tree

The dataset by Ndou et al. (2020) was a synthetically created dataset (based on SA student data) of 50000 students with 2000 students sampled for the purposes of the study. This dataset is the only publicly available dataset. The other datasets identified either focus on an individual course, a small number of courses, or are a collection of students from a variety of degrees or colleges. Many studies, for example Fynn and Adamiak (2018), consider the student instances and attributes as a whole rather than based on the degrees that they are doing or the courses that they are registered for. It cannot be assumed that all colleges and degrees are the same and that the same

attributes can be applied for prediction. For example, a student's mathematics result may be a better predictor at a Science-based college than at an Arts college.

2.9 Chapter Summary

The chapter provided an overview of LA, first covering the concepts (Section 2.2) and then different processes involved in LA (Sections 2.3 to 2.6). Section 2.7 covered the most commonly identified tools used for LA based on a survey of the literature.

One of the key points noted was the lack of LA/EDM studies conducted in Africa when compared to the rest of the world. It was evident and noted by other authors that Africa has been slow to take advantage of the benefits of LA, however this has been due to other challenges such as lack of technological infrastructure.

The development of datasets and making them available would further encourage application and evidence-based research in LA. This would then allow for further research on LA implementation studies within HEIs in Africa.

All studies identified have individually focused on specific aspects of LA such as ethical issues, data preparation, pre-processing or learning algorithm application. No studies were identified that systematically cover the full LA process from the data acquisition (including ethical clearance, data collection, preparation and preprocessing) to application of learning algorithms and artificial intelligence techniques with discussion of results.

The next chapter covers the research methodology for this research, including the adopted LA research model that covers the entire LA process, aspects of data acquisition, and choice of application that will be used to further the knowledge in the LA field.

Chapter 3 – Research methodology

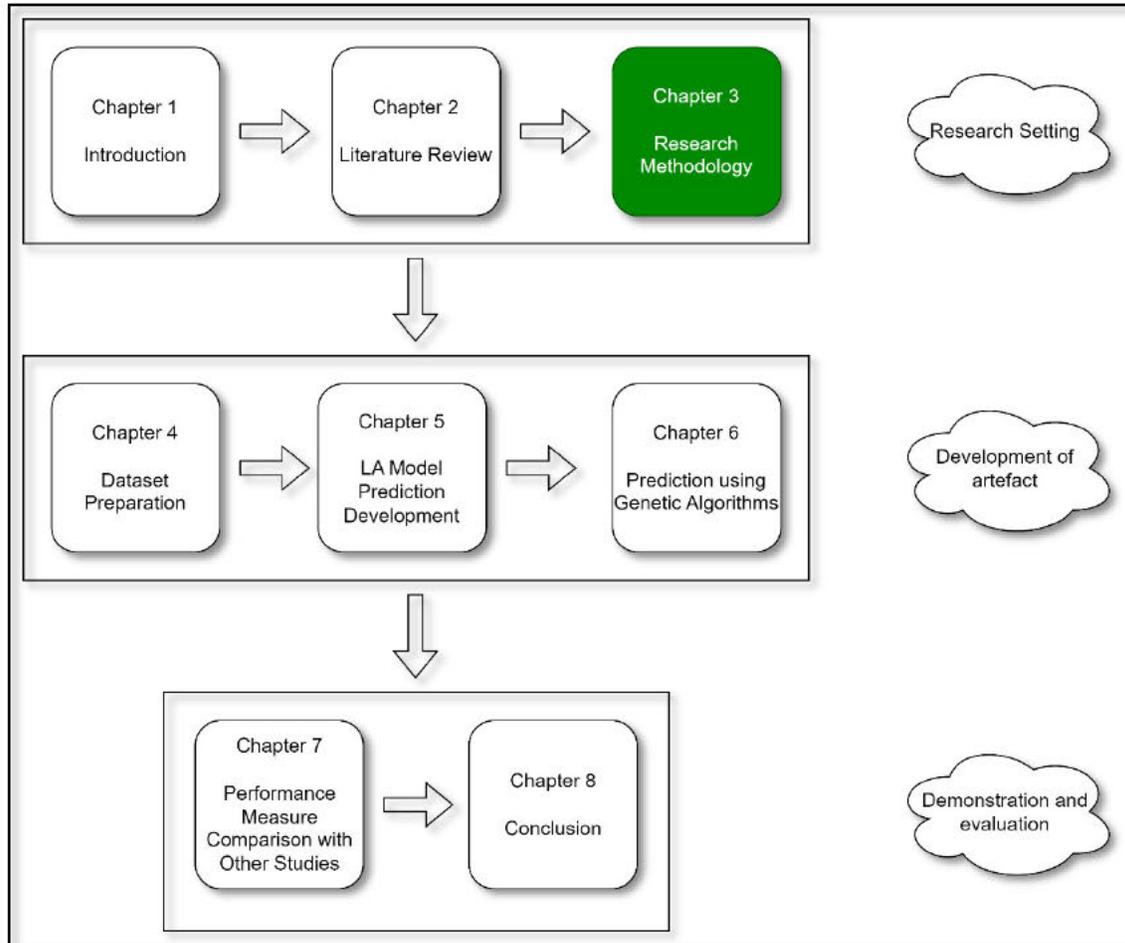


Figure 3.1: Thesis structure

3.1. Introduction

According to Rajasekar and Verma (2013), research is defined as a process of the logical and/or systematic discovering of novel information about a particular topic of interest. The objective of the research covered in this dissertation was to better understand how learning analytics can be applied to three UKZN datasets consisting of IS&T student data. Chapter 3 is the final chapter used to establish the setting of the research (see Figure 3.1) and describes the research methodology.

According to Creswell (2014), a research approach (methodology) is a set of plans and procedures for a research project. This involves the description of assumptions and detailed steps related to collection of data, data analysis, interpretation of results and all other steps involved in the research project. As described in the first two chapters, this study focuses on learning analytics (LA) and, as with any other study, also requires a research approach. An overview of the chapter is illustrated in Figure 3.2 below and described thereafter.

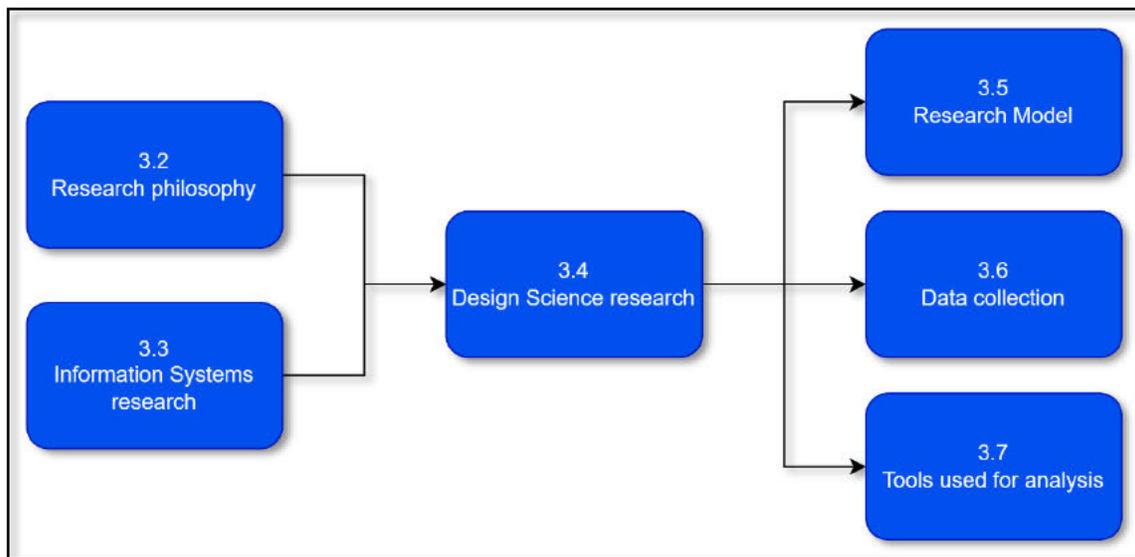


Figure 3.2: Map for Chapter 3 coverage

Section 3.2 covers the common research philosophies used in research that are often used to guide the direction of the research. Learning analytics can fall under the domain of Information Systems (IS) research where two approaches are most commonly followed: Behavioural Science and Design Science (Hevner, March, Park & Ram, 2004). Section 3.3 discusses these two approaches, and why, in this case, the Design Science approach is most appropriate. The Design Science approach is then covered in relation to this study (Section 3.4), i.e. how the research objectives align with the design science model that is being used. This is followed by the research model to be used (Section 3.5), method of data collection (Section 3.6) as well as how analysis and prediction methods were conducted in order to meet the objectives of the study (Section 3.7).

3.2. Research philosophies

According to Žukauskas, Vveinhardt and Andriukaitienė (2018), the research philosophy can guide the researcher towards choosing the most appropriate strategy, problem formulation, mode of data collection, processing and analysis. This section, therefore, covers four commonly identified research philosophies, that being positivist, interpretivist, realistic, and pragmatic philosophies.

The positivist research philosophy assumes that the world can be viewed objectively and the researcher can work independently and not be biased (Žukauskas et al., 2018). The opposite of positivism is the interpretivist/constructivist philosophy where it is understood that the world position is subjective in nature (Žukauskas et al., 2018). A pragmatic research philosophy is one that is dependent on the research problem where the researcher approaches the problem in the best manner required to solve the problem (Žukauskas et al., 2018). Finally, the realistic research philosophy is one that combines the positivist and interpretivist philosophies. Table 3.1, adapted from Žukauskas et al. (2018), outlines the research philosophy as well as the procedure and tools used for data collection.

Table 3.1: Philosophies, research methods and suggested instruments

Research philosophy	Research method	Research instrument examples
Positivism	Quantitative	Experiments Tests Scales
Interpretivism	Qualitative	Interview Observation Document/file study Image data analysis
Pragmatism	Qualitative and/or quantitative	Instruments from positivism as well as interpretivism
Realism	Qualitative, quantitative and mixed methods	Variety of measures to reduce bias

As can be seen in Table 3.1, with a positivist viewpoint, a quantitative research approach is generally preferred as only quantifiable data is considered as evidence (Giddings & Grant, 2006).

A large amount of data collected is preferable in order to improve the likelihood of statistical significant correlation (Giddings & Grant, 2006). With regard to the interpretivist viewpoint, a qualitative research method is preferred where the data collected is in the form of perspective. A number of different perspectives, in combination with the researcher's perspective, allows for a more holistic truth of the subject of study (Giddings & Grant, 2006). The viewpoint of the realist would result in following a combination of qualitative, quantitative and mixed methods as this would depend on contextual and historical factors (Žukauskas et al., 2018). Finally, a pragmatic viewpoint will follow a methodology dependent on what is required to solve the problem (Žukauskas et al., 2018).

With the research philosophies described above, the next section discusses an area that LA falls under, that being Information Systems research.

3.3. Information Systems research

The objective of IS within an organization is to ensure the continuous improvement of efficiency and effectiveness of processes within that organization (Hevner et al., 2004). Thus, the objective of any research endeavor in IS is to improve the body of knowledge that assists in the improvement in the application of Information Technology in the relevant organization (Hevner et al., 2004). The research paradigms of behavioural science and design science play a role in improving the IS body of knowledge (Hevner & Chatterjee, 2010).

Behavioural Science stems from the methods used for natural science research. From an IS perspective, this research paradigm is used when testing and/or justifying theories in order to explain the IS related activities (analysis, design, implementation and/or use) within an organization (Hevner & Chatterjee, 2010). The importance of this area of research is that it provides practitioners with relevant information regarding the actions of people, technology and organizations and how these actions should be managed to improve efficiency and effectiveness (Hevner & Chatterjee, 2010). In fact, according to Hevner and Chatterjee (2010), this form of research has been dominant in the IS discipline where the majority of studies try to understand the impact of artefacts (for example design models and technology) and its effect on people and organizations.

Design Science research is often seen as a research area that falls within the computer science and engineering disciplines. Some of the problem characteristics where design science research is suitable include unstable requirements, constraints defined within an uncertain environmental context, complex interactions amongst the problem subcomponents, difficulty in adapting to change, and a critical dependence on human cognitive abilities to produce effective solutions (Hevner & Chatterjee, 2010). Design science results in the development of new artefact(s) which can be evaluated using behavioural science methodologies. Thus, behavioural science and design science have an important relationship that is necessary for the continuing development and enhancement of the knowledge base in IS. This relationship is illustrated in Figure 3.3.

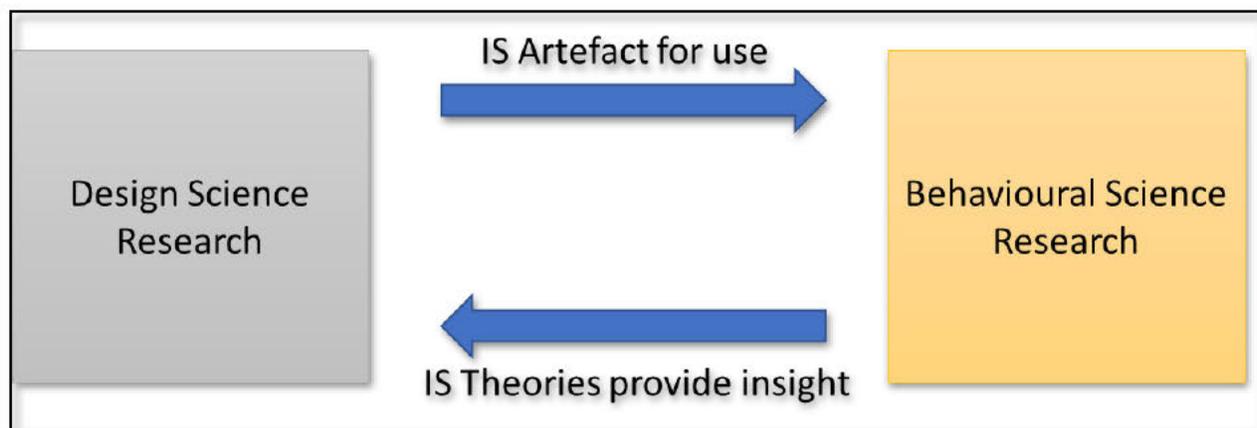


Figure 3.3: Complementary relationship between design science and behavioural science research areas (Hevner & Chatterjee, 2010)

In the context of this study, the Design Science research methodology with a pragmatic philosophy was used and an artefact was developed. By addressing the research objectives of this study, the artefact developed for this study was in the form of a process model that can be used as part of analysis of a student profile, as well as for the prediction of student academic performance based on the aspects of the student profile. The next sections provide more detail on the Design Science research model as well as details on the development of this artefact.

3.4. Design Science research approach

According to Hevner and Chatterjee (2010), Design Science falls under a pragmatic research paradigm, meaning that the main goal is to address the research objectives. A pragmatic philosophy allows for the research objectives to be dealt with using any philosophy (such as

positivism, post modernism or critical realism) that would best allow for the requirements of the objectives to be met (Saunders, Lewis & Thornhill, 2019). There are two main objectives in Design Science, these being the development of an artefact(s) as well as the evaluation and fitting of the artefact(s) to solve the problem (Cronholm & Göbel, 2016).

To assist with carrying out of Design Science research, Hevner et al. (2004) provided a set of seven (7) guidelines. These guidelines, when followed with an appropriate Design Science methodology, will allow for the implementation of an effective research project. The guidelines are summarized in the table below:

Table 3.2: Design Science guidelines by Hevner et al. (2004)

Guideline	Description
1. Design as an artefact	The Design Science research project must produce a complete, usable artefact.
2. Relevance to the problem	The Design Science research project must result in the development of a solution to the relevant problem.
3. Evaluation of the design	The usability and quality of the developed artefact must be determined through demonstration and compared against some performance criteria
4. Contribution of the research	The Design Science research conducted must have a significant and verifiable contribution towards the artefact, its design or developments as well as the methodologies used.
5. Research rigor	Methodologies used must be carried out in a disciplined manner during both the development and the evaluation of the artefact.
6. Design as a search process	The development of a viable, quality artefact requires the effective use of any or all legal resources available.
7. Research communication	The Design Science research implemented must be effectively presented to all relevant stakeholders.

There are a number of methodology models identified for Design Science research. This includes the Design Science Research Model (DSRM) by Peffers, Tuunanen, Rothenberger and Chatterjee (2007), the Design Science Research Process Model by Vaishnavi, Kuechler and Petter (2004), the Design Science Research method for Decision Support Systems development (Arnott, 2006), Soft Design Science Methodology (Baskerville, Pries-Heje & Venable, 2009) and the Learning Analytics Information Systems (LAIS) Design Methodology by (Nguyen, Gardner & Sheridan, 2020). Each of these models have similar activities such as identification of the problem, suggesting of solutions, development of an artefact and evaluation or comparison of the artefacts

For this research, the DSRM as proposed by Peffers et al. (2007) was followed when developing the artefact. While the other models had similar stages, the model by Peffers et al. (2007) suggests an iterative approach, thus allowing for a pragmatic philosophy being adopted. A pragmatic philosophy was also suggested Hevner et al. (2004) in terms of guidelines 6 (see Table 3.2). The LAIS proposed by Nguyen et al. (2020) also follows an iterative approach but is more focused on LA implementation at an institutional level and includes architectural and service based activities that are outside the scope of this study. Thus, the DSRM model by Peffers et al. (2007) is well suited due to its ability to work through what can be classified as a practical, real-world problem. The iterative nature of the model (as shown in Figure 3.4), allows for the development of effective artefacts that can be used by teaching staff to improve teaching and learning outcomes (Chatti et al., 2012).

The DSRM is made up of six (6) main activities. The authors of the DSRM identified these activities based on common steps followed by leading design science researchers. The description of these activities as well as how they relates to this research is described in subsections 3.4.1 to 3.4.6.

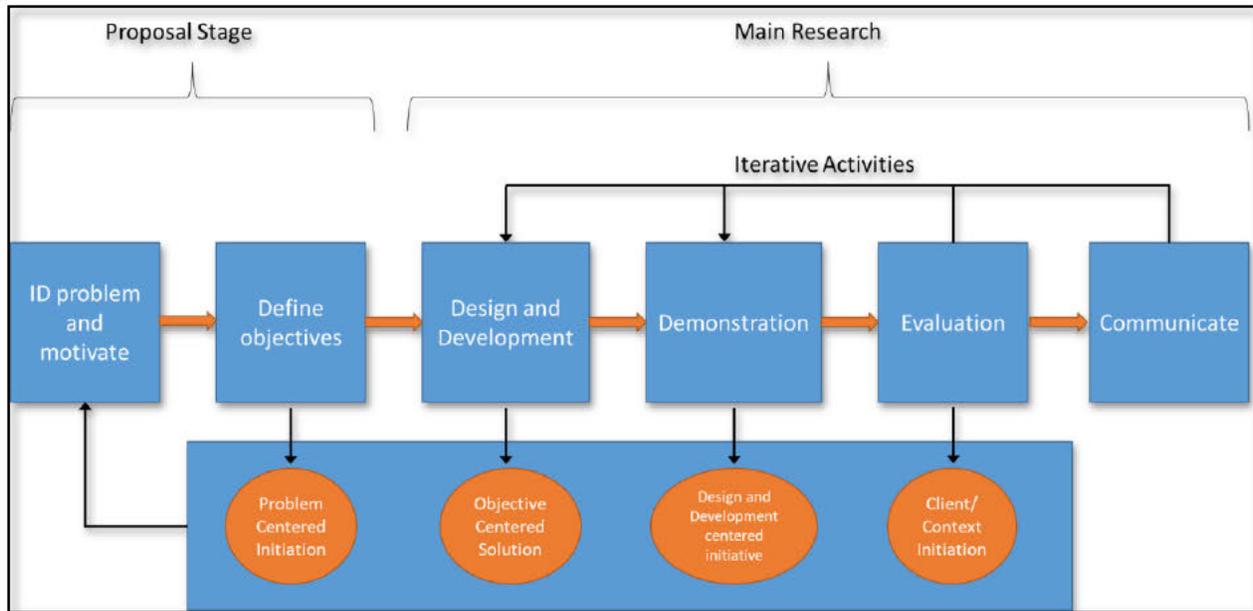


Figure 3.4: DSRM Framework (Peppers et al., 2007)

3.4.1. Problem identification and motivation

This activity involves forming the definition of the problem as well as the rewards or benefits of the solution. This is necessary as it assists in identifying what artefact(s) must be developed. Peppers et al. (2007) suggest that the problem be broken down conceptually so that the artefact(s) can meet the complex requirements. This stage was similarly described in other design models as problem awareness (Vaishnavi et al., 2004), problem recognition (Arnott, 2006) or problem situation (Baskerville et al., 2009). Problem identification arises from new developments in the problem domain and are usually in the form of situations that need to be resolved rather than explanations to unanswered questions (Vaishnavi et al., 2004).

From the perspective of this study, the problem statement outlined in Section 1.3 as well as the motivation for solving the problem has been established. The problem has also been broken down into research questions that outline the complexity of the problem. The problem statement, taken from Section 1.3 is as follows:

In order to predict and understand student academic performance in higher education institutions, the use of technology in learning analytics has become increasingly important due to limited resources and an ever-increasing number of students.

The motivation for this study, as described in Section 1.7, is to better use the large amounts of data continuously stored at higher education institutions with the objective of better understanding and predicting student academic performance, thereby improving student learning outcomes.

3.4.2. Defining the objectives for a solution

From the problem statement, the objectives of the study are conceptualized (Section 1.5). The objectives to the solution can be quantitative, such as measures that could determine when a solution is better than the current scenario, or qualitative in nature, such as how the new solution would solve the current problem scenario (Peffer et al., 2007). This stage was also referred to as “Suggestion” (Arnott, 2006; Vaishnavi et al., 2004).

Peffer et al. (2007) note that the objectives for a Design Science research study focus on design and development. In the context of this study, the research objectives also focus on design and development, which when completed, will result in the completion of the artefact. The research objectives for this study (from Section 1.5) are listed below:

1. To integrate the relevant university data sources in preparation for classification.

To achieve this objective, an artefact is developed that guides the collection and organization of the data in a form such that the data can be applied to machine learning algorithms and/or artificial intelligence techniques.

2. To extract, clean and classify the integrated data.

In order to address this objective, aspects of data preparation are addressed. This includes dealing with the removal of redundant or duplicate data as well as converting the data into an appropriate format in order to ensure effective processing and prediction.

3. To train the data in order to determine patterns and useful information for student performance prediction.

This objective is addressed by utilizing feature selection in combination with machine learning and/or artificial intelligence methods in order to develop prediction models.

4. To determine the effectiveness of the training techniques by evaluating their accuracy in terms of how they predict student performance.

This will be accomplished by performing a comparison between the results of the artefact with data and results that have already been generated to determine how accurately the artefact can make predictions.

5. To evaluate the results generated by the artefact against other similar artefacts.

This will be accomplished by comparing the performance of the model generated through the artefact against that of other LA studies in the literature based on accuracy and other assessment metrics.

3.4.3. Design and development

This activity involves the creation of the artefact. As described in the previous section, an artefact can be a construct, model, method, instance or any technical resource or information that contributes towards enhancing the function, effectiveness and/or efficiency of an organization, team or people involved in Information Systems (Peppers et al., 2007). This activity also includes determining the functionality of the artefact and the use of resources required to develop the artefact. These resources include the knowledge base that can be used to develop the artefact. Table 3.3 provides the most common artefact types that result from Design Science research (Vaishnavi et al., 2004).

Table 3.3: List of potential types of artefacts (Vaishnavi et al., 2004)

Artefact type	Description
1. Constructs	Theoretical concepts within the domain of study
2. Models	Set of relationships between constructs
3. Frameworks	Output to support or guide through a process
4. Architectures	High level system structures
5. Design principles	Core rules or philosophies to guide through a design process
6. Methods	Step-by-step guide to follow for task completion
7. Instantations	Output that is the result of implementing other artefacts such as methods, frameworks, models, design principles or constructs
8. Design theories	Combination of one or more artefacts that results in prescriptive statements for meeting an objective

From the perspective of this study, one of the artefacts developed was a process model, outlining the steps required to analyze and predict student academic performance based on the data sources provided. This process model combines the characteristics of artefact type 6 and artefact type 3 (see Table 3.3) in that it provides a step by step process to guide LA researchers. The process model is presented in Section 3.5 while the aspects considered during the design and development of this model are covered in the next three chapters. Chapter 4 covers the data collection, cleaning and preparation stages. Chapter 5 and Chapter 6 cover the description of the machine learning and artificial intelligence techniques used in this study to develop and test the student academic performance prediction models. In this case, the techniques are the Decision Tree and Random Forest algorithms (machine learning techniques) as well as the Genetic Algorithm and Optimized Forest algorithm (artificial intelligence).

In addition, a further artefact developed from this study was in the form of a dataset. In the case of Table 3.3, this artefact would fall under artefact type 7 where the data collected in this study was anonymized, cleaned and formatted for use for analytics. Future LA researchers may use this dataset as part of testing for new techniques or algorithms to improve student performance analysis.

3.4.4. Demonstration

This activity involves application of the artefact to solve the problem. This stage can be perceived as part of the testing of the artefact to determine how well it solves the problem. Demonstration can involve the use of case studies, experiments, simulations or any other activity (Peffer et al., 2007).

The implementation or application of the artefact, together with the techniques used to anonymize data as well as the algorithms used for prediction are covered in Chapters 4, 5 and 6. In Chapter 4, the data collection process is described as well as what techniques are used to effectively anonymize, clean and prepare the data for prediction or classification. Chapter 5 and Chapter 6 cover the demonstration of the Decision Tree, Random Forest, Genetic and Optimized Forest algorithms and how these algorithms performed when applied to the UKZN ISTN dataset.

3.4.5. Evaluation

This activity involves measuring how well or to what extent the artefact solves the problem. This can be accomplished by comparing the objectives of the study against observed results from the previous activity (Demonstration). This activity requires knowledge of evaluation techniques and relevant metrics. Evaluation can be in the form of a document outlining how the objectives have been met, quantitative evaluations such as statistics and graphs, or quantifiable measures such as response times.

In the case of this study, two forms of evaluation were considered. Firstly, the prediction models developed by the process model was tested against unseen data instances, and these evaluations are discussed in Chapter 5 and Chapter 6. Secondly, Chapter 7 covers a comparison of the best performance measure values obtained in this study against performance measure values reported in other learning analytics or electronic data mining studies identified in the literature.

3.4.6. Communication

The final activity of the DSRM is to convey the details of the artefact and its importance. These details include the results of demonstration and evaluation, and the benefits of the artefact to researchers and other relevant stakeholders. This dissertation, in addition to two research outputs (see page iii), formed the communication medium for the artefact, its importance and relevance to

the use and application of LA in higher education, how the study is carried out resulting in the artefact development, and the application of the artefact to the problem domain and its resultant ability to predict student performance.

It should be noted that the activities of the DSRM may not necessarily be executed in the order specified. The activities, if needed, could iterate between development, demonstration, evaluation and communication (Peppers et al., 2007). This observation is also in line with other authors that feel that design science follows a pragmatic approach, such as Hevner and Chatterjee (2010); and from an LA perspective, Gibson and Lang (2018) have stated that a pragmatic approach is ideal for an LA study.

3.5. Research model

This section provides the details of the artefact developed that addresses the research objectives of the study. Firstly, in section 3.5.1, an overview of previous frameworks and models related to learning analytics is covered. Section 3.5.2 discusses the process model developed for this study.

3.5.1. An overview of learning analytics models and frameworks

Sections 3.5.1.1 to 3.5.1.5 describes different LA methodologies and frameworks that have been identified in the literature. Section 3.5.1.6 describes the LA model that has been adopted for this study.

3.5.1.1. Five stage LA model (Campbell, DeBlois & Oblinger, 2007)

This LA model suggests five (5) stages for developing a model for generating information for very large data sets. This model is shown in Figure 3.5.

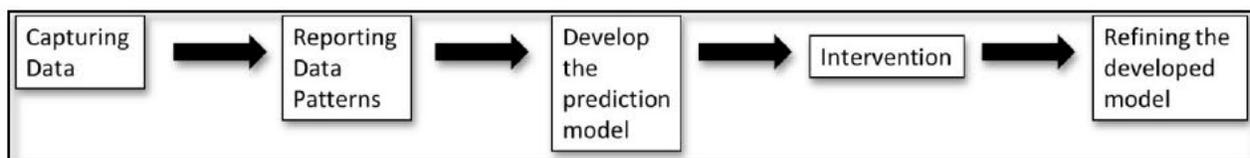


Figure 3.5: LA model suggested by Campbell et al. (2007)

The first stage of the model, Capturing Data, involves the selection of available data sources and capturing of data relevant to the LA initiative being performed. In the second stage, statistical and analytical techniques are applied to the data obtained from stage one in order to produce

information. In the case of LA, this information relates to student academic performance with the intention to identify student weaknesses and provide feedback to students and relevant stakeholders (Haggag et al., 2018). For stage three, a model can be defined using a researcher's preferred technique. The model may incorporate a number of variables or data attributes from the data set and each variable can be associated with weightings based on its importance to predicting the target value. This model can then be tested to determine its accuracy of how the model predicts the target value (Haggag et al., 2018). The intervention stage (stage four) involves the teaching staff taking action with students that have been predicted to struggle and intervening by changing their learning habits (Haggag et al., 2018). Finally, stage 5 (refining the model) involves assessing all aspects of the LA initiative and seeking improvements for future iterations. This includes looking at the data sources and how data is represented and captured, evaluating the techniques used for information generation, and identifying improvements to the prediction model based on importance of variables/attributes (Haggag et al., 2018).

3.5.1.2. Sequence model for learning analytics (Mahzoon et al., 2018)

In the study by Mahzoon et al. (2018), each student is represented as a sequence of nodes (see Figure 3.6). The initial node contains the student's demographic data which includes their age, gender, and employment status, amongst others. The subsequent nodes respectively represent the set of activities performed by a student during a semester. The number of subsequent nodes in this case is the total number of semesters that a student is part of. The final node is an outcome node that contains the final status of the student (such as graduated, inactive or withdrawing, as well as the date of this outcome). Mahzoon et al. (2018) identified the benefits of this model as being a greater focus on time-based events, separation of events (in this case by semester), as well as improved story telling.

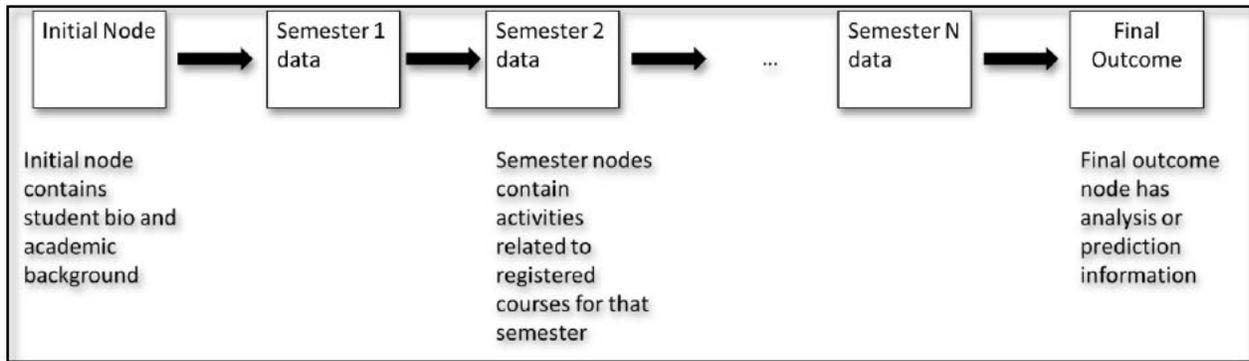


Figure 3.6: Sequence model for student representation in LA project proposed by (Mahzoon et al., 2018)

3.5.1.3. Learning analytics cycle

The LA cycle is an iterative, cyclical process that is made up of six (6) steps, illustrated in Figure 3.7. Explanations of the LA cycle have also varied, with some authors such as Clow (2012) only using four (4) steps, these being the student (learner), data, metrics and intervention. The cycle in Figure 3.7 expands upon the steps to include learning activities that students participate in, resulting in the collection of data. This data is processed, stored and analyzed. The resultant analysis can be visualized for easier understanding, allowing for action to be taken, further resulting in more learning activities (Chatti & Muslim, 2019).

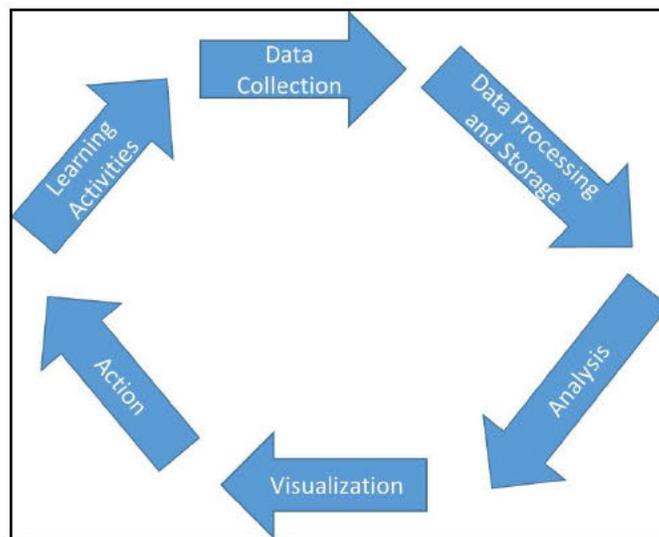


Figure 3.7: LA Cycle adapted from Chatti and Muslim (2019)

3.5.1.4. Learning analytics model by Siemens (2013)

The LA model proposed by Siemens (2013) consists of seven (7) components, these being collection and acquisition, storage, cleaning, integration, analysis, representation and visualization, and action. The model is presented in Figure 3.8.

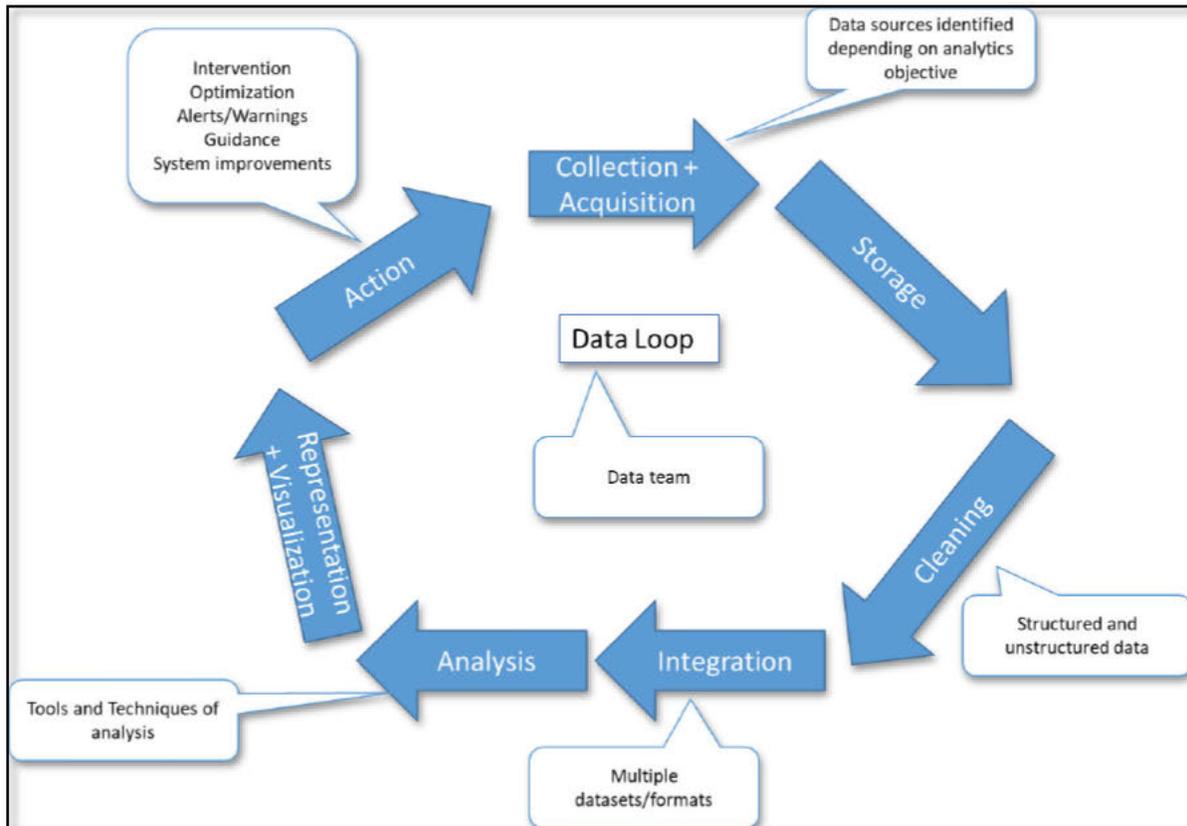


Figure 3.8: Learning analytics model by Siemens (2013)

Siemens (2013) describes the model as providing a system wide approach to analytics. The model indicates processes, as with other learning analytics process models. However, where necessary, interventions and resources are specified in order to assist with implementation (Siemens, 2013). The first stage usually begins with data collection and acquisition, where data sources are identified depending on the objective of the learning analytics initiative. The objectives could be related to marketing, learning, administration or research amongst others. During the cleaning stage, structured and unstructured data need to be considered, while the integration stage involves the consideration of multiple data formats. The stage of analysis involves the application of various tools and techniques such as concept development, prediction and risk determination. The final

stage of action involves the relevant administration or teaching staff to implement changes as per the interpretation of the analysis and visualization stages. This may include interventions, resource management, optimizations and system improvement, amongst others. The final component is the inclusion of the data loop which incorporates the team of individuals that may be involved in the learning analytics initiative (Siemens, 2013).

3.5.1.5. Learning analytics models and frameworks that focus on conceptual and physical implementation

While this is not the focus of this study, for the sake of completeness and an appreciation of LA-based frameworks and models, this section describes the most commonly identified models that are aimed at conceptual and physical implementation of learning analytics at academic institutions.

The reference model proposed by Chatti et al. (2012) is aimed at providing a classification schema of LA initiatives. The reference model is made up of four dimensions covering data source requirements (what?), individuals and stakeholders (who?), purpose and objective of LA implementation (why?), and finally, the techniques required to implement the LA project (how?).

A similar framework is proposed by Greller and Drachsler (2012) but also includes two additional dimensions, these being internal and external limitations. These limitations are aspects that may affect how the LA initiative is carried out and the scope of its functions (Greller & Drachsler, 2012). The external limitations refer to the aspects around the environment that may affect the implementation, such as the ethical and privacy aspects of storing, using and disseminating digital data, as well as local and international laws regarding data regulation. In terms of the internal limitations, the authors identify human-related factors within the organization, for example the competency of individuals and their ability to use the LA system as well as interpret the results produced by the system. Competency can also affect acceptance of the system. A poor understanding of LA could result in users rejecting the initiative. Greller and Drachsler (2012) emphasize the importance of behavioural science and propose an updated Technology Acceptance Model (TAM) that should be used to evaluate the use and acceptance of LA initiatives.

The original ROMA framework was used to assist in strategic and policy development in the field of international development. This framework was adapted by Ferguson et al. (2014) to assist in

the implementation of LA based initiatives. The framework consists of seven steps that include: defining a clear set of policy objectives; identifying barriers to LA implementation; identifying stakeholders; identifying and understanding the purpose behind the LA initiatives; developing strategies for meeting LA requirements; considering capacity and ability of staff; and developing and evaluating the LA initiative. Once LA has been implemented, the system must be monitored and adjusted in order to maintain its effectiveness and to improve the system for the future (Ferguson et al., 2014).

The Let's Talk Framework was introduced by West et al. (2016) with the objective of providing guidance to institutions of higher education with regard to implementation of an LA initiative. The first aspect (domain) of the framework is to provide the institutional parameters that will dictate what is feasible or unfeasible for implementation at the institution. Some of these parameters may include location, size or structure of the institution as well as student and staff demographics (West et al., 2016). The remaining five domains are transitional institutional elements (culture, size, demographic and strategy considerations), LA infrastructure (technology and expertise), transitional retention elements (LA effect on current institutional policies), LA for student retention discussion, and intervention and reflection.

The objective of the Personalization and Learning Analytics (PERLA) framework is to guide the development of an effective LA system that is capable of determining effective indicators for personalization learning (Chatti & Muslim, 2019). The framework is made up of two layers with the inner layer based on the LA reference model discussed above and an outer layer representing the process of identifying indicators for personalized learning.

3.5.1.6. Learning analytics model adopted for this study

The LA model shown below is influenced by the model proposed by Siemens (2013) described in Section 3.5.1.4. The model adopted for this study is illustrated in Figure 3.9 below:

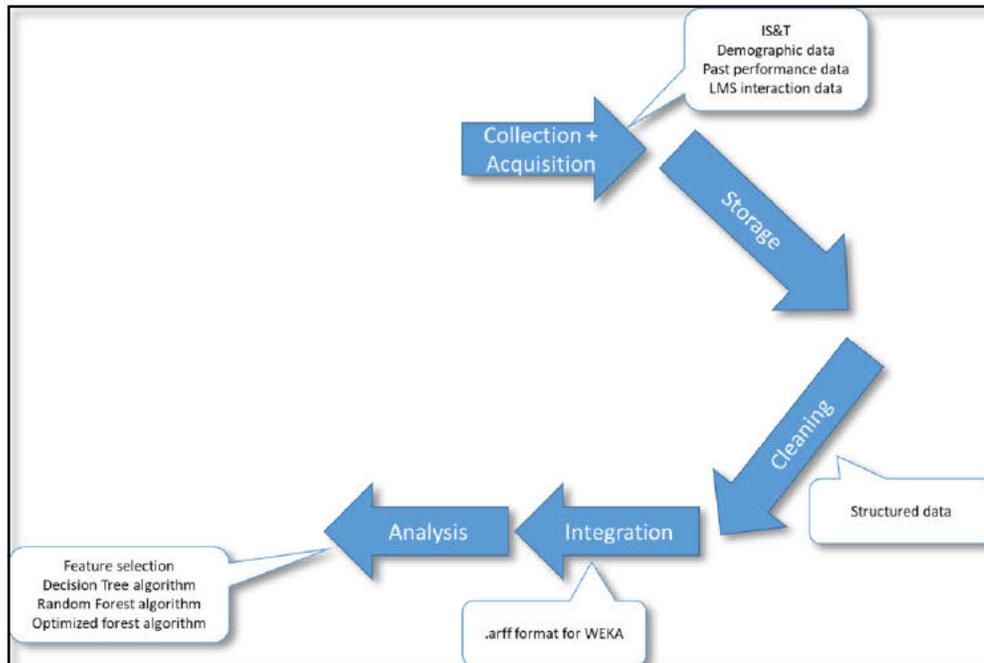


Figure 3.9: Proposed LA model for the study

As shown in Figure 3.9, the model proposed in this study covers five similar stages outlined in the model proposed by Siemens (2013). As the dissertation forms an individual effort, the Data team aspect is removed. The collection and acquisition stages, as well as the storage stages (discussed in Section 3.5.1.4) involve the attainment and storage of IS&T student data. The data is then cleaned and integrated in anticipation for analysis.

As stated in Section 2.9, Africa is still fairly new to the Learning Analytics and a thorough understanding of data collection, acquisition, cleaning and analysis must be undertaken before addressing the stages of visualization and undertaking student intervention strategies. Thus, the areas of visualization is addressed as future work (see Section 8.7) while a discussion related to using LA for student monitoring and intervention is discussed in Section 8.4.

3.5.2. Process model used in this study

For this study, the developed artefact is in the form of a process model outlining the steps involved for an LA approach to analyzing and predicting student academic performance at UKZN, i.e. predictive analytics. The process model is presented in the form of an adapted data flow diagram. The symbolic notation for the process model is described in Figure 3.10.

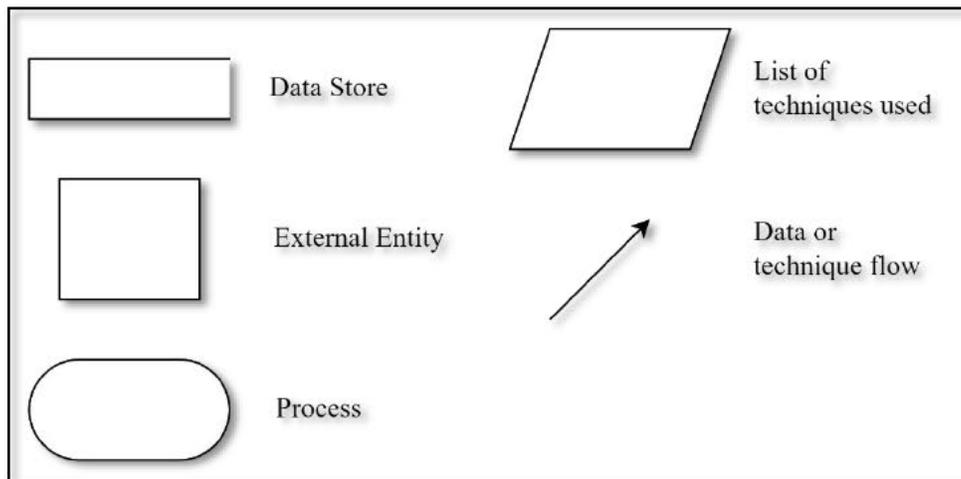


Figure 3.10: Symbolic notation for process model

With the exception of the parallelogram, the data store, external entity, process and data flow are the standard symbols used for a dataflow diagram (Satzinger, Jackson & Burd, 2015). The data store symbol represents a data file or database that stores information regarding a data entity. The rectangle symbol represents an external entity, which is an individual or entity outside the system that provides data into the system. The curved rectangle represents a process within the data flow diagram, when data input is processed resulting in an output from that process. Processes are given a name, usually in a verb-noun form and are associated with a process number. This number is used for reference purposes and does not necessarily indicate the order of the processes (Satzinger et al., 2015).

Differing from the standard dataflow diagram symbols is the parallelogram where, in the context of this process model, the symbol indicates techniques that are used as part of a process in order to produce output from a given input. Figure 3.11 shows the process model developed in this study.

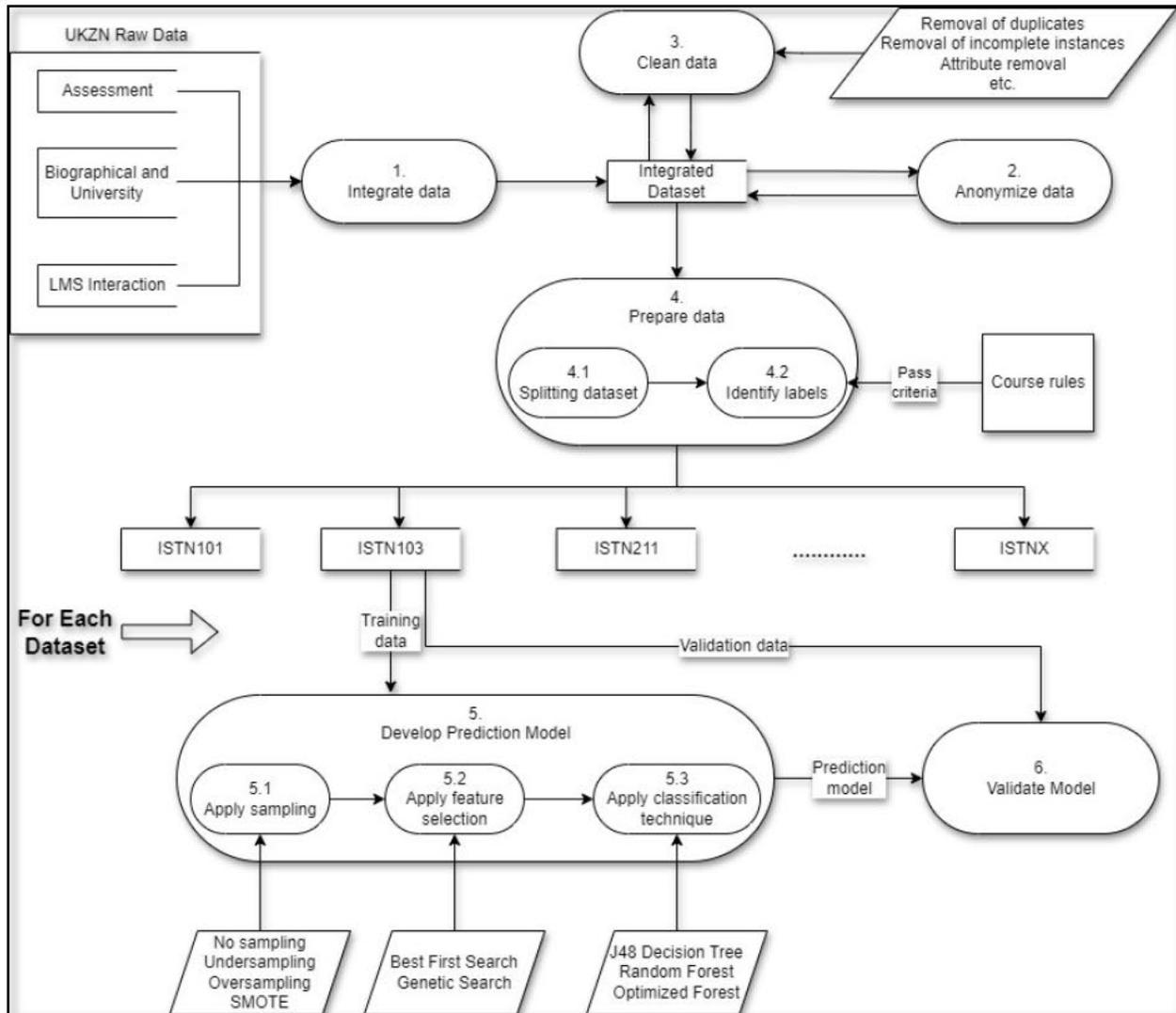


Figure 3.11: Process model developed in this study

The process begins with the acquisition of data. The process of acquiring the assessment data, biographical and university data and LMS interaction data in its raw form is explained in section 3.6. Thereafter, the data integration process (process 1) is performed where the three datasets are combined into a single dataset.

The integrated dataset then undergoes the process of data anonymization (process 2) and cleaning (process 3). The cleaned and anonymized integrated dataset is then prepared for prediction (process 4) where the researcher chooses the available courses resulting in the splitting of the integrated dataset into respective courses identified. In addition, the pass criteria are supplied for

the creation of labels of the classes for each of the split datasets (in this case, pass or fail). The data anonymization, cleaning and preparation processes are discussed in Chapter 4. Process 4 results in the integrated dataset being split into individual course datasets.

Each course dataset is split into a training data-subset (all data before 2021) and a validation data-subset (2021 data) with the training dataset being input for process 5 (Develop Prediction Model) and the validation dataset being input for process 6 (Validate model). The training dataset undergoes the processes of sampling (process 5.1 that is covered in Chapter 5) and feature selection (process 5.2 that is covered in Chapters 5 and 6). Process 5.3 covers the application of classification technique(s), discussed in Chapter 5 and Chapter 6 that are applied to the sampled (or non-sampled) dataset resulting in the development of a prediction model. This prediction model serves as input to process 6 which involves the application of the model to the validation data-subset. Here, the objective is to compare the accuracy obtained during process 5 with the accuracy of the resultant prediction model when applied to an unseen dataset.

This model is similar to those described in sections 3.5.1.1 to 3.5.1.5 in that all the models have similar processes (such as data collection, pre-processing and analysis). The model in this study, through the use of dataflow diagram (level-0) notation, shows how the data is transformed between processes. Furthermore, the model in this study includes a notation to specify which techniques are applied within each process. Level-1 dataflow diagrams were not considered due to the potential increase in complexity of the diagram through the addition of further processes and dataflows.

3.6. Data collection

This section outlines the steps taken to collect data required for the study. Before any study relating to data use can be conducted at UKZN, ethical clearance must first be granted. This aspect is covered in section 3.6.1. Section 3.6.2 describes the final datasets used for the study. Section 3.6.3 discusses the data validity and reliability.

3.6.1. Ethical clearance

Ethical clearance (EC) was applied for with the intention of acquiring student biographical and past academic data as well as Moodle LMS interaction data from the UKZN Institutional Intelligence (II) division. As part of the EC process, a gatekeeper letter was obtained from the UKZN registrar. The ethical clearance letter is included in Appendix A.

The data was approved to be released by both II and the registrar on condition that the data was anonymized in accordance with the POPI act introduced in South Africa. To facilitate the acquisition of the data as well as meeting the requirements specified by POPIA, a non-disclosure agreement (NDA) was signed between the researcher and UKZN. This NDA stated that the data would be anonymized and only available to authorized individuals involved in the study at UKZN.

3.6.2. Data used for the study

The data obtained for this research is secondary data in the form of a dataset of UKZN students from the discipline of IS&T. The data for the study was initially made up of three datasets. The first dataset contains student demographics, registration data and academic performance data. The dataset (biographical, registration and assessment data) initially contained data related to 50 courses and approximately 14000 students registered to ISTN courses from 2014 to 2021. The second dataset contains student marks obtained in high school. Both of these datasets were provided in MS-Excel format by the UKZN II division.

The final dataset consists of student Moodle LMS interactions with the different IS&T courses that they were registered for. Access to each Moodle site was obtained by permission of the respective course co-ordinators and was manually downloaded by going to the moodle site and downloading the data (in MS-Excel format). The Moodle interaction data was obtained from Moodle IS&T course sites from 2017 through to 2021. Previous LMS course interaction data were no longer available for download. The description of the datasets as well as the process of data anonymization, integration, cleaning and preparation is described in detail in the next chapter.

3.6.3. Reliability and validity

Thanasegaran (2009) defines reliability as the degree to which measures are free from error and produce consistent results such that the results can be repeated or duplicated. Validity is defined

as the degree or the extent to which a method measures what it is meant to be measuring (Thanasegaran, 2009).

From the perspective of the study, the reliability of the data was addressed through the process of data cleaning and preparation (described in Chapter 4). Further, the predictive validity of the model is determined by dividing the dataset for each course into two parts (George, Osinga, Lavie & Scott, 2016). The first part is the training set and the second part is the validation dataset which the learning algorithm has not interacted with. K-fold cross validation was used as part of the development of the predictive model and this model was then applied to the validation dataset to assess the capability of the model in making predictions on unseen data. Validity is confirmed if the accuracy obtained by the model via the training dataset is equivalent when compared to the accuracy of the model applied to the validation dataset (George et al., 2016). In addition, the validity of the predictive models is also determined using a set of commonly identified performance metrics. These metrics are described in section 5.3.5.

3.7. Tools used for data analysis

The raw datasets were integrated by linking each datafile based on the student number or in the case of the Moodle LMS interaction data, the name of the students. The integration of the datasets was performed using Microsoft Excel.

Data anonymization was performed using functions available in Microsoft Excel. From the researcher's perspective, this software was available and the researcher was familiar with the functions and process required to anonymize the data using this software. Further details related to anonymization of the data is covered in Chapter 4. Cleaning of the data was also performed using Microsoft Excel.

Data preparation was performed using the filter function in MS-Excel to separate the dataset based on the different ISTN courses. Labelling of data attribute values were also performed using MS-Excel.

The WEKA (Waikato Environment for Knowledge Analysis) data mining tool was used to perform data analysis and prediction. As stated in Section 2.7.1, WEKA is a Big Data analytical tool that has several pre-processing functions and machine learning algorithms to assist users in conducting analysis and prediction. It is also one of the most commonly used applications for conducting LA studies. The prediction models are then tested using unseen data instances using WEKA.

3.8. Chapter summary

Methodology refers to the theoretical assumptions and principles that underpin a particular research approach. It guides a research study on how to state the research questions as well as what processes and/or methods to use. The details of the research methodology are important as they provide transparency to all the facets of the research being conducted. From the perspective of this research study, this chapter covered the aspects of the methodology applied. The study follows the Design Science research methodology that will result in the development of an artefact. This artefact is in the form of a process model that was introduced in Section 3.5.2. The chapter also covered the data collection approach; in this case, the collection of data relating to student demographics, academic performance and LMS interaction. The following Chapters 4 through to 7 describe the execution of the described methodology with the next chapter covering the first aspect of data preparation.

Chapter 4 – Dataset preparation

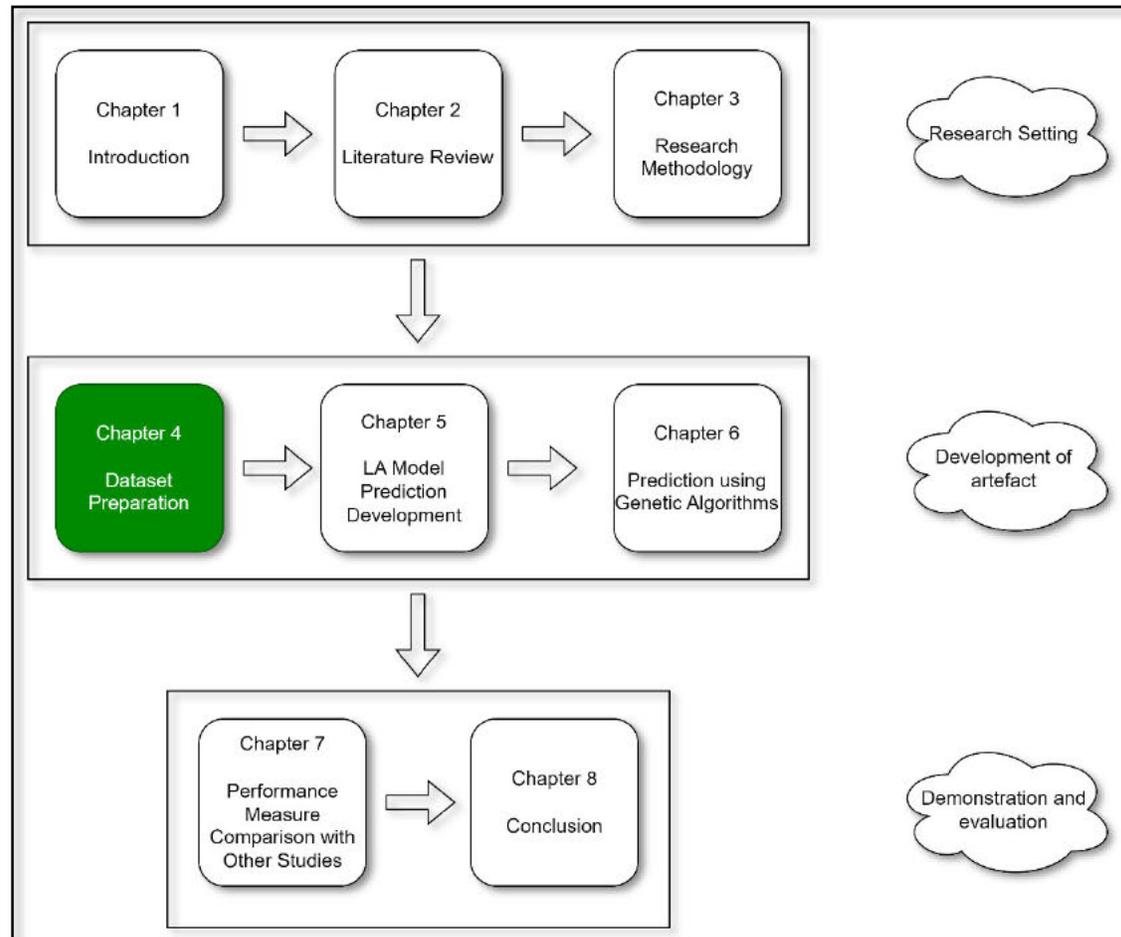


Figure 4.1: Thesis structure

4.1. Introduction

This is the first chapter of the thesis that addresses the solving of the research questions and objectives through the development of the artefact (see Figure 4.1).

The initial step(s) of all LA initiatives involves the collection and preparation of data (Chatti et al., 2012; Munk et al., 2017; Romero et al., 2014). Depending on the quantity, quality and presentation of the data, this can be a complex task to complete, as was determined in several studies on Big

Data and LA (Adejo & Connolly, 2017b; Oussous et al., 2018). In this chapter, the focus is on addressing the first two research questions/objectives of the study, these being:

RQ1 - How can the data from the relevant data sources (SMS, Moodle logs, registers etc.) be integrated?

RQ2 - How can the integrated data be organized in preparation for data analysis?

From the perspective of the DSRM described in Chapter 3, this chapter focuses on the design and development of the artefact with a focus on the acquisition of data and the preparation of this data for analysis and/or prediction. Section 4.2 covers the process from data acquisition to the preparation process. A breakdown of Section 4.2 is depicted in Figure 4.2 below:

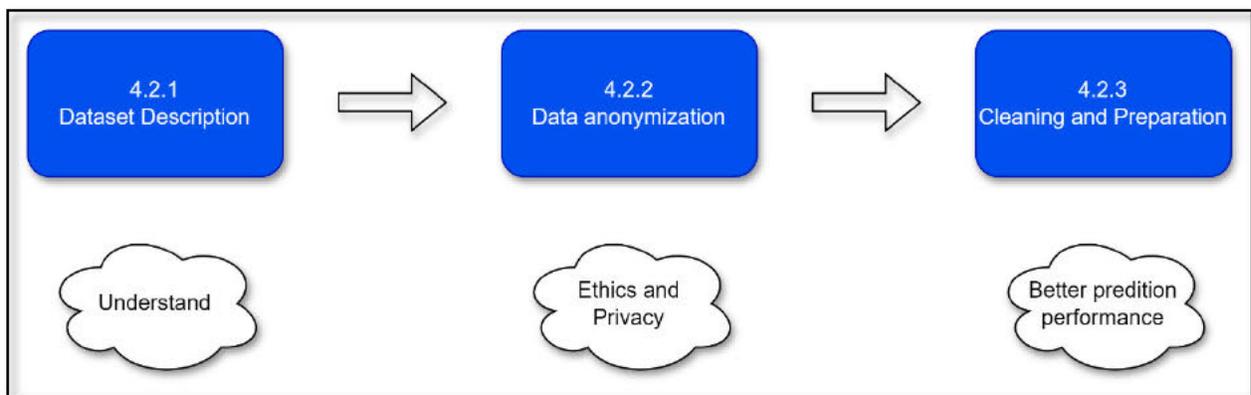


Figure 4.2: Map for Section 4.2 coverage

4.2. From data acquisition to preparation

For this research, data was obtained related to students from the University of KwaZulu-Natal in the form of three datasets. This first dataset consists of students that registered for any IS&T courses from 2014 and includes aspects of their biographical and university related information, and class mark, exam mark and final mark for the IS&T courses that they had registered for. A second, related dataset is the high school results that were submitted as part of the student's application to University. The third dataset is made up of a number of datafiles, with each file consisting of all the interactions made by students on the Moodle LMS site for the individual IS&T course(s) that they were registered for (from 2017-2021). This includes the Moodle logs for any IS&T course site that could be acquired as well as, where possible, activity completion reports

listing the different activities on the Moodle site and whether or not the student completed these activities.

4.2.1. Description of datasets

The demographics dataset and high school dataset are initially separate data files with a common student number attribute. Where available, each IS&T Moodle course was accessed and the log files as well as activity completion reports were obtained. This is represented in Figure 4.3. Four datasets were provided for this study and are named in Figure 4.3 as DS1, DS2, DS3 and DS4. As was discussed in the introductory chapter, there is a distinct separation between these files with the only common attribute being the student number.

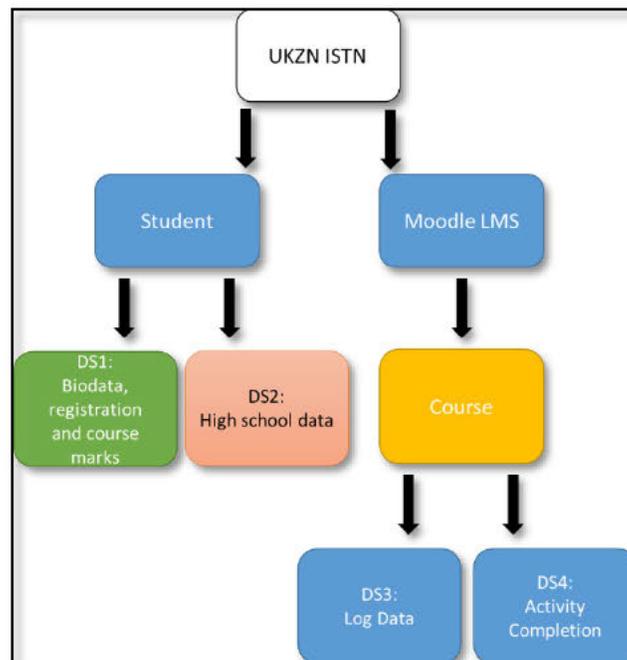


Figure 4.3: Initial dataset hierarchy structure

Each of the datasets are described in sections 4.2.1.1 (DS1), 4.2.1.2 (DS2) and 4.2.1.3 (DS3 and DS4).

4.2.1.1. Biographical and registration dataset (DS1)

This first dataset consists of student biographical, registration, and course marks data (DS1). The purpose of using this dataset is to understand if the biographical, registration and assessment

attributes play a role in prediction of student academic performance. In section 2.3, student demographics were commonly identified as attributes tested in LA or EDM studies in the past.

Initially, the dataset was provided in MS-Excel format, consisting of 44106 rows with each row representing a student instance of their biographical data, registration for a course and the marks achieved for that course. The dataset contains instances from registrations for 50 IS&T courses. These courses are from all levels of undergraduate (1st, 2nd and 3rd year of study) and honours study. The Excel file provided was composed of the following attributes (features) listed in Table 4.1.

Table 4.1: Attribute description for dataset DS1

Attribute	Description	Data range
YEAR	Student year of registration for course.	2014...2021
BC	Student semester of registration for course.	1 – Semester 1 2 – Semester 2 0 – Both semesters
OT, OTDESC and CAMP	Campus abbreviation, campus name and campus code.	HA – Howard College (1) PA – Pietermaritzburg (2) WA – Westville (4)
COLL and COLLEGE	College code and name.	24 – College of Law and Man Studies
DEPT and DEPTNAME	Department code and name.	2484 – School of Man Info Tech & Gov
QUAL and QUALDESC	Qualification that the student was doing when registered for the course.	
UGPG	Whether the student falls under undergraduate and postgraduate.	UG – Undergraduate PG – Postgraduate
SELFFUNDED	Whether the course is self-funded or not.	Y – Yes N – No
Continued on next page...		

Table 4.1 continued

Attribute	Description	Data range
STNO	Unique student number created to distinguish one student from another. Required for associating a student record to their respective Moodle activity that was separately collected.	
BIRTHDATE	Student date of birth in YYYY/MM/DD format .	
GENDER	Gender of student.	M – Male F – Female
RACE	The ethnic group that the student falls under.	A – African C – Coloured I – Indian O – Other W – White
ALIENYN	This attribute indicates whether or not the student is from South Africa or not.	Y – Yes N – No P – Permanent Residence
RELIGIONDESC	What religion the student falls under.	
COUNTRYCITZCODE AND COUNTRYCITZDESC	The allocated code and country of origin.	
COUNTRYPERMCODE AND COUNTRYPERMDESC	Permanent residence code and country of student.	
HOOMELANGCODE AND HOMELANGDESC	Code and language spoken at home as specified in initial application.	
MARITALSTATUS	Whether the student is married, single, divorced or widowed.	S – Single M – Married D – Divorced W – Widowed
Continued on next page...		

Table 4.1 continued

Attribute	Description	Data range
MATRICTYPE, MATRICDATE, MATRICPOINTS, MATRICRANGE	Matric type, date of matric, points achieved, and range	
QUINTILE	South African government schools are placed into one of five quintiles, mainly for the purpose of financial resource allocation	1...5 (1 being poorest quintile and 5 being least poor quintile) NA – Not applicable (in the case of private or overseas schools)
SECONDARYSCHOOLCODE, SECONDARYSCHOOL	Name of school that the student attended	
ADDRPCODE and AREA	Code and name of area of residence	
RESYN	Whether the student is in University or private accommodation (residence)	Y – Yes N – No
RESBLDNAME and RESBLDOWNER	If RESYN is Y, then this indicates the residence code, residence name and owner	
BURSARYYN	Whether the student has a bursary or not	Y – Yes N – No
COUNCILLOANYN	Whether the student has been given a loan by some council	Y – Yes N – No
HIGHERDEGREEREMISSIONYN	Whether the student has been provided with remission of fees	Y – Yes N – No
NSFASBURSARYYN	Whether the student has been given a bursary through NSFAS	Y – Yes N – No
NSFASLOANYN	Whether the student has been given a loan through NSFAS	Y – Yes N – No

Continued on next page...

Table 4.1 continued

Attribute	Description	Data range
SCHOLARSHIP	Whether the student has been given a scholarship	Y – Yes N – No
UNCATEGORIZEDYN	Whether the student has received some other form of funding	Y – Yes N – No
FUNDINGTOTALPAID	Total amount of funding that has been paid for the given year	
SUBJ and SUBJDESC	The course code and name that the student is registered for that year	
M_YMARK, M_EMARK, M_FMARK, M_ERES	Year (class) mark, exam mark, final mark and result of the course for that student for that year.	
SUBJREGDATE	When the student registered for the course in YYYY/MM/DD format	
SUBJCANCDATE, SUBJCANCREASON, SUBJCANCREASOND	When the student cancelled their registration for the course and the reason for the cancellation (if applicable otherwise blank)	
EXEMPTYN	Whether the student was given an exemption from doing the course	Y – Yes N – No
WEBREGYN	Whether the student registered via the online registration system or not	Y – Yes N – No
SUPPREG	Whether the student registered for a supplementary exam	0 - No 1 – Yes
S_YMARK, S_EMARK, S_FMARK, S_ERES, SUBJFMARK, SUBJERES	Related to Supplementary exams. Year (class) mark, exam mark, final mark and result of course for that student for that year.	

4.2.1.2. Dataset consisting of high school marks (DS2)

A separate file was provided containing the student high school data (DS2). As described in Section 2.3, high school marks were identified by studies in the literature as factors that could contribute to performance prediction. This MS-Excel file consisted of 112 773 rows with each row indicating the student and a high school subject that they did as well as the marks and/or grades achieved. For this file, the following attributes were included (Table 4.2):

Table 4.2: Attribute description for dataset DS2

Attribute	Description
STNO	Unique student number created to distinguish one student from another. Required for associating a student record to their respective Moodle activity that was separately collected.
SUBJECT and SUBJDESC	Matric subject code and name
GR11GRADE, GR11PERC, GR11SYMBOL	Level of subject (higher or standard), percentage obtained, symbol obtained for that grade 11 subject
TRGRADE, TRPERC, TRSYMBOL	Level of subject (higher or standard), percentage obtained, symbol obtained for matric trial examination for that subject
MATRICGRADE, MATRICPERC, MATRICSYMBOL	Level of subject (higher or standard), percentage obtained, symbol obtained for matric trial examination for that subject

4.2.1.3. Moodle LMS course data and activity completion datasets (DS3 and DS4)

The datasets extracted from the Moodle LMS are the activity logs (DS3) and activity completion reports (DS4). As noted in Section 2.3, student learning activities and participation is seen as an important factor in student academic prediction. The attributes for the log files are listed in Table 4.3 and the activity completion attributes are listed in Table 4.4. With respect to Moodle logs, not all log data were available for all courses for all years. The only complete log data that were available on the UKZN servers were from 2017 to 2021. In terms of the activity completion reports, only 2020 and 2021 reports were used where available as the activity completion report feature was only implemented in the UKZN Moodle LMS from 2020.

Table 4.3: Attributes and descriptions for dataset DS3

Attribute	Description
Time	Date and time of activity
User full name	Name of the individual performing the action
Affected user	Individual affected by action (if any)
Event context	Relates to an event that the student is participating in
Component	Category that the event context falls under
Event name	Name of the action that has occurred
Description	Description of the activity performed
Origin	Where the action originated from
IP Address	IP address where the action occurred

Table 4.4: Attributes and descriptions for dataset DS4

Attribute	Description
Student name (column not labelled in file)	Name of individual
Username (STNO)	Individual (Staff or student) ID given by university during registration
ID number (STNO)	Individual (Staff or student) ID given by university during registration
Email address	University allocated email address
Institution	Campus that student resides in
Activity 1 ... N (multiple attributes)	Set of activities specified in LMS course site and whether these are completed or not
Activity completion date/time	If the activity is completed, date and time of completion

4.2.2. Data anonymization

The first step for the process of anonymization was to de-identify the data as required by POPIA (discussed in Section 2.4). The datasets DS1 and DS2 supplied by the UKZN Institutional Intelligence department (II) did not contain any names to identify students but did contain the student's university allocated student ID. This student number is the common attribute that is used to uniquely identify a student's biographical data, the academic performance data (high school marks and ISTN course marks) as well as the Moodle course activity completion reports. While this provides some form of anonymity, an individual with access to the UKZN Student Management System (SMS) would still be able to view a student number in the dataset and access the SMS to get the student details. Thus, the student number needed to be replaced with an alternative unique reference to ensure anonymity. The technique of pseudonymization via hashing, as specified by Khalil and Ebner (2016), was used where the student number is replaced into a special key value. Algorithm 4.1 was followed to anonymize all UKZN student numbers (attribute named STNO) in the student biographical and course registration dataset:

Input: Dataset DS1	
Output: Anonymized Dataset DS1	
1	Copy the STNO column values to another column
2	Remove all duplicate values, i.e. multiple instances of the same student record in a file
3	On a new column, starting with STUD0000001, copy the values down such that the number increments continuously to the end of the student number list, i.e. STUD0000002, STUD0000003, STUD0000004, etc.
4	Replace the student number value in the STNO column with the corresponding STUD number.

Algorithm 4.1: Anonymization algorithm for UKZN student numbers

The replacement student number (STUD number) must also replace the corresponding student numbers in the DS2 and DS4 datasets. Once this was accomplished for all datasets, the original student numbers were removed. This ensured that the actual student numbers were not available for any individuals accessing this dataset. This was in line with the requirements specified by II, the UKZN registrar and the ethical clearance application.

In the case of the Moodle course interaction data (Dataset DS3), the names of the students were associated with the different actions performed within the logs. As part of the anonymization process, these names were linked to their respective UKZN student ID. Once done, the student name attribute was removed and the student number was replaced by the new associated anonymized STUD number.

Further descriptions related to de-identification of the data are described in sections 4.2.3.2. and 4.2.3.3.

4.2.3. Cleaning and preparation of datasets

This section describes aspects required to improve the quality of the datasets in preparation for analysis and prediction. This is an important task required to ensure a better-quality prediction model (Jayaprakash et al., 2014).

4.2.3.1. Handling missing data

This section describes how missing data items in respective attributes were handled. As stated in section 2.5.1, this is an important aspect of the data preparation stage. The attributes that needed to be addressed were only for the DS1 dataset. These attributes as well as the action taken are described in Table 4.5.

Table 4.5: Addressing missing values in different attributes within the dataset

Attribute/Feature	Action taken with justification
RELIGION	It is assumed that where the input is blank, student chose not to divulge this information. Blank values were replaced with the value “Not Specified” as it would be inappropriate to determine or guess these values for any individual students in the dataset.
QUINTILE	The value is left blank in the event that that the specified school does not fall under that of being an ordinary South African public school (such as schools outside South Africa or private funded schools). In this case the blank value is replaced by “NA” (Not applicable).
AREA	Where the area of residence was not specified, a value of “Not specified” was used.
RESBLDNAME AND RESBLDOWNER	For students that were not in residences (i.e. staying in their own homes or private accommodation), a value of “NA” (Not applicable) was given.

4.2.3.2. Additional attributes for analysis and prediction purposes

This section outlines additional attributes that were added to the datasets. These attributes were included for the purposes of de-identification and/or to allow for improved analysis and prediction.

According to Ali et al. (2013), age was a statistically significant factor that affected student academic performance. Therefore, the student age was added as an attribute for analysis and prediction of student performance. The age of a student is determined by subtracting the student’s date of birth from the year that the student registered for the course. Based on the calculated age, the value is assigned based on one of two age classes (values) and is further elaborated upon in section 4.2.3.3.

To determine the role of computer science or technology-based subjects' impact on academic performance, an attribute was created by checking whether students had chosen one or more computer science or technology subjects in school. If a student did do this subject type, a value of Y was allocated, or else a value of N was allocated.

The Moodle LMS used at UKZN keeps track of all user interactions in the form of logs. To better understand if Moodle usage plays a role in student performance, the logs were summarized and resultant attributes were included that counted the total number of interactions performed by each student for the course for that year. This approach was also adopted by Mwalumbwe and Mtebe (2017), who kept track of the number of student logins, items downloaded, peer and forum interactions and exercises performed.

Each Moodle site can also be set up to keep track of specific activities that students have interacted with and completed via Activity Completion reports (DS4). Activities on the course site are marked using a checkbox next to the activity. Each activity can be marked as complete either manually by the student, or automatically based on the student performing a certain task(s). In the case of courses run during the years 2020 and 2021, attributes were created to count the number of activities completed by students based on the activity completion report generated by Moodle for the different IS&T courses. Three attributes were created, these being number of activities recorded as complete, the number of activities recorded as not complete, and the percentage of activities completed.

Courses at 2nd and 3rd year level have pre-requisite requirements that must be met in order to be able to register for that course. These pre-requisite requirements are in the form of previous year courses that a student is required to pass. For the 2nd and 3rd year courses, the pre-requisite course symbols have also been included to determine the role that these courses play in whether the student will pass or fail that course.

4.2.3.3. Data discretization

To assist with analysis and prediction as well as to further the process of data de-identification (Khalil & Ebner, 2016), the student age attribute has been categorized in two groups based on their

age (15-20, 21 and Above). These two groups were chosen as the majority of students' ages were in the ranges of either 15 to 20 or 21 and above.

For attributes related to student marks, the mark range was narrowed down to symbols, thereby reducing the number of possible attribute values from one hundred (100) to five (5). This is shown in Table 4.6.

Table 4.6: Ranges for marks

Mark Range	Symbol
0 – 49	F
50 – 59	D
60 – 69	C
70 – 79	B
80 – 100	A

For the AREA attribute, the initial values indicated specific areas where the student was from, e.g., Northern KwaZulu-Natal, Gauteng Pretoria Tshwane, KwaZulu-Natal Midlands, Eastern Free State, etc. These values were summarized into the main provincial areas e.g. KZN, FS, EC, WC, etc., in order to reduce the number of nominal values for this attribute.

4.2.3.4. Removal of unnecessary attributes and instances

As this study involves students and their interactions, log entries and attributes related to automated system events (for example, automated addition and removal of students from the course) and staff interactions (for example, adding lecture slides and content creation) were removed.

Students that were withdrawn from courses as well as instances where students were given exemptions from courses were removed from the dataset. This was done as students that withdrew from courses did not have complete results or data and student exemptions were just duplicated records for two different years. This strategy was also adopted in previous studies relating to predicting student academic performance, such as Minaei-Bidgoli et al. (2003) as well as Waddington et al. (2016).

From the perspective of the DS1 dataset, attributes where data was found to be duplicated (for example, country of origin and country of permanent residence) as well as attributes that were abbreviations of other attributes (e.g. campus name and campus code) were removed.

When downloaded, the Moodle log file (DS3) contains the time and date of the activity, the name of the individual that performed the activity (or to whom the activity is related to) and the type of activity performed. As there is no student ID included in the downloaded file, there is no possibility to differentiate between students with identical names. For example, if there are two or more students named Andile Dlamini, it is impossible to differentiate within the log file which Andile Dlamini is performing an activity recorded within a log file. Thus, any activities involving students with duplicate names were not considered.

Finally, records related to courses that were no longer offered in the discipline were not considered. These were mainly 3rd year courses that were eventually merged together as part of new university directives.

When working with Activity Completion reports (DS4), a limitation noted was that the activity completion feature will only record the date and time that the requirements were completed for the activity and not the extent to which the student immersed themselves into the activity. For example, Moodle will record a file access activity as complete when the student clicks on the file and it is viewed or downloaded. The LMS cannot, however, determine whether or not the student has actually looked at the document and understood the content within.

4.2.3.5. Data integration

As the WEKA application only accepts a single file, it was necessary to merge the DS1, DS2, DS3 and DS4 datasets. The Moodle logs (DS3) were added to the demographics and performance dataset (DS1) in the form of total clicks made by the student, as well as the number of times the student had interacted with different activities (such as files, folders, quizzes, H5P videos etc.). In terms of DS4, a record was made of the count and percentage of activities completed.

Further to the above, the merged data file was separated into multiple files based on courses (i.e., each file contained instances of registration for each of the courses offered in the IS&T discipline). This was done as each course is run independently of the other. In addition, each course data file was divided into three file variations based on date. The first file variation (VAR1) contained data with no Moodle activity as Moodle activity was only available from 2019 onwards. The second

file variation (VAR2) consisted of only instances that had Moodle data, i.e., data from 2019 onwards. The final file variation (VAR3) only contained 2020 and 2021 data as these courses were taught during the COVID-19 lockdown and all the content was taught online.

WEKA only accepts files in `.arff` or `.csv` formats. Microsoft Excel has a facility to convert data from the standard MS-Excel format (`.xls`) to `.csv` format. Further, WEKA has built in functionality to convert `.csv` files to `.arff` files that the application (WEKA) prefers to use. As WEKA cannot distinguish between commas used to separate attributes and commas within text values (for example: “Durban, KZN”), this needed to be addressed. WEKA has a similar issue with quotation marks and apostrophes.

Once the files have been converted to `.arff` format, the file is ready to be applied to the WEKA application.

4.3. Chapter summary

This chapter addressed the first two research questions, i.e. integrating the data sources into a dataset and then preparing the dataset for analysis and prediction purposes. The data sources were initially identified. As per the requirements set by the UKZN registrar, POPIA and II, the data was then anonymized by creating a new student identifier as well as categorizing certain attributes. The dataset was also cleaned by removing unnecessary attributes, providing values for missing data that was not available or specified, and performing discretization to certain numeric and string attributes. Finally, the Moodle log data (DS3) and activity completion reports (DS4) were summarized, in terms of number of times certain items were accessed and number of activities completed, respectively. These summary attributes were merged with the DS1 and DS2 datasets resulting in each row indicating the student, their demographics, the details regarding their registration for a particular IS&T course (including their performance for that course) as well as a summary of their Moodle interactions for that course (the Moodle interaction details and activities completed where applicable). Now that the data was prepared for analysis and prediction, it could be applied to the selected tools. This is discussed in the next chapter. As a summary, the specification table of the dataset is provided in the Table 4.7.

Table 4.7: Description of UKZN ISTN dataset

Subject Area	Information Systems and Technology
Type of data	Demographics, registration, class and examination marks, LMS interactive data
How data was acquired	For the years 2014 to 2021, the demographic and academic performance data was obtained from UKZN institutional intelligence (II). The Moodle LMS data was obtained by accessing each course (where permission was given) and downloading the log and activity completion files.
Data source location	The data obtained relates to IS&T students at the University of KwaZulu-Natal, Pietermaritzburg and Westville campuses, KwaZulu-Natal, South Africa.
Data Format	Cleaned, preprocessed, and divided into courses. Each course dataset is further divided into training data and validation data.
Data Access	In order to further the development of the LA field in Africa, these datasets will be made publicly available in a Microsoft Excel format.

Figure 4.4 illustrates the portion of the framework (described in section 3.5.2) that was discussed in this chapter.

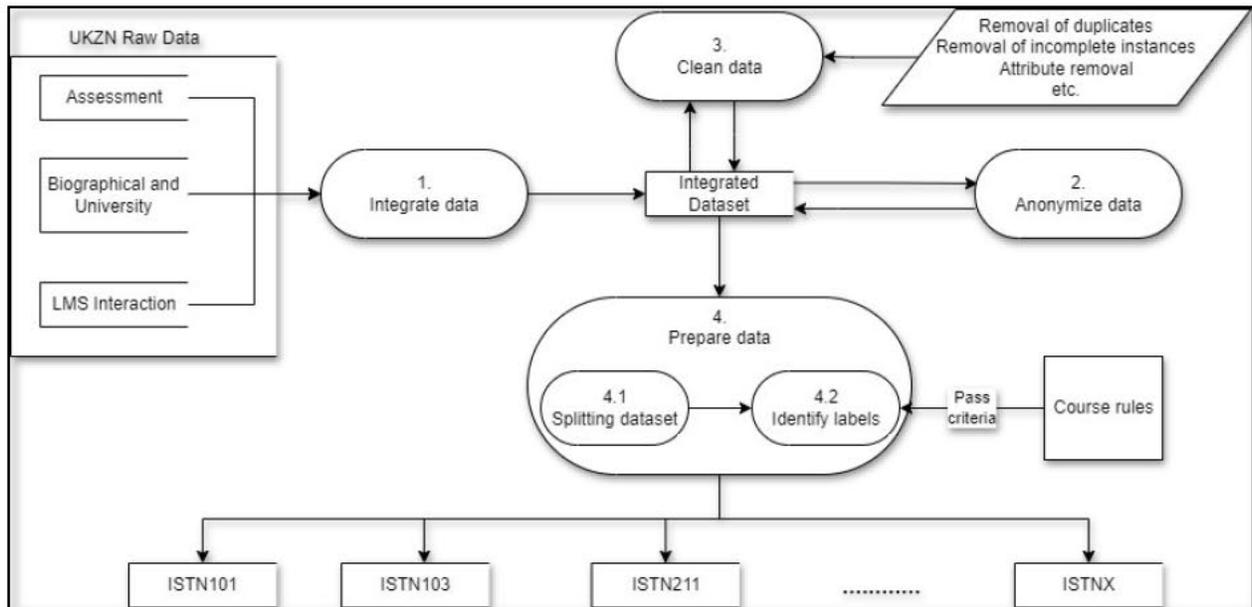


Figure 4.4: Data preparation to integration

Chapter 5 – LA model prediction development

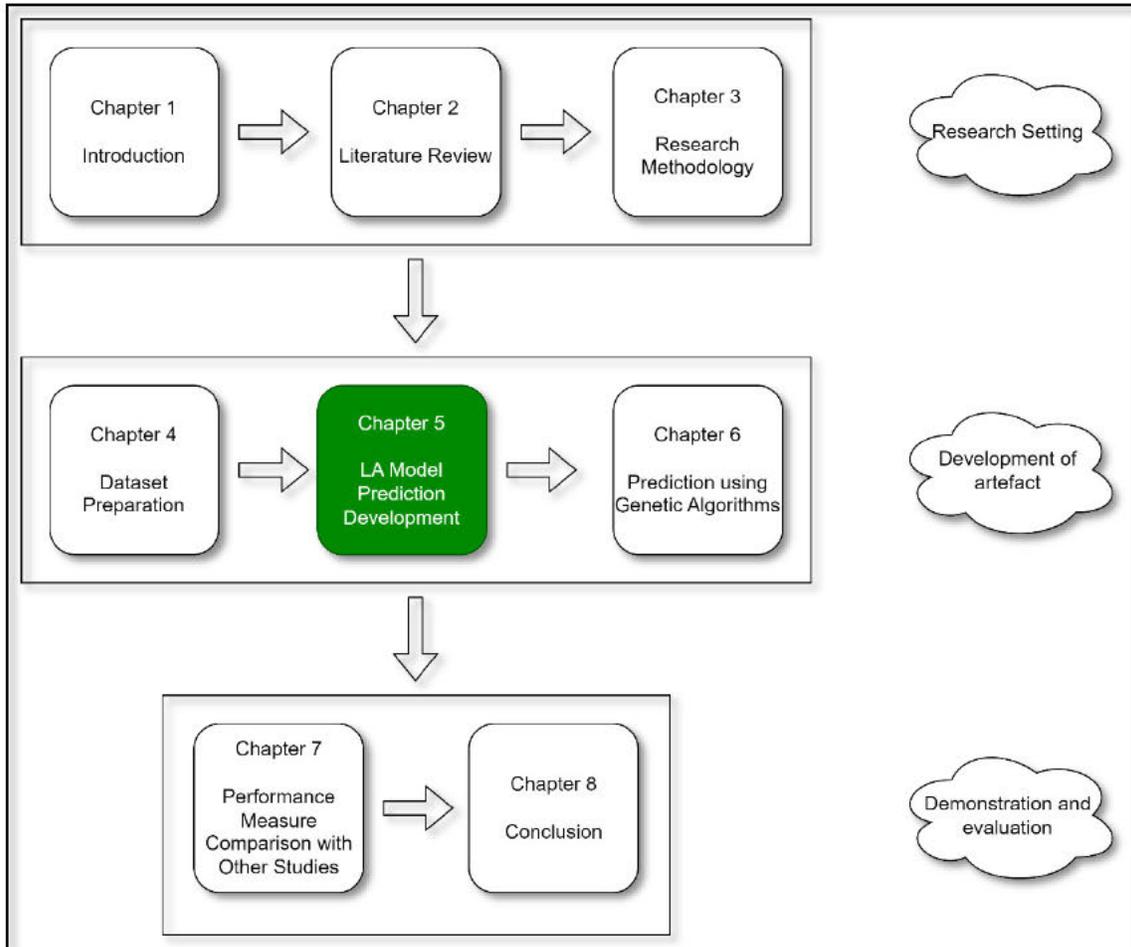


Figure 5.1: Thesis structure

5.1. Introduction

This is the second chapter that form part of the artefact development process (see Figure 5.1). In Chapter 4, the process of data preparation and cleaning was described and discussed, where the separate datasets were amalgamated into a single dataset containing student biographical data, registration data, academic performance for each course registered and counts of LMS interactions. This dataset is now referred to as the UKZN ISTN dataset. The next step is the application of learning algorithms to the dataset in order to predict student performance. Thus, this chapter discusses the process of data analysis and prediction of academic performance. This is identified in research questions 3 and 4, which are listed next:

RQ3 - How can the data be used for identifying learning patterns (training)?

RQ4 - How can the trained data be used to predict student academic performance?

From the perspective of the DSRM described in Chapter 3, this chapter focuses on the design and development of the artefact with a focus on applying learning algorithms to the now prepared dataset with the objective of predicting student performance. The structure of this chapter is as follows and is depicted in Figure 5.2. The entire dataset cannot be used as input into WEKA as each course is different in terms of the outcomes, mode of teaching, type of content and forms of assessment. Thus, as stated when describing the process model in Section 3.5, the integrated dataset is divided into ten (10) course datasets. Section 5.2 describes each of the course datasets and aspects to consider before analysis and prediction. Section 5.3 describes how the data is trained and aspects to consider during training. Section 5.4 covers the results of the experiments conducted for each course dataset and the results of the experiments. Finally, Section 5.5 concludes the chapter with a summary of the findings overall.

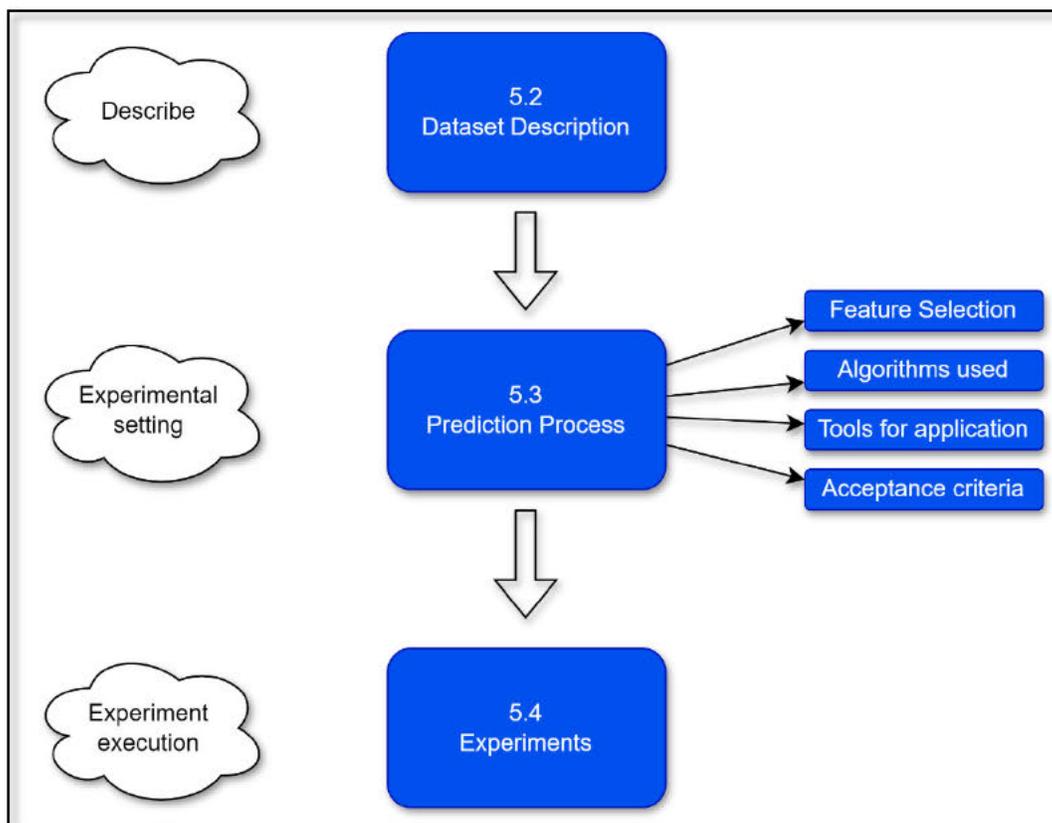


Figure 5.2: Map for Chapter 5 coverage

5.2. Course dataset description

This section describes each of the course datasets that were used as input into the WEKA application. Section 5.2.1 covers the division of each course dataset into variations based on the data available while Section 5.2.2 covers separating the dataset into training data and validation data. Validation is necessary in order to test the performance of the prediction models generated in each of the experiments. Section 5.2.3 covers the section on dealing with data imbalance for each of the course datasets. Finally, a description of each of the courses as well as their level of imbalance are described in section 5.2.4.

5.2.1. Testing variations of the course datasets

During training, three variations of the course dataset were considered. The first variation, referred to as Variation 1 (VAR1), tested just the demographic and assessment data. Variation 2 (VAR2) is a dataset variation with demographic, assessment data and Moodle interaction data. The above two variations are considered to better understand the impact of including LMS interaction data when attempting to predict academic performance.

The third and final variation, Variation 3 (VAR3), contains only 2020 demographic, assessment data and 2020 Moodle interaction data. VAR3 relates to data collected during the COVID-19 pandemic and a move from face-to-face learning to online learning. It was observed during the administration of these courses that working with LMS data during COVID-enforced online learning greatly differed when working with LMS data during the pre-COVID era.

5.2.2. Establishment of the training and validation sets

In order for a learning algorithm to be effective, it must be provided with a sufficient set of examples or cases. The set of examples is referred to as the training set (Smith & Frank, 2016). The learning algorithm then learns from the training set, resulting in the development of a model. In order to test this model, a separate set of examples or cases must be used where predictions are made and compared to what results are known. This set of examples is referred to as the test or validation set (Smith & Frank, 2016).

Smith and Frank (2016) identify four (4) techniques to establish the training and validation datasets. The first technique is to use the entire dataset for training as well as validation. While

this strategy makes most use of the available data and may lead to a more representative evaluation, it is not preferable as the model developed would be specific for that set of data and may not fit any other dataset (Smith & Frank, 2016). Another technique, commonly known as the holdout method, is to keep the training set and test set separate while the most common technique is to use a single dataset and split the dataset into training and test subsets (Ghorbani & Ghousi, 2020). When splitting the data into training and test data, the researcher must specify the percentage split, for example, 70% of the data is used for training and the remaining 30% to be used for testing (Smith & Frank, 2016). The final technique is commonly referred to as k-fold validation. Here, the data is divided into a set of k subsets of equal size. A single fold occurs when a single subset is used as test data while the other subsets are used to train the data. This occurs k times with each subset being given a chance to be the test data while the other subsets are used to as the training dataset. This is widely regarded as the most reliable method of establishing training and test datasets (Gudivada et al., 2017) as each data instance is allowed to be in the test dataset at least once. The disadvantage of this method is the increase in computation time when compared to the holdout method.

For this study, a combination of the holdout method and k-fold validation was used. The most recent data acquired for this study is that for the year 2021. This portion of the dataset was held back and used as the validation dataset in order to understand how well the models obtained during training performed against unseen data. A similar approach was followed by Gray, McGuinness, Owende and Hofmann (2016) where the most recent set of data was used for testing while data from the subsequent years was used for training. For the remaining data used for training, WEKA applied 10-fold validation and a resultant model was developed and assessed. The model was then tested against the unseen validation dataset.

5.2.3. Dealing with imbalanced dataset

As discussed in Section 2.5.6, imbalanced datasets are a major challenge with regard to any analytics initiative. In the case of the UKZN ISTN dataset, there is a significantly larger portion of students that have passed than that of students that have failed. Imbalance can be measured by using equation 5.1 (Madasamy & Ramaswami, 2017):

$$Imbalance = \frac{N_{major}}{N_{minor}} \quad (5.1)$$

where, N_{major} represents the number of major class instances and N_{minor} represents the number of minor class instances. According to Ortigosa-Hernández, Inza and Lozano (2017), the imbalance measurement formula (equation 5.1) is suitable for datasets with only two classes (in this case, pass and fail). A greater imbalance value indicates a more complex dataset.

The issue of data imbalance can be addressed from a data perspective as well as an algorithmic perspective. From a data perspective, four sampling techniques were used and assessed, that being no sampling, undersampling, oversampling and the synthetic minority oversampling technique (SMOTE). Section 2.5.6 described the characteristics of the latter three sampling techniques. The preprocessing filter function in WEKA allows for these three sampling techniques to be applied to the training dataset. Addressing the data imbalance problem from an algorithmic perspective is discussed in Chapter 6.

5.2.4. Course descriptions, characteristics and imbalance levels

This section describes the characteristics for each of the courses of the UKZN ISTN dataset. Table 5.1 provides the course code and semester when it is offered, the title of the course and a general background to the course. Understanding of these course details provides context for the course, enabling better understanding of the course dataset.

Table 5.1: Details of each course in the UKZN ISTN dataset

Course code and semester	Title	Course description
ISTN100 (Both semesters)	End User Computing	Teaches concepts relating to hardware, software, networks and the use of MS Office. Usually taken by students searching for an elective course or is a requirement for certain degrees.
ISTN101 (Semester 1)	Information Systems and Technology for Business	Introduces information systems concepts to the students and covers the use of MS Office. Students registered for a non-BSc degree in IS&T must pass this course to qualify for 2 nd year ISTN courses.
ISTN103 (Semester 2)	Development and Application Fundamentals	Further covers information systems related concepts with a greater focus on problem solving in information systems. Some of the topics covered are Information Systems Management, Systems Analysis and Design, logical problem solving and programming, as well as covering advanced concepts in MS-Excel. The course is a pre-requisite for 2nd year ISTN courses (with an exception for students doing B.Sc. IT degrees, where it is an elective course).
ISTN2IP (Semester 1)	Introductory Programming for Information Systems	Aimed at non-Computer Science students with the objective of improving student programming and is thus a requirement for students majoring in ISTN. Covers introductory programming concepts to students and builds on programming concepts covered in ISTN103.
ISTN211 (Semester 1)	Systems Analysis and Design	Second year course for all students majoring in IS&T. As the name suggests, the course covers the important aspects of initial analysis and design of an Information System and builds upon concepts covered in ISTN103.
ISTN212 (Semester 2)	Databases and Programming	This course builds on what was covered in ISTN2IP and prepares the student for programming at 3 rd year level. The course also introduces concepts related to database design and application, including the drawing of ERDs and creation of databases and manipulation of data using SQL Server.
ISTN3SA (Semester 1)	Advanced Systems Analysis and Design	Third year course that expands on content covered in ISTN211 and focuses on Object Oriented analysis and design.
Continued on next page...		

Table 5.1 continued		
Course code and semester	Title	Course description
ISTN3AS (Semester 1)	Applied Systems Implementation 1	A practical based 3 rd year course that covers project management, advanced programming and provides students with experience on front-end systems development. Involves the development of a windows-based application project.
ISTN3SI (Semester 2)	Applied Systems Implementation 2	This course is a continuation of the ISTN3AS course. Students expand on the project in ISTN3AS by developing a website and incorporating reporting into their systems.
ISTN3ND (Semester 2)	Networking and Databases	Provides students with knowledge of the technical background of information systems in a web and enterprise environment. To enable students to design and manage databases in a business context.

The characteristics and imbalance level of each course dataset including when considering the variations is described in Table 5.2.

Table 5.2: Characteristics of course datasets

Course		Number of instances	Years	Number of passes	Number of fails	Imbalance
ISTN100	All instances	3540	2014-2021	3044	496	6.13
	Variation 1	3416	2014-2020	2940	476	6.18
	Validation	124	2021	104	20	5.20
ISTN101	All instances	9479	2014-2021	8082	1399	5.78
	Variation 1	8729	2014-2020	7397	1332	5.55
	Variation 2	2026	2019-2020	1791	235	7.62
	Variation 3	971	2020	922	49	18.82
	Validation	738	2021	672	66	10.18
ISTN103	All instances	9046	2014-2021	7619	1427	5.34
	Variation 1	8224	2014-2020	6866	1358	5.06
	Variation 2	2204	2019-2020	1769	435	4.07
	Variation 3	1088	2020	960	128	7.50
	Validation	813	2021	745	68	10.96
ISTN2IP	All instances	631	2016-2021	517	114	4.54
	Variation 1	486	2016-2020	385	101	3.81
	Variation 2	358	2018-2020	293	65	4.51
	Variation 3	143	2020	132	11	12.00
	Validation	143	2021	130	13	10.00
ISTN211	All instances	1905	2014-2021	1812	93	19.48
	Variation 1	1576	2014-2020	1501	75	20.01
	Variation 2	768	2018-2020	729	39	18.69
	Variation 3	238	2020	229	9	25.44
	Validation	327	2021	309	18	17.17
ISTN212	All instances	1875	2014-2021	1558	317	4.91
	Variation 1	1576	2014-2020	1272	304	4.18
	Variation 2	750	2018-2020	652	98	6.65
	Variation 3	246	2020	236	10	23.60
	Validation	297	2021	284	13	21.85
ISTN3SA	All instances	1114	2016-2021	1031	83	12.42
	Variation 1	884	2016-2020	805	79	10.19
	Variation 2	527	2018-2020	470	57	8.25
	Variation 3	191	2020	189	2	94.50
	Validation	230	2021	226	4	56.50
ISTN3AS	All instances	1077	2016-2021	1054	23	45.83
	Variation 1	850	2016-2020	834	16	52.13
	Variation 2	494	2018-2020	481	13	37.00
	Variation 3	183	2020	179	4	44.75
	Validation	227	2021	220	7	31.43

Continued on next page...

Course		Number of instances	Years	Number of passes	Number of fails	Imbalance
ISTN3SI	All instances	1079	2016-2021	1031	48	21.48
	Variation 1	854	2016-2020	813	41	19.83
	Variation 2	508	2018-2020	474	34	13.94
	Variation 3	184	2020	174	10	17.40
	Validation	225	2021	218	7	31.14
ISTN3ND	All instances	1196	2016-2021	1013	183	5.54
	Variation 1	963	2016-2020	785	178	4.41
	Variation 2	586	2018-2020	470	116	4.05
	Variation 3	210	2020	205	5	41.00
	Validation	233	2021	228	5	45.60

The Moodle interaction data was not available for the ISTN100 course, thus there is only one variation, VAR1, and the validation dataset that is included for prediction. For the ISTN101 and ISTN103 courses, Moodle LMS data was only available from 2019 to 2021.

For the 2nd year courses, two extra attributes were included, that being the student's ISTN101 and ISTN103 performance in the form of a symbol (A, B, C, D, E or NA). These were included as passing these courses (or equivalent courses in the case of students with NA) are prerequisites in order to register for 2nd year courses. The ISTN2IP course was only introduced in 2016. For all 2nd year courses, Moodle interaction data was only available from 2018 to 2021.

For the 3rd year courses, three extra attributes were included, that being the student's ISTN2IP, ISTN211 and ISTN212 performance in the form of a symbol (A, B, C, D, E or NA). These were included as passing these courses (or equivalent courses in the case of students with NA) are prerequisites in order to register for 3rd year courses. The selected ISTN 3rd year courses were only introduced in 2016 and the only Moodle interaction data available was from 2018 to 2021.

5.3. Processing of datasets for prediction

This section discusses how the datasets are processed resulting in a prediction model. The literature has shown that identification of important attributes reduces computation time as well as improves accuracy through reduced overfitting (see Section 2.5.4). The process of feature selection is discussed in Section 5.3.1.

In terms of training of the data, two machine learning algorithms were chosen based on their success in other studies. The Decision Tree (DT) algorithm was used for this study as it was found to be one of the most commonly used and successful algorithms from the literature, specifically for performance prediction (see Table 2.9). In addition, the advantages of this algorithm include fast computation time and a generated model that is easy to understand and follow. The second algorithm used was an ensemble algorithm that has been used previously for addressing the dataset imbalance problem. In the case of this study, the Random Forest (RF) ensemble algorithm was tested as it had also been used successfully in the literature (see Table 2.8). Each of these algorithms are discussed in sections 5.3.2 and 5.3.3 respectively.

5.3.1. Feature selection

According to Zaffar, Hashmani and Savita (2017), feature selection algorithms analyze data with the objective of removing irrelevant data attributes to improve the performance of classifier algorithms. In addition, feature selection reduces the complexity of learned results (Zaffar et al., 2017).

The WEKA `WrapperSubsetEval` function was used for feature selection using best first-forward and best first-backward search methods respectively. This function ran multiple iterations of the specified algorithm to determine the combination of attributes that produce the best accuracy. Once the attributes were identified, the learning algorithm was executed using only the specified attributes. Salal et al. (2019) also followed this approach of performing feature selection followed by the application of a learning algorithm using the identified attributes.

5.3.2. Decision Tree algorithm

A Decision Tree is a model that can be followed sequentially, usually in a top to bottom approach. It is created by combining a number of logic tests where each test compares a numeric value against a group of ranges or a nominal value against a group of possible values (Kotsiantis, 2013). An example of a Decision Tree model with the relevant terminology is shown in Figure 5.3.

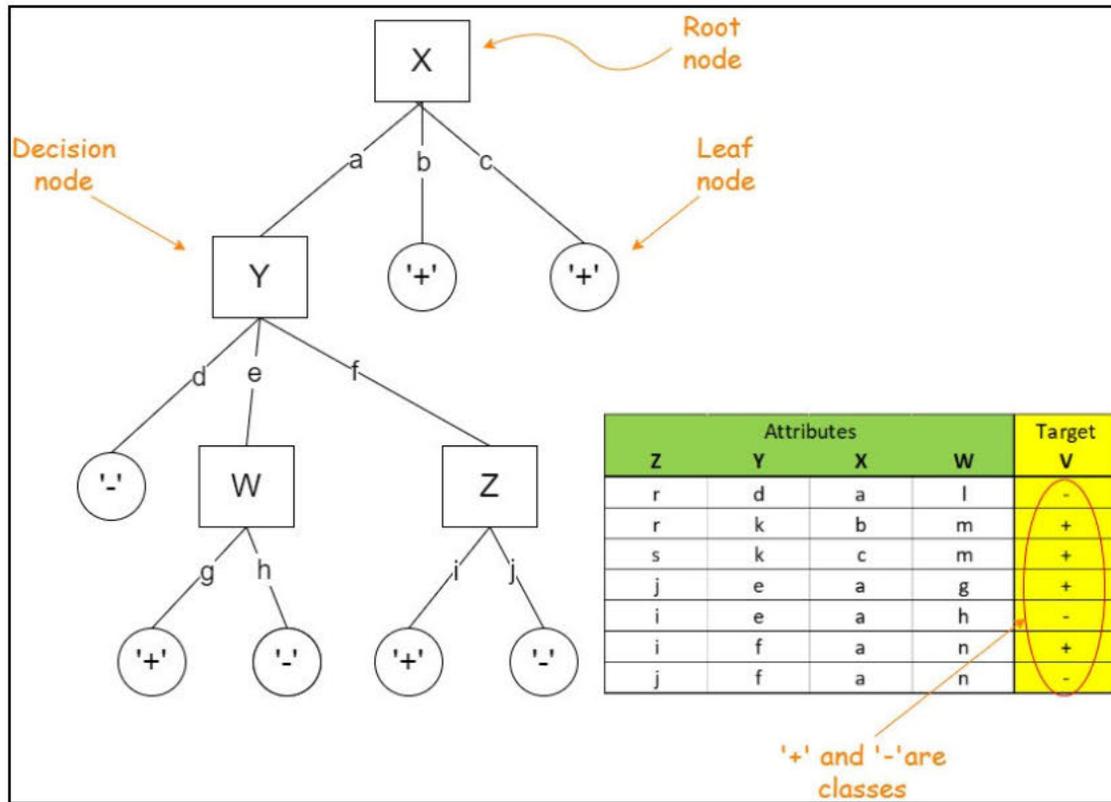


Figure 5.3: Decision tree example and concepts

The decision nodes of the Decision Tree signify the different attributes. The branches stemming from the nodes are the possible values that these attributes can have in the observed samples, and the leaf (terminal) nodes tell us the classification of the target variable.

Algorithms have been developed that allow for the creation of Decision Tree models. For this study, the J48 Decision Tree algorithm implemented in WEKA was used. This algorithm, which is an extension of the C4.5 Decision Tree algorithm is a statistical classifier that uses the concept of information entropy. Information entropy, in this context, refers to the level of uncertainty in the dataset (Kotsiantis, 2013).

Assume a training dataset $DS = \langle DS_1, DS_2, \dots \rangle$ of existing classified data. Each sample DS_i consists of a p -dimensional vector (A_1, A_2, \dots, A_p) where the A_j represents the attribute values of the related sample, as well as the class in which the sample falls. In order to gain the highest classification accuracy, the best attribute to split on is the attribute that provides the best information.

The Decision Tree algorithm follows a greedy approach and the construction of the Decision Tree is in a top-down, recursive manner (Anuradha & Velmurugan, 2015). For each recursive iteration, the algorithm chooses the attribute of the dataset that most effectively splits its set of samples into subsets of one class value or the other. The criteria for splitting are calculated from the information gain (difference in entropy). The attribute with the highest normalized information gain is chosen as the decision node. The algorithm then recurses on the partitioned subtree utilizing a divide-and-conquer approach and creates a Decision Tree based on the greedy algorithm (Rokach & Maimon, 2005). The pseudocode for the Decision Tree algorithm is shown in Algorithm 5.1:

Input: Integrated dataset	
Output: Decision tree model	
1	Function CreateDecisionTree(instances, attributes)
2	{
3	If all instances are in one class C
4	return leaf node with label C
5	else
6	if the set of attributes is empty
7	return the most common class value as leaf node
8	else
9	Select an attribute A and create a node R for it
10	For each possible value V_i of A:
11	Let Instances _i be the subset of instances that have value v_i for A
12	Add an outgoing branch B to Node R labeled with the value v_i
13	If Instances _i is empty
14	Attach leaf node to branch B labeled with most common class value in Instances
15	Else
16	Call CreateDecisionTree(instances, attributes - {A} and attach resulting tree as subtree under branch B
17	Return subtree rooted at R
18	}

Algorithm 5.1: Pseudo code for Decision Tree algorithm

5.3.3. Random Forest algorithm

According to Pal (2005), an ensemble classifier combines decisions made by multiple individual classifiers using a weighted voting mechanism to classify unseen instances of a classification problem. The Random Forest classifier is an ensemble algorithm that combines multiple decision trees using the concept of bagging. The objective of bagging is to improve accuracy via the

creation of a composite classifier made from a number of unstable classifiers (Amrieh, Hamtini & Aljarah, 2016). From the outputs of the classifiers, a single prediction is generated. Figure 5.4, adapted from Amrieh et al. (2016), shows the bagging process used in the Random Forest learning algorithm, and Algorithm 5.2 explains the Random Forest algorithm. In Figure 5.4, the ISTN101 dataset is used as an example. Here, multiple sub-datasets are created using randomly selected instances from the ISTN101 dataset resulting in the Sub datasets 1 to n. In addition, each sub-dataset uses a randomly chosen set of attributes. The classifier algorithm (in this case, the DT algorithm) is then applied to the sub-datasets. Once all the classifiers have been completed and predictions have been determined, the results from each of the classifiers are tallied with a majority voting scheme implemented, i.e., the class value that was most predicted by the classifier is chosen as the prediction value of the overall random tree algorithm (Amrieh et al., 2016).

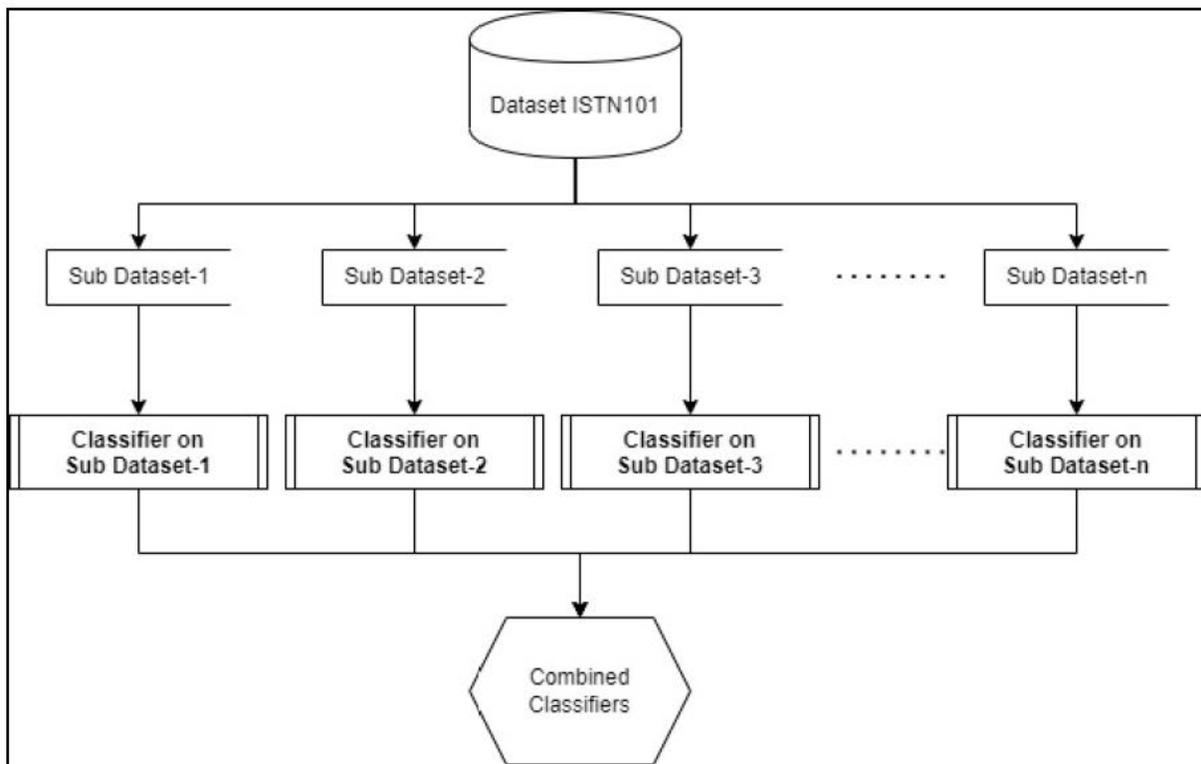


Figure 5.4: Bagging process followed in Random Forest ensemble algorithms

Input: Training dataset DS, unseen dataset	
Output: Set of predictions for given instances within unseen dataset	
1	Function RandomForest
2	{
3	Create n subsets DS1...DSn consisting of sample instances from dataset DS
4	For each subset
5	Choose a random set of attributes that will be used for decision tree creation
6	For each subset
7	Create a decision tree using the set of attributes identified for that subset
8	For each constructed decision tree
9	Make predictions based on unseen dataset
10	Prediction = class value that appears the greatest number of times
11	}

Algorithm 5.2: Random forest algorithm

5.3.4. Tools and techniques

This section describes the different functions and parameters used when applying the learning algorithms to the UKZN ISTN dataset. The tool used to run the experiments of applying the learning algorithms to the UKZN ISTN dataset is WEKA.

WEKA has a number of filter functions that can be used as part of the pre-processing or data preparation stage. The `Resample` function was used to apply oversampling to the dataset while the `SpreadSubsample` function was used to apply undersampling to the dataset. Finally, the `SMOTE` function was the only function not built-in to the standard install of WEKA. The only `SMOTE` function found in WEKA's package manager repository was downloaded and installed. This `SMOTE` function is based on that described by Chawla, Bowyer, Hall and Kegelmeyer (2002). Table 5.3 outlines the important parameters considered when applying the filter functions:

Table 5.3: WEKA Filter functions and parameters used for sampling techniques

Resample – used for oversampling	
biasToUniformClass	A value of 0 leaves the class distribution as is. For this study, the value 1 was used when oversampling was applied and the class distributions were the same.
sampleSizePercent	The subsample size as a percentage of the original set. This value is set to $Y*2$ where Y is the percentage proportion of instances that belong to the majority class.
SpreadSubSample – used for undersampling	
DistributionSpread	Indicates the maximum class distribution spread. This value was set to 1 to ensure a uniform distribution with the minority class, thus resulting in undersampling being applied.
SMOTE	
Percentage	The percentage value to increase the minority class by. For this study, the value varied for each dataset variation and the aim was to get as close as possible to uniform distribution.

Feature selection was accomplished using the `WrapperSubsetEval` function. This function evaluates sets of attributes with the objective of finding the best accuracy. `WrapperSubsetEval` is based on the approach used by Kohavi and John (1997) and determines the most useful attributes based on the classifier provided. The function searches through a feature (attribute) search space by iteratively applying a classifier using a subset of features. The number of attributes used for each classifier is determined by a best-first search engine. This search can operate in a forward or backward manner. A forward search begins with an empty set of attributes and every iteration result in the addition of attribute(s) to find the ideal set of attributes that produces the best accuracy. With a backward search, all the attributes of the problem are included and each iteration results in the removal of attributes(s) with the objective of producing the best accuracy (Kohavi & John, 1997). The important parameters for this function are listed in the Table 5.4:

Table 5.4: WEKA WrapperSubsetEval parameters

Parameters	Description
Classifier	This parameter specifies the learning algorithm that will be used when determining the best attribute set.
Folds	Number of folds required when implementing k-fold validation when running each classifier. For all experiments, the number of folds is set to 10.
Search Method Direction	Forward search or Backward search

As described earlier in the section 5.3 introduction, the J48 decision tree (DT) algorithm and the Random Forest (RF) algorithm were used for this study.

5.3.5. Assessment metrics

In order to understand how well the algorithm performs against the UKZN ISTN dataset, assessment metrics were used to determine the ability of the generated models to make predictions. The most commonly used assessment metrics are discussed in the following subsections.

5.3.5.1. Accuracy

The most common method of determining how well a model performs is the accuracy, which is defined as the count of the number of objects that have been correctly predicted by the model (Asif, Merceron, et al., 2017). This is the most commonly identified method of measuring the performance of the algorithm. The equation to calculate accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5.2)$$

In equation 5.2, *TP* (True Positive) and *TN* (True Negative) are counts of the number of correct classifications for each of the respective classes while *FP* (False Positive) and *FN* (False Negative) are the counts of incorrect classifications for each of the respective classes (Anuradha & Velmurugan, 2016). While popular due to its simplicity, accuracy cannot be the only measure for model performance as it does not consider correct predictions that occur by chance (Ben-David, 2008).

5.3.5.2. *Kappa statistic*

Another performance measure that was identified in the literature was the Kappa statistical value. Kappa was used as a performance measure by Anuradha and Velmurugan (2016), Asif, Merceron, et al. (2017), as well as Adekitan and Salau (2019). The Kappa value indicates the probability of whether or not the prediction occurs by chance, i.e., the chances of the algorithm guessing the class value. A recent study by Delgado and Tibau (2019), however, found that the Kappa statistic exhibits abnormal behaviour, especially when imbalanced datasets are taken into consideration and thus should not be considered as a measure for model performance, especially when other more reliable measures are available.

5.3.5.3. *Receiver operator characteristics (ROC)*

The ROC curve is also commonly used to measure the predictive performance of a classifying algorithm (Jayaprakash et al., 2014). Davis and Goadrich (2006) stated that it is commonly used to assess performance in binary decision problems. This performance measure was used by Jayaprakash et al. (2014), Hashim, Talab, Satty and Talab (2015), Kumar and Singh (2017), as well as Umar (2019). The ROC curve is a graph that displays and compares the number of correctly classified instances against the number of incorrectly classified instances, respectively, determined by the learning algorithm (Davis & Goadrich, 2006). Related to the ROC curve is the area under the ROC curve, a value in the range of 0 to 1, where the closer the value is to 1, the better the performance of the algorithm, i.e., a generated model that can make an accurate prediction (Mandrekar, 2010).

5.3.5.4. *Precision, recall and F-measure*

Another common measure for performance assessment is that of precision, recall and F-measure. This was included in algorithm performance analysis in studies by Algur et al. (2016), Hamoud et al. (2018), Jalota and Agrawal (2019), Ribot et al. (2020), as well as Silva et al. (2022), amongst others. In this context, precision is defined as the proportion of instances that have been correctly classified as positive while recall is defined as the proportion of only positive instances that are correctly classified (Abdullah, Malibari & Alkhozai, 2014). While both precision and recall have been identified as performance measures for learning algorithms, Ma and He (2013) state that the goal of maximizing recall and precision can often be conflicting objectives (for example, an increase in true positives – increased recall - may also result in an increase in false positives –

reduced precision). As a result, the F-measure was introduced as a metric that combines precision and recall into a single score (Sandoval et al., 2018). With the UKZN ISTN dataset being an imbalanced dataset, it should be noted that both Ma and He (2013) as well as Davis and Goadrich (2006) identified precision and recall as being more reliable in measuring the performance of an algorithm than the ROC curve. The equations for precision, recall and F-measure are shown as equations 5.3, 5.4 and 5.5 respectively:

$$Precision = \frac{TP}{TP + FP} \quad (5.3)$$

$$Recall = \frac{TP}{TP + FN} \quad (5.4)$$

$$F - Measure (F1) = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5.5)$$

Precision and recall have also been plotted on a graph similar to ROC, resulting in the Precision Recall Curve (PRC). PRC values (also identified as AUC-PRC values) can also be used to evaluate the performance of an algorithm and are commonly used as a measure for imbalanced datasets (Saito & Rehmsmeier, 2015). Similar to the ROC value, a PRC value is in the range 0 to 1, with 0.5 indicating that the algorithm is guessing. No studies have identified an acceptable range of values for PRC as this is context-dependent, depending on whether the objective is consistency (recall) or accuracy (precision).

5.3.5.5. Assessment metrics used for this study

Based on the discussions in sections 5.3.5.1 to 5.3.5.4, the performance measures and acceptance criteria used for this study are summarized in Table 5.5.

Table 5.5: Assessment metrics to be used for this study with acceptance criteria

Assessment metric	Acceptance Criteria	Justification
Accuracy	In range 80% to 98%	Majority of studies produced accuracy in this range. Minimum of 80% accuracy was hypothesized by Ifenthaler and Widanapathirana (2014). Accuracy greater than 98% should be questioned as this may indicate that the model overfits the data.
Accuracy difference between training and validation datasets	Difference should be less than 10%	A difference of less than 10% indicates accuracy from training and accuracy through validation are similar. A difference greater than 10% indicates that the model does not fit unseen data instances.
ROC	Greater than 0.7 as per Bharati, Rahman and Podder (2018)	Not seen as reliable when considering imbalanced datasets (Ma & He, 2013) thus will be considered when sampling is applied.
Precision, Recall, PRC value and F-Measure	F-measure: ≥ 0.8 Precision: ≥ 0.7 Recall: ≥ 0.7 PRC: ≥ 0.7	Han, Kamber and Pei (2012) suggested that precision and recall values of 0.7 or more are considered good models and the F-measure should be 0.8 or more. A similar suggestion was made by Schütze, Manning and Raghavan (2008). With regard to PRC, no recommended value was reported in the literature, thus for this study, an acceptable PRC value would also be in a range of 0.7 to 1.

5.4. Results of experiments conducted

This section provides the results from applying the machine learning algorithms described in Algorithm 5.1 and Algorithm 5.2 to the UKZN ISTN dataset. As described in Chapter 4, the

dataset was divided based on the IS&T courses that are required for the main ISTN major. Each of the subsections 5.4.1 to 5.4.10 covers experiments conducted on each course dataset. Figure 5.5 shows the format used for labeling each of the experiments.

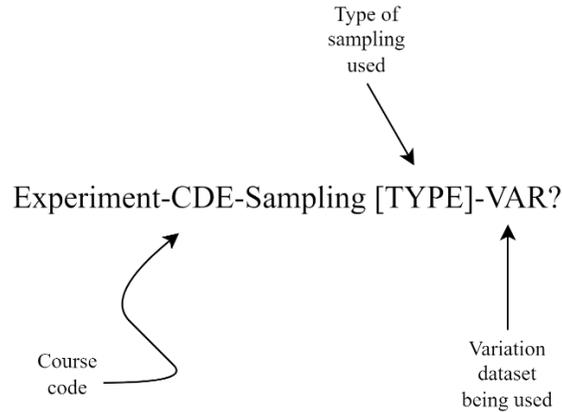


Figure 5.5: Experiment notation

In Figure 5.5, CDE represents the last three characters of the course dataset. For example, Experiment-101 indicates that the experiments are being conducted on the ISTN101 course dataset. [TYPE] indicates the type of sampling that will be applied in the experiment, either no sampling ([None]), undersampling ([US]), oversampling ([OS]) or [SMOTE]. Finally, the experiment notation ends by indicating the dataset variation being used (VAR1, VAR2 or VAR3).

An example of how the assessment metrics are presented in a tabular format is shown in Figure 5.6. The analysis generated when applying the DT algorithm is shown in the top half of the table, while the results of the RF algorithm are shown in blue section of the table (bottom half). The column labelled “10-Fold” relates to the analysis from training using 10-fold validation. The resultant model generated is then applied to the validation dataset and the analysis for this is shown in the column labelled “Validation”. A “?” value indicates that a value for this assessment metric could not be calculated.

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search													
Backward Search													
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search													
Backward Search													

Figure 5.6: Table format for presenting performance analysis

5.4.1. Experiments for the ISTN100 dataset

For the ISTN100 course, the Moodle interaction (log) data was not available for any years and thus was not included. As a result, only VAR1 is considered for the experiments for the ISTN100 dataset.

5.4.1.1. Experiment-100-Sampling [None]

For both the DT AND RF algorithms, `WrapperSubsetEval` was applied for feature selection in WEKA using both the forward and backward searches respectively. A summary of the performances is shown in Table 5.6.

Table 5.6: Summary analysis for RF generated model – Experiment-100-Sampling [None]-VARI

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	86	83.8	0.49	0.5	0.76	0.72	0.74	?	0.86	0.83	0.79	?
Backward Search	None	86	83.8	0.49	0.5	0.76	0.72	0.74	?	0.86	0.83	0.79	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	86.4	83.8	0.53	0.5	0.78	0.73	0.85	?	0.86	0.83	0.8	?
Backward Search	28	85.8	83.8	0.6	0.7	0.82	0.85	0.82	0.79	0.85	0.83	0.82	0.79

For the DT algorithm, feature selection yielded no specific attributes for either search, and the best accuracy obtained was 86%. The resultant DT model was composed of a single leaf (terminal)

node, i.e., “P” (pass). This is due to the imbalance of the ISTN100 course data where the majority class is “P”. The precision and F-measure values that could not be calculated also suggest that the generated prediction model would not be useful.

Applying the RF algorithm yields a similar analysis in terms of the assessment metrics. While feature selection did identify a set of attributes, an accuracy of 86.4% and 85.8% is similar to that of the model produced by the DT algorithm. For the forward search, precision and F-measure values could not be calculated, while for the backward search, an acceptable model was found.

5.4.1.2. Experiment-100-Sampling [US]

In this experiment, undersampling using the `spreadSubSample` filter was applied to the dataset with the objective of mitigating the imbalance issue. Instances were removed from the majority class resulting in an equal number of instances that have passed (“P”) and failed (“F”) the course. Once again, feature selection via `WrapperSubsetEval` was applied, followed by application of the respective learning algorithms. The analysis of each algorithm is shown in Table 5.7. For these experiments, the `DistributionSpread` parameter is set to 1, thus applying undersampling to the dataset.

Table 5.7: Summary analysis for RF generated model – Experiment-100-Sampling [US]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	49.8	83.8	0.5	0.5	0.5	0.72	0.49	?	0.49	0.83	0.49	?
Backward Search	None	49.8	83.8	0.5	0.5	0.5	0.72	0.49	?	0.49	0.83	0.49	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	11	66	59.6	0.67	0.62	0.64	0.79	0.66	0.76	0.66	0.59	0.66	0.65
Backward Search	18	66.4	53.2	0.68	0.6	0.65	0.78	0.66	0.77	0.66	0.53	0.66	0.59

Both algorithms exhibit models with poor accuracy, with 49.8% and 66% for DT algorithm and RF algorithm, respectively. Similar to *Experiment-100-Sampling [None]*, the DT algorithm generated model was composed of a single terminal leaf that represents the “P” class. In addition,

the ROC values indicate poor predictive capability from the generated models. While the model generated by the RF algorithm is slightly better according to ROC value, the accuracy obtained indicates that undersampling did not improve the performance of the algorithms.

5.4.1.3. Experiment-100-Sampling [OS]

For this experiment, oversampling was applied using the `Resample` filter method. Feature selection was then applied to determine the optimal attributes to use when applying the learning algorithms. For oversampling to be applied to this dataset, the `biasToUniformClass` parameter was set to 1 and the `sampleSizePercent` parameter was set to 172 as the proportion of passes was calculated to 86% (see ISTN100 data in Table 5.2 and `sampleSizePercent` formula in Table 5.3). Once the two algorithms were applied to the oversampled dataset using the specified attributes, the following analysis was generated (Table 5.8):

Table 5.8: Summary analysis for RF generated model – Experiment-100-Sampling [OS]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	14	91.6	80.6	0.96	0.62	0.94	0.78	0.92	0.79	0.91	0.8	0.91	0.8
Backward Search	14	91.7	80.6	0.96	0.62	0.94	0.78	0.92	0.79	0.91	0.8	0.91	0.8
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	16	90.9	62.9	0.98	0.69	0.98	0.82	0.91	0.77	0.9	0.62	0.9	0.67
Backward Search	28	92.2	64.5	0.9	0.6	0.98	0.79	0.92	0.75	0.92	0.64	0.92	0.68

The accuracy achieved by the algorithms is better when oversampling was applied to the dataset during the training phase. There was, however, a reduction in accuracy when the generated DT and RF models were applied to the unseen validation dataset. This can also be verified when seeing the difference between the ROC, PRC, Precision, recall and F-Measure values (10-fold vs validation), thus confirming that the generated models do not fit when applied to unseen data.

5.4.1.4. Experiment-100-Sampling [SMOTE]

For this experiment, the SMOTE filter was applied to the data, followed by application of feature selection and the learning algorithms based on the attributes identified. Table 5.9 summarizes the analysis of the models generated by the DT and RF algorithms when applied to the SMOTE sampled dataset. For these experiments, the Percentage value parameter (see Table 5.3 for SMOTE) was set to 515.

Table 5.9: Summary analysis for RF generated model – Experiment-100-Sampling [SMOTE]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	11	89.3	84.6	0.9	0.6	0.86	0.78	0.89	0.82	0.89	0.84	0.89	0.82
Backward Search	16	89.1	83.8	0.9	0.6	0.87	0.79	0.89	0.81	0.89	0.83	0.89	0.82
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	87.4	77.4	0.94	0.6	0.93	0.79	0.87	0.76	0.87	0.77	0.87	0.76
Backward Search	29	89.8	80.6	0.9	0.6	0.94	0.82	0.89	0.77	0.89	0.8	0.89	0.78

In this experiment, the accuracy obtained after training the algorithms are similar, ranging from 87.4% to 89.8%. It would appear, however, that the model generated using the DT algorithm was a better fit for the validation data than the RF generated model. The precision, recall and F-measure are all in the acceptable range while the ROC value generated during validation is slightly less than the acceptable value of 0.7 specified in Table 5.5.

5.4.1.5. Analysis of experiments conducted

When no sampling was used, the only viable prediction model was generated when the RF algorithm was used. When undersampling was used, poor prediction models were generated with prediction accuracy of less than 70%. This was due to the removal of a large number of instances that could have been useful in predicting student performance.

On the other hand, the use of oversampling resulted in acceptable accuracies greater than 90%, but similar accuracy could not be replicated for the validation dataset, where the accuracy obtained

was about 80% (DT algorithm model) and 62% (RF algorithm model) respectively. As described in section 2.5.6, increasing the number of instances from the minority class (failure instances) removed the imbalance problem but caused a resultant bias towards the number of failures, and the model could not adequately predict unseen instances (as seen with the validation accuracy).

Finally, as seen in section 5.4.1.4, the use of SMOTE resulted in the best performance of the algorithms where three of the four experiments conducted yielded viable prediction models. The DT algorithm using forward search feature selection had the closest difference in accuracy, as well as acceptable precision, recall and F-measure values. The SMOTE method of creating synthesized instances of the minority class resulted in a balanced dataset with less bias than when using oversampling.

5.4.2. Experiments for the ISTN101 dataset

Unlike the experiments run on the ISTN100 dataset, the Moodle activity log data was made available for this course and thus the three variations (VAR1, VAR2, VAR3) discussed in section 5.2.1 are considered in the experiments.

5.4.2.1. Experiment-101-Sampling [None]

When no sampling was used, the model generated by the DT algorithm was a tree with a single terminal leaf (“P”), resulting in the accuracy produced being the same as the percentage of students that passed the course. This, along with undefined precision and F-measure values indicate an unacceptable model.

The performance measures for the model generated by the RF algorithm are shown in Table 5.10. Feature selection using forward search for RF algorithm identified three attributes. The RF model generated from the 24 attributes (backward search) had a better ROC value, indicating a better model than the RF algorithm generated using only three attributes. The F-measure and PRC areas were also better with the RF generated model. However, the validation accuracy of 98.1% falls outside the acceptable range for this study.

Table 5.10: Summary analysis for RF generated model – Experiment-101-Sampling [None]-VAR1

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	85.2	91.3	0.56	0.5	0.76	0.83	0.85	0.92	0.85	0.91	0.78	0.87
Backward Search	24	83.4	98.1	0.66	0.99	0.81	0.99	0.78	0.98	0.83	0.98	0.8	0.98

An improvement in performances of both the algorithms was observed when they were applied to the VAR2 dataset (see the performance measures in Table 5.11). The resultant analysis showed an improvement in accuracy, PRC value and F-Measure when compared to the performance measures of the algorithms applied to the VAR1 dataset (Table 5.10).

Table 5.11: Summary analysis for RF and DT generated models – Experiment-101-Sampling [None]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	88.8	92.1	0.62	0.58	0.83	0.86	0.86	0.92	0.88	0.92	0.85	0.89
Backward Search	14	89	92	0.63	0.58	0.84	0.86	0.87	0.92	0.89	0.92	0.85	0.88
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	32	88.6	91.5	0.77	0.84	0.77	0.93	0.85	0.92	0.88	0.91	0.84	0.88
Backward Search	29	88.7	91.5	0.76	0.83	0.88	0.92	0.86	0.9	0.88	0.91	0.84	0.88

The VAR3 dataset contains data collected during the first year of the COVID-19 pandemic where teaching and learning had moved to an online learning platform using the Moodle LMS. The analysis of models generated when the learning algorithms were applied to the VAR3 dataset is shown in Table 5.12.

Table 5.12: Summary Analysis for RF and DT generated models – *Experiment101-Sampling [None]-VAR3*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	94.9	91	0.49	0.5	0.9	0.83	?	?	0.95	0.91	?	?
Backward Search	None	94.9	91	0.49	0.5	0.9	0.83	?	?	0.95	0.91	?	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	38	94.9	91	0.76	0.84	0.93	0.92	?	?	0.95	0.91	?	?
Backward Search	9	95.3	92.6	0.75	0.8	0.94	0.93	0.94	0.92	0.95	0.92	0.93	0.9

Once again, the DT model generated was just a single terminal leaf (“P”). The improved accuracy obtained was due to the higher pass rate (during COVID-19) when compared to the other two variations. This is due to the movement to the online mode of assessments where students could use their downloaded lecture content and communicate with other students during assessments. The high pass rates resulted in a greater dataset imbalance with respect to the number of passes and failures. The PRC values of the RF generated model (backward search) are better than that of the DT generated model, indicating a more reliable model when using the RF algorithm. As seen in the literature, the Random Forest algorithm is also known to provide better models when applied to imbalanced datasets than when using decision tree algorithms (Bekkar & Alitouche, 2013).

5.4.2.2. *Experiment-101-Sampling [US]*

For the next three experiments, the process of undersampling was applied to the three dataset variations. Feature selection was then applied and using the specified attributes, the two learning algorithms (DT and RF) were applied to the undersampled data.

For VAR1, the model performance measures were poor with an accuracy of below 70% when using the test data (see Table 5.13).

Table 5.13: Summary analysis for RF and DT generated models – Experiment-101-Sampling [US]-VARI

Decision Tree (J48)													
Variation 1	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
	Forward Search	8	66.4	71.3	0.7	0.72	0.68	0.89	0.66	0.87	0.66	0.71	0.66
Backward Search	18	66	68.9	0.68	0.77	0.65	0.9	0.66	0.9	0.66	0.68	0.66	0.75
Random Forest													
Attribute Count	Accuracy %	ROC		PRC Area		Precision		Recall		F-Measure			
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	65.3	80.2	0.68	0.73	0.66	0.89	0.65	0.88	0.65	0.8	0.65	0.83
Backward Search	22	64.9	70.9	0.69	0.95	0.67	0.96	0.65	0.93	0.64	0.7	0.64	0.77

When the learning algorithms were applied to VAR2, an improvement in accuracy, while still below 80%, was achieved (see Table 5.14). Applying the resultant model to the validation data achieved an accuracy score of about 93%. However, the difference between accuracy for the validation and 10-fold (training) of between 17% to 20% indicates that the model would be unpredictable when applied to unseen data instances.

Table 5.14: Summary analysis for RF and DT generated models – Experiment-101-Sampling [US]-VAR2

Decision Tree (J48)													
Variation 2	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
	Forward Search	7	74.25	93.2	0.71	0.75	0.67	0.9	0.75	0.92	0.74	0.93	0.73
Backward Search	19	76.59	93	0.78	0.73	0.74	0.91	0.76	0.92	0.76	0.93	0.76	0.92
Random Forest													
Attribute Count	Accuracy %	ROC		PRC Area		Precision		Recall		F-Measure			
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	16	76.3	93.6	0.8	0.81	0.79	0.93	0.76	0.93	0.76	0.93	0.76	0.92
Backward Search	23	73.1	90.9	0.78	0.86	0.76	0.93	0.73	0.9	0.73	0.9	0.73	0.9

For VAR3, the difference between validation and testing (10-fold) was about 10% to 12% (see Table 5.15). The model produced using RF (backward best-first search) was the only model with close accuracy between training and validation datasets.

Table 5.15: Summary analysis for RF and DT generated models – Experiment-101-Sampling [US]-VAR3

Decision Tree (J48)													
Variation 3	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation										
	Forward Search	3	74,48	86,85	0,65	0,77	0,6	0,89	0,74	0,91	0,74	0,86	0,74
Backward Search	5	80,6	70	0,83	0,65	0,8	0,87	0,8	0,87	0,8	0,7	0,8	0,76
Random Forest													
Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	5	75,5	63,95	0,65	0,61	0,61	0,85	0,75	0,86	0,75	0,64	0,75	0,71
Backward Search	26	81.6	79.1	0,87	0,83	0,85	0,92	0,81	0,89	0,81	0,79	0,81	0,83

All the experiments discussed in this section did not produce prediction models with acceptable accuracy for both training and validation.

5.4.2.3. Experiment-101-Sampling [OS]

This section covers the analysis generated when the two learning algorithms were applied to the three dataset variations with oversampling applied to address the class imbalance issue. VAR1 analysis is shown in Table 5.16. For VAR1, the `sampleSizePercent` parameter value is 170 in order to increase the number of instances from the minority class.

Table 5.16: Summary analysis for RF and DT generated models – Experiment-101-Sampling [OS]-VAR1

Decision Tree (J48)													
Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	18	89.2	84	0.94	0.96	0,92	0,96	0,89	0,94	0,89	0,84	0,89	0,87
Backward Search	22	89.4	84.4	0.94	0.96	0,92	0,96	0,9	0,94	0,89	0,84	0,89	0,87
Random Forest													
Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	16	94	91.3	0.99	1	0,98	1	0,94	0,95	0,94	0,91	0,94	0,92
Backward Search	27	94.5	92.6	0.99	1	0,99	1	0,94	0,96	0,94	0,92	0,94	0,93

With oversampling, a vast improvement in accuracy was noted when compared to the undersampled variations and was similar to that of when no sampling was used. In this case

however, the ROC, PRC and F-Measure values indicate that the models generated are better than the respective values generated using no sampling and undersampling, respectively.

In the second oversampling experiment, the learning algorithms were applied to the oversampled VAR2 dataset. As with the experiments using VAR1, for VAR2, models have been generated that produced acceptable accuracy with both the training and validation datasets. The results are shown in Table 5.17. For VAR2 experiments, the `sampleSizePercent` parameter value is 196.

Table 5.17: Summary analysis for RF and DT generated models – Experiment-101-Sampling [OS]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	92.9	82.2	0.97	0.63	0,96	0,86	0,93	0,85	0,93	0,82	0,93	0,83
Backward Search	23	95.8	84.5	0.97	0.64	0,96	0,86	0,96	0,87	0,95	0,84	0,95	0,85
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	19	98.8	92	1	0.73	1	0,9	0,98	0,92	0,98	0,92	0,98	0,88
Backward Search	28	98.4	92.6	1	0.8	1	0,93	0,98	0,93	0,98	0,92	0,98	0,9

Overall, the RF generated model produces accuracy of over 98% with models that fit well with the validation dataset. This is confirmed by the high ROC, PRC and F-measure values. In the case of the models generated by the RF algorithms, as the accuracy falls outside the acceptance criteria, the models were not accepted.

The third experiment focused on the application of DT and RF algorithms to the oversampled VAR3 data (with feature selection). The analysis of the model performance measures is shown in Table 5.18. The `sampleSizePercent` value for the VAR3 experiments is 194.

Table 5.18: Summary analysis for RF and DT generated models – *Experiment-101-Sampling [OS]-VAR3*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	98.6	91.4	0.99	0.65	0,98	0,87	0,98	0,9	0,98	0,91	0,98	0,9
Backward Search	27	98.2	89.8	0.98	0.49	0,98	0,83	0,98	0,82	0,98	0,89	0,98	0,86
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	26	81.6	79.1	0.87	0.83	0,85	0,92	0,81	0,89	0,81	0,79	0,81	0,83
Backward Search	8	99.6	93.3	1	0.79	1	0,92	0,99	0,93	0,99	0,93	0,99	0,91

As with the previous two experiments on oversampling, high accuracy is achieved for the COVID-19 dataset of 2020 (VAR3) and the resultant models were also able to perform predictions with a high degree of accuracy for the validation dataset as well. This is confirmed with high ROC, PRC and F-Measure values. The accuracy produced, however, falls outside the acceptable range and thus the models were not acceptable.

5.4.2.4. Experiment-101-Sampling [SMOTE]

The final three experiments applied for this course dataset covered the application of SMOTE to the three variations. Once applied, feature selection determined the best attributes to use and the learning algorithms were applied to only these attributes of the dataset. The prediction model analysis generated when applying the algorithms to VAR1 is shown in Table 5.19. In order to implement SMOTE, the Percentage parameter (see Table 5.3) is set to 475.

Table 5.19: Summary analysis for RF and DT generated models – *Experiment-101-Sampling [SMOTE]-VAR1*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	25	84.6	83	0.88	0.79	0,88	0,92	0,84	0,87	0,84	0,83	0,84	0,85
Backward Search	11	86.8	88.5	0.88	0.91	0,85	0,95	0,86	0,91	0,86	0,88	0,86	0,89
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	16	87.3	98.9	0.94	0.99	0,93	0,98	0,87	0,99	0,87	0,98	0,87	0,99
Backward Search	27	87.6	98.1	0.94	0.99	0,94	0,99	0,87	0,98	0,87	0,98	0,87	0,98

The accuracy, while acceptable, is not as high as when oversampling was applied to this variation. The other measures (ROC, PRC, Precision, Recall and F-Measure) used for assessing the quality of the model indicate that the resultant model from both the DT algorithm and the RF algorithm are useful for making good predictions.

From the observation of the results obtained from the VAR2 and VAR3 (Table 5.20) experiments, it once again appears that the inclusion of the Moodle interaction data enhanced the ability of the algorithm to generate models that can predict student performance with an acceptable accuracy. For VAR2 and VAR3 experiments, the Percentage parameter was set to 650 and 1750 respectively.

Table 5.20: Summary analysis for RF and DT generated models – *Experiment-101-Sampling [SMOTE]-VAR2 and Experiment-101-Sampling [SMOTE]-VAR3*

Decision Tree (J48)													
Variation 2	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation										
	Forward Search	9	91.6	85.7	0.91	0.67	0,89	0,87	0,91	0,86	0,91	0,85	0,91
Backward Search	27	89.5	76.8	0.91	0.65	0,88	0,87	0,89	0,86	0,89	0,76	0,89	0,8
Random Forest													
Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	3	89.2	81.7	0.93	0.6	0,92	0,86	0,89	0,84	0,89	0,81	0,89	0,82
Backward Search	27	92.4	90.9	0.97	0.76	0,97	0,9	0,92	0,88	0,92	0,9	0,92	0,88
Decision Tree (J48)													
Variation 3	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation										
	Forward Search	5	96.8	79.5	0.96	0.61	0,95	0,86	0,96	0,84	0,96	0,79	0,96
Backward Search	13	96.4	89,56	0,97	0,63	0,96	0,63	0,96	0,86	0,96	0,86	0,96	0,87
Random Forest													
Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	3	96.5	86.3	0.97	0.6	0,97	0,87	0,96	0,85	0,96	0,86	0,96	0,85
Backward Search	30	97.5	89	0.99	0.8	0,99	0,91	0,97	0,84	0,97	0,89	0,97	0,86

5.4.2.5. Analysis of experiments conducted

From the experiments conducted in sections 5.4.2.1 to 5.4.2.4, three models were identified as the best performing models from no sampling, oversampling and SMOTE datasets, respectively. When undersampling was used, no viable models were generated by either of the algorithms. This was mostly likely due to reducing the number of instances to match the minor class (Fail) resulting in the loss of useful instances that would have contributed to better prediction. This is a common issue with undersampling and thus was noted as a potential disadvantage by Fernández et al. (2018).

All three of the models were generated using the Random Forest algorithm and are listed in Table 5.21. The first model (named VAR3-None), was generated using the VAR3 dataset with no sampling. The VAR1-OS model was generated using the oversampled VAR1 dataset. Finally, the VAR2-SMOTE model was developed using the VAR2 dataset with SMOTE applied.

Table 5.21: Best three models generated from *Experiment-101*

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR3 - None	95,3	92,3	0,75	0,8	0,94	0,93	0,94	0,92	0,95	0,92	0,93	0,9
VAR1 - OS	94,5	92,6	0,99	1	0,99	1	0,94	0,96	0,94	0,92	0,94	0,93
VAR2 - SMOTE	92,4	90,9	0,97	0,76	0,97	0,9	0,92	0,88	0,92	0,9	0,92	0,88

A comparison of accuracy between the three models is illustrated in Figure 5.7. The difference between the training accuracy (10-fold) and the validation accuracy for the VAR3-None prediction model is 3. This is higher than the other two models, most likely because the dataset used was the most imbalanced dataset (VAR3 had an imbalance value of 18.82 as shown in Table 5.2). This, coupled with no sampling applied, indicates that while the RF algorithm did produce an acceptable model, the use of sampling techniques can assist in reducing bias and produce better models. The difference in accuracies between the VAR1-OS model and the VAR2-SMOTE model were 1.9 and 1.3 respectively.

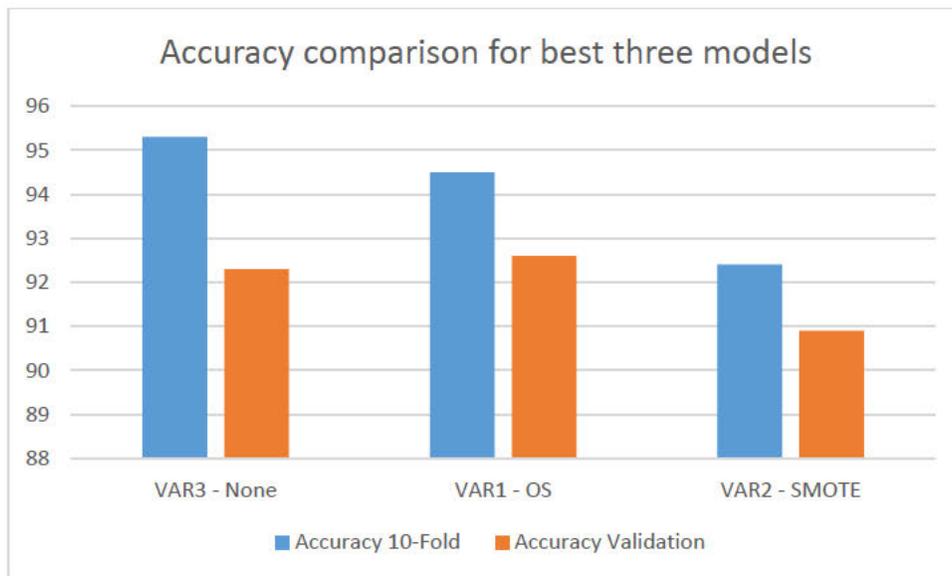


Figure 5.7: Accuracy comparison of best three models

The remaining assessment measures are compared and shown in Figure 5.8. The ROC values, as expected, are lower for the VAR3-None algorithm due to imbalance of the dataset. The PRC, precision, recall and F-measure values all fall within the acceptable range for all three models.

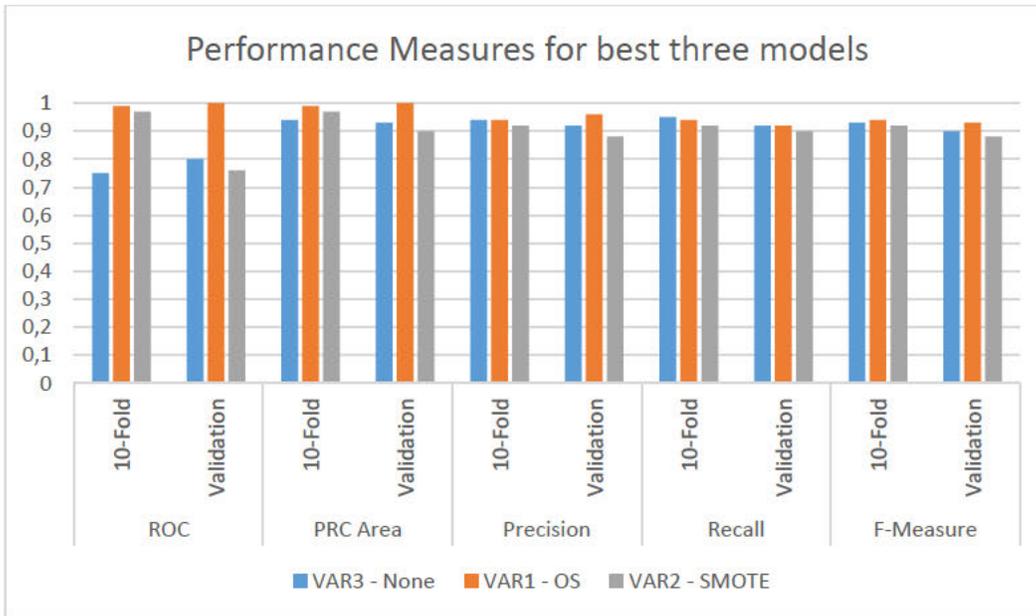


Figure 5.8: Performance measure comparison of best three models

5.4.3. Experiments for the ISTN103 dataset

This section covers the experiments conducted for the DT and RF algorithms applied to the ISTN103 course data.

5.4.3.1. Experiment-103-Sampling [None]

For the VAR1 dataset, the model generated from the DT algorithm contained only a single leaf (“P”). Thus, the accuracy for both the training and validation datasets are determined by the pass rate of the ISTN103 dataset. The ROC value of 0.5 also indicates that for any instance, a guess of “P” will more often than not produce a correct prediction. The RF model generated was better than that of the DT (based on the ROC and PRC scores), but the accuracy of both models generated from the forward and backward search algorithms are similar.

Table 5.22 shows the analysis of the learning algorithms’ respective model performance when applied to the VAR2 dataset. The PRC values indicate that the models are better when compared to the VAR1 analysis. In this case, the addition of Moodle interaction data has played a role in the generation of better prediction models. However, with the exception of the backward-search DT generated model, the model accuracy difference is greater than 10, meaning that the models cannot be accepted based on the acceptance criteria.

Table 5.22: Summary analysis for RF and DT generated models – *Experiment-103-Sampling [None]-VAR2*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	85,1	95,6	0,77	0,84	0,83	0,94	0,84	0,95	0,85	0,95	0,83	0,95
Backward Search	24	84,4	94,3	0,7	0,72	0,79	0,91	0,83	0,93	0,84	0,94	0,82	0,93
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	80,8	93,3	0,68	0,83	0,78	0,93	0,77	0,92	0,8	0,93	0,77	0,92
Backward Search	24	84,5	95,8	0,83	0,91	0,88	0,96	0,83	0,95	0,84	0,95	0,82	0,95

With the results from VAR3 (Table 5.23), the observation was made that the performance measures between the training data and validation data were closer to each other than when compared to that of the measures seen for VAR1 and VAR2. The PRC and F-measure values also indicate that the models were good at making predictions despite the imbalance of the dataset.

Table 5.23: Summary analysis for RF and DT generated models – *Experiment-103-Sampling [None]-VAR3*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	93,1	94,5	0,73	0,82	0,88	0,93	0,93	0,94	0,93	0,95	0,92	0,94
Backward Search	14	93,1	95,5	0,72	0,74	0,88	0,91	0,93	0,95	0,93	0,95	0,92	0,94
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	10	93	95,5	0,81	0,89	0,91	0,96	0,92	0,95	0,93	0,95	0,92	0,95
Backward Search	32	92,6	95,9	0,86	0,9	0,93	0,96	0,92	0,95	0,92	0,95	0,91	0,95

5.4.3.2. *Experiment-103-Sampling [US]*

The learning algorithms were applied to the undersampled variations. For VAR1, the generated models from both algorithms yielded poor accuracy (ranging from 63% to 67%). Accuracy was found to have improved when the algorithms were applied to the VAR2 dataset. However, when applied to the validation dataset, a difference of about 20% between training and validation accuracies indicate that the model would be unreliable when applied to unseen instances. For

VAR3, while there was better alignment in accuracy between the training and validation dataset, the accuracies indicated were between 76.5% and 83%. As a result, no acceptable models were generated when undersampling was applied to the dataset variations.

5.4.3.3. Experiment-103-Sampling [OS]

When oversampling is applied to VARI (with a `sampleSizePercent` value of 166), the resultant models generated by the learning algorithms for training were in the range of 86.4% to 93.5%. However, when applied to the unseen data, accuracies were in the range 63.3% to 69.7% (see Table 5.24). This indicates overfitting of the models to the training data and predictions cannot be made for unseen data instances.

Table 5.24: Summary analysis for RF and DT generated models – Experiment-103-Sampling [OS]-VARI

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	17	88,4	69,2	0,93	0,61	0,92	0,86	0,88	0,87	0,88	0,69	0,88	0,75
Backward Search	24	86,4	63,3	0,93	0,62	0,92	0,86	0,87	0,88	0,86	0,63	0,86	0,71
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	93,1	68,8	0,98	0,71	0,98	0,89	0,93	0,88	0,93	0,68	0,93	0,75
Backward Search	26	93,5	69,7	0,98	0,75	0,98	0,9	0,94	0,88	0,93	0,69	0,93	0,76

An improvement in performance was shown when Moodle interaction data was included, as shown by the performance of the algorithms when applied to the VAR2 dataset using a `sampleSizePercent` of 160 (Table 5.25). This once again, re-emphasizes the role of student interactions when predicting student performance. The accuracy difference for the models generated by the DT algorithm were greater than 10, meaning that these models are unacceptable. The RF algorithm generated models were better with close accuracy between training and validation data. The acceptable performance measures (ROC, PRC, precision, recall and F-measure) also confirm the reliability of the models generated by the RF algorithm.

Table 5.25: Summary analysis for RF and DT generated models – Experiment-103-Sampling [OS]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	95,3	80,8	0,98	0,69	0,98	0,88	0,95	0,87	0,95	0,8	0,95	0,83
Backward Search	30	97,3	85,1	0,97	0,78	0,96	0,91	0,97	0,91	0,97	0,85	0,97	0,87
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	12	96,5	95,8	0,99	0,9	0,99	0,96	0,96	0,95	0,96	0,95	0,96	0,95
Backward Search	21	96,8	95,6	0,99	0,89	0,99	0,96	0,96	0,95	0,96	0,95	0,96	0,95

The performance measures for the models generated when the algorithms were applied to the oversampled VAR3 dataset (using `sampleSizePercent` of 176) are shown in Table 5.26. The training accuracy for the RF generated models is greater than 98%, indicating that the models are not acceptable based on the acceptance criteria of the study. The backward search DT algorithm did produce an acceptable model in terms of all the performance measurements.

Table 5.26: Summary analysis for RF and DT generated models – Experiment-103-Sampling [OS]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	91,5	79,7	0,92	0,59	0,89	0,86	0,91	0,86	0,91	0,79	0,91	0,82
Backward Search	30	89,4	88,3	0,91	0,85	0,89	0,93	0,89	0,92	0,89	0,88	0,89	0,89
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	99,5	94,2	0,99	0,84	0,99	0,95	0,99	0,94	0,99	0,94	0,99	0,94
Backward Search	29	99,5	94,8	1	0,89	1	0,96	0,99	0,94	0,99	0,94	0,99	0,94

5.4.3.4. Experiment-103-Sampling [SMOTE]

The results of the three experiments are shown in the tables starting with Table 5.27 covering VAR1. Here, the `Percentage` parameter value is set to 400.

Table 5.27: Summary analysis for RF and DT generated models – Experiment-103-Sampling [SMOTE]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	84,5	79	0,87	0,56	0,83	0,86	0,84	0,85	0,84	0,79	0,84	0,82
Backward Search	14	83,9	76,9	0,88	0,6	0,85	0,86	0,84	0,86	0,83	0,77	0,83	0,8
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	86,2	80	0,93	0,73	0,93	0,9	0,86	0,88	0,86	0,8	0,86	0,83
Backward Search	25	86,8	79,5	0,93	0,72	0,93	0,9	0,86	0,87	0,86	0,79	0,86	0,83

The RF generated model performed better than the DT generated model in terms of all the performance measures when applied to the VAR1 dataset. Improved models were generated once again when the Moodle interaction data was included as shown in Table 5.28 (VAR2 – Percentage parameter value is 300) and Table 5.29 (VAR3 – Percentage parameter value is 650).

Table 5.28: Summary analysis for RF and DT generated models – Experiment-103-Sampling [SMOTE]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	85,6	85,1	0,86	0,69	0,82	0,88	0,85	0,88	0,85	0,85	0,85	0,86
Backward Search	20	85,8	90,4	0,87	0,83	0,84	0,93	0,85	0,92	0,85	0,9	0,85	0,91
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	81,5	91,2	0,83	0,76	0,81	0,91	0,83	0,9	0,81	0,91	0,81	0,9
Backward Search	28	89	95,5	0,95	0,92	0,95	0,97	0,89	0,95	0,89	0,95	0,89	0,95

In the case of VAR2, all four experiments yielded acceptable prediction models with the DT generated models having better accuracy difference. In terms of the other assessment measures, the models generated using the RF algorithms were deemed to be more reliable models.

Table 5.29: Summary analysis for RF and DT generated models – *Experiment-103-Sampling [SMOTE]-VAR3*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	92,3	83,3	0,93	0,59	0,9	0,86	0,92	0,86	0,92	0,83	0,92	0,84
Backward Search	21	89,2	91,5	0,91	0,84	0,9	0,93	0,89	0,93	0,89	0,91	0,89	0,92
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	92,2	26,9	0,95	0,76	0,94	0,91	0,92	0,92	0,92	0,26	0,92	0,32
Backward Search	17	95,6	93,8	0,98	0,85	0,98	0,93	0,95	0,93	0,95	0,93	0,95	0,93

The accuracies produced by the models generated using oversampled VAR3 were better than those when oversampled VAR2 was used. The RF backward search generated model performed best in terms of all the assessment measures.

5.4.3.5. Analysis of experiments conducted

As with the ISTN101 dataset, the use of undersampling did not produce any valid models from the experiments conducted and described in Section 5.4.3.2. The best model identified when the other sampling techniques were used are listed in Table 5.30. All three models were generated using the RF algorithm, although the DT algorithm also managed to produce acceptable models.

Table 5.30: Best three models generated from *Experiment-103*

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR3 - None	93	95,5	0,81	0,89	0,91	0,96	0,92	0,95	0,93	0,95	0,92	0,95
VAR2 - OS	96,5	95,8	0,99	0,9	0,99	0,96	0,96	0,95	0,96	0,95	0,96	0,95
VAR3 - SMOTE	95,6	93,8	0,98	0,85	0,98	0,93	0,95	0,93	0,95	0,93	0,95	0,93

In terms of accuracy, all the models produce more than 90% accuracy when training and similar accuracy was obtained when the models were applied to the validation dataset. The VAR2-OS model appears to be the best model in terms of accuracy and closeness (see Figure 5.9).

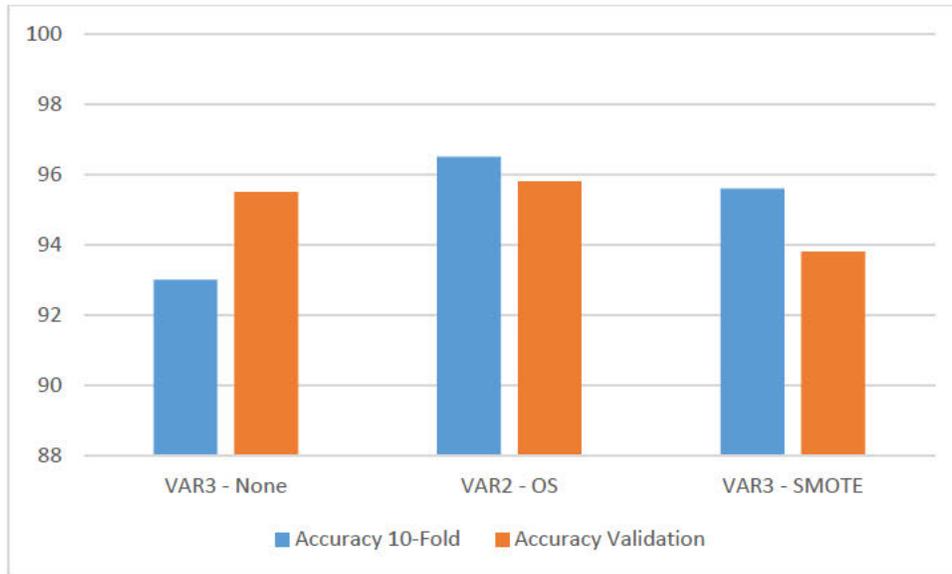


Figure 5.9: Accuracy comparison for best three models

From the perspective of the other assessment measures, the values for each of the models are greater than 0.75, which meets the acceptance criteria for being a viable prediction model in this study.

5.4.4. Experiments for the ISTN2IP dataset

For these datasets, two additional attributes were added to the dataset, these being the student’s ISTN101 and ISTN103 symbol. A value of “NA” is given in the event that the student has no mark for these courses. This would occur in the event that the student had done an equivalent course, was given an exemption, or did not do the course for some other reason.

5.4.4.1. Experiment-2IP-Sampling [None]

The DT algorithm was not able to generate a viable model for this dataset. Only a one leaf tree (“P”) was generated, indicating that the model adopted a guessing approach. For the RF algorithm, the accuracy obtained was similar to that of the model generated using the DT algorithm. The PRC and F-measure values were much improved for the RF model (Table 5.31).

Table 5.31: Summary analysis for RF and DT generated models – *Experiment-2IP-Sampling [None]-VAR1*

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	80	91,6	0,64	0,79	0,75	0,89	0,78	0,92	0,8	0,91	0,72	0,88
Backward Search	23	79	85,3	0,6	0,59	0,73	0,86	0,74	0,84	0,79	0,85	0,74	0,85

By including Moodle interaction data (VAR2), there appeared to be an improvement in the models developed by the two algorithms. This is evident by viewing not only the accuracy, but the PRC and F-Measure values (see Table 5.32). For both forward search algorithms, the viability of the models cannot be accepted as the precision and F-measure values could not be calculated.

Table 5.32: Summary analysis for RF and DT generated models – *Experiment-2IP-Sampling [None]-VAR2*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	84	90,9	0,64	0,5	0,77	0,83	0,84	?	0,81	0,9	0,84	?
Backward Search	2	84,9	92,3	0,65	0,61	0,79	0,87	0,83	0,92	0,84	0,92	0,81	0,89

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	82,6	90,9	0,61	0,76	0,75	0,88	0,81	?	0,82	0,9	0,75	?
Backward Search	28	85,4	90,9	0,8	0,87	0,86	0,94	0,84	0,87	0,85	0,9	0,82	0,87

When analyzing the algorithms applied to the COVID-19 dataset (VAR3), the RF generated models performed better than the DT generated models in terms of closeness between the accuracy of the training dataset (10-fold) and the validation dataset. The PRC, precision, recall and F-measure values from both algorithms fall within the acceptance criteria.

Table 5.33: Summary analysis for RF and DT generated models – *Experiment-2IP-Sampling [None]-VAR3*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	38	94,4	85,3	0,7	0,59	0,9	0,85	0,94	0,96	0,94	0,85	0,94	0,86
Backward Search	3	95,8	79,7	0,79	0,59	0,92	0,85	0,95	0,88	0,95	0,79	0,95	0,83
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
Backward Search	14	94,4	88,1	0,78	0,66	0,93	0,87	0,94	0,82	0,94	0,88	0,92	0,85

5.4.4.2. *Experiment-2IP-Sampling [US]*

For VAR1 and VAR2, the DT and RF algorithms developed models with an unacceptable accuracy (less than 80% accuracy observed during training of the algorithms). The model produced using the VAR3 dataset did produce acceptable accuracy (90.9%) when applied to the validation data, however, the difference between the training and validation accuracy is more than 20%, indicating that the model cannot be guaranteed to perform well against unseen data. For VAR3, the models developed are suitable, specifically for the training data (overfitting), with almost thirty percent difference in accuracy between application of the training data and the validation data.

5.4.4.3. *Experiment-2IP-Sampling [OS]*

Table 5.34 shows the analysis of the models generated using the learning algorithms when applied to the oversampled VAR1 dataset. For this experiment, the `sampleSizePercent` parameter value is 158.

Table 5.34: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [OS]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	87,9	72	0,94	0,52	0,93	0,84	0,88	0,84	0,88	0,72	0,88	0,77
Backward Search	19	89,5	74,8	0,94	0,6	0,93	0,85	0,89	0,88	0,89	0,74	0,89	0,79
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	20	95,3	79	0,99	0,7	0,99	0,89	0,95	0,88	0,95	0,79	0,95	0,82
Backward Search	23	96,3	80,4	0,99	0,69	0,99	0,89	0,96	0,86	0,96	0,8	0,96	0,83

The accuracy obtained when training the dataset exceeds 87% for the DT algorithm and exceeds 95% for the RF algorithm. However, the accuracy obtained when the respective models are applied to the validation dataset was at least 15% less than that of the training accuracy. By observing the recall value, it appeared that the algorithm could not produce consistent accuracy when applied to unseen data.

In Table 5.35, when including the Moodle interaction data (VAR2), an improvement was noted in terms of accuracy difference when compared to the analysis for Experiment-2IP-Sampling [OS]-VAR1. In the case of the VAR2 dataset (using `sampleSizePercent` value of 164), the RF algorithm generated acceptable prediction models while the DT algorithm was not able to.

Table 5.35: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [OS]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	88,3	74,8	0,89	0,68	0,85	0,88	0,88	0,87	0,88	0,74	0,88	0,79
Backward Search	13	82,8	79	0,84	0,68	0,8	0,88	0,83	0,85	0,82	0,79	0,82	0,81
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	12	97,6	88,8	0,99	0,75	0,98	0,9	0,97	0,88	0,97	0,88	0,97	0,88
Backward Search	24	97,2	89,5	0,99	0,89	0,99	0,93	0,97	0,89	0,97	0,89	0,97	0,89

For the COVID-19 dataset (VAR3), it appears that the generated model overfits on the training dataset, resulting in poorer accuracy when the model is applied to the validation dataset. For this course, the pass rate was much higher for the years 2020 and 2021 than when the course was run before COVID-19. A greater number of instances with improved diversity with respect to attribute values may assist in improving the performances of the algorithms should this course continue with an online model in the future.

5.4.4.4. Experiment-2IP-Sampling [SMOTE]

When SMOTE is applied to the VAR1 dataset (using a Percentage parameter value of 275), the resultant training accuracy is not as good as when oversampling was used. The accuracy between training and validation is, however, much closer than when compared to the oversampled variations. Valid models are found for the backward search DT and RF models respectively.

Table 5.36: Summary analysis for RF and DT generated models – Experiment-2IP-Sampling [SMOTE]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	86,3	76,9	0,85	0,57	0,81	0,86	0,86	0,84	0,86	0,76	0,86	0,8
Backward Search	16	82,6	81,8	0,86	0,77	0,82	0,89	0,82	0,89	0,82	0,81	0,82	0,84
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	85,7	65,7	0,89	0,63	0,87	0,86	0,85	0,87	0,85	0,65	0,85	0,73
Backward Search	22	85,9	81,8	0,92	0,65	0,92	0,86	0,86	0,86	0,86	0,81	0,86	0,84

For VAR2, accuracy for training was in an acceptable range of 85.9% to 90.3% for each of the learning algorithms (DT and RF). However only the model generated using the RF algorithm (backward search) produced similar accuracies for both training (90.3%) and validation datasets (87.4%). The ROC (0.96 and 0.82), PRC (0.96 and 0.91), F-measure (0.9 and 0.87), precision (0.9 and 0.87) and recall values (0.9 and 0.87) were also close, indicating that the model could be used to predict performance for unseen instances.

The performance of the algorithms when applied to VAR3 were similar to that of oversampling, where training of the algorithm resulted in very high accuracies that were not replicated when the models were applied to the validation dataset.

5.4.4.5. Analysis of experiments conducted

The best models identified from each of sections 5.4.4.1 to 5.4.4.4 are listed in Table 5.37. As with previous datasets, the use of undersampling did not result in the production of any acceptable models. All models listed in Table 5.37 were developed using the RF algorithm and are named VAR3-None, VAR2-OS and VAR3-SMOTE respectively.

Table 5.37: Best three models generated from *Experiment-2IP*

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR3-None	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
VAR2-OS	97,2	89,5	0,99	0,89	0,99	0,93	0,97	0,89	0,97	0,89	0,97	0,89
VAR2-SMOTE	90,3	87,4	0,96	0,82	0,96	0,91	0,9	0,87	0,9	0,87	0,9	0,87

The accuracy comparison for the three models is shown in Figure 5.10. The VAR2-OS model produced the highest accuracy for both training (10-fold) and validation but the VAR2-SMOTE model produced the smallest accuracy difference.

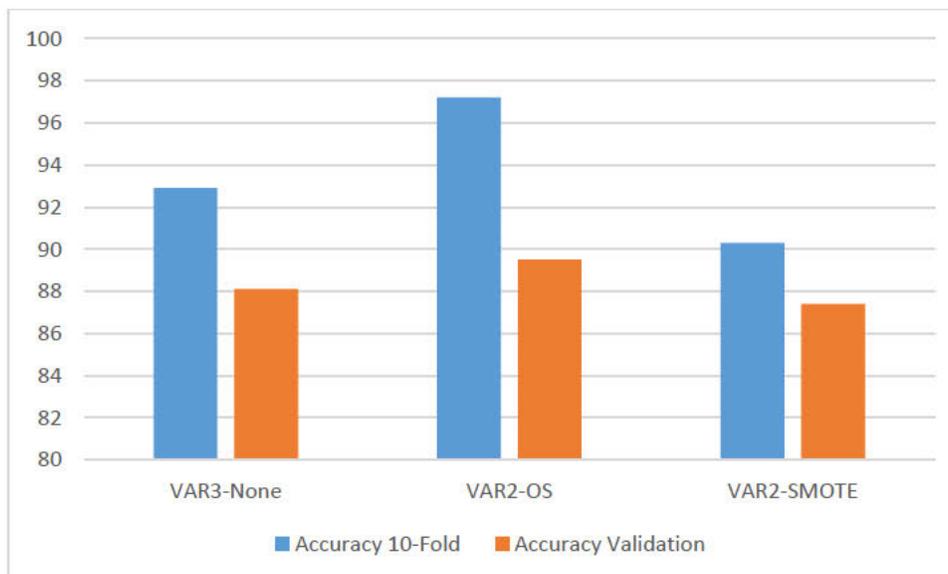


Figure 5.10: Accuracy comparison for best three models

A comparison of the remaining assessment measures is illustrated using the bar graph in Figure 5.11. As expected, the ROC values for VAR3-None were lower due to ROC being affected by imbalanced datasets. These are much improved when sampling is applied in the case of VAR2-OS and VAR2-SMOTE. The other performance measures are all above 0.8, indicating the reliability of these models.

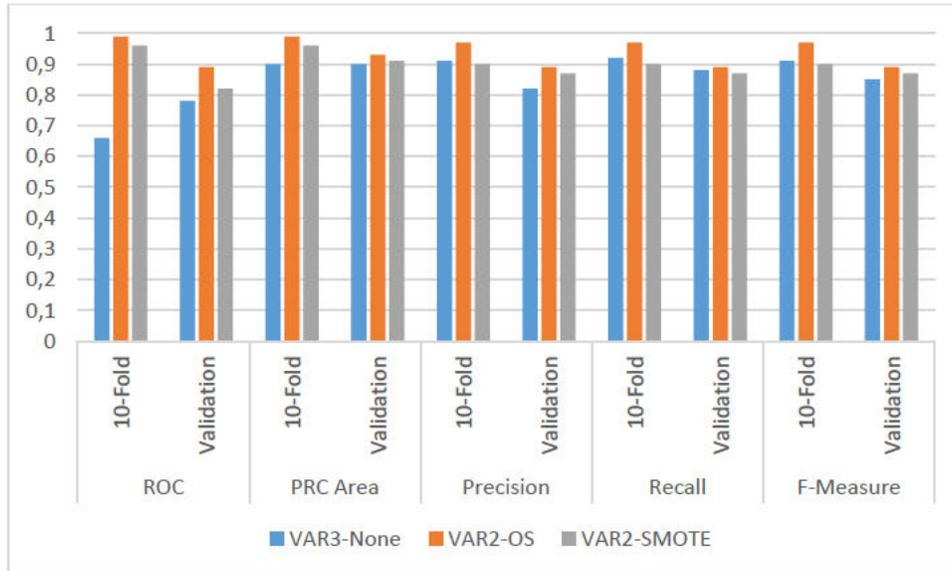


Figure 5.11: Performance measure comparison of three best models

5.4.5. Experiments for the ISTN211 dataset

This section covers experiments for the ISTN211 dataset, which has the highest imbalance score as per Table 5.2.

5.4.5.1. Experiment-211-Sampling [None]

With no sampling applied, the DT algorithm was not able to produce any viable models for all three dataset variations. The RF algorithm was able to produce acceptable prediction models as seen in Table 5.38.

Table 5.38: Summary analysis for RF generated models – *Experiment-211-Sampling [None]-VAR1 and Experiment-211-Sampling [None]-VAR2 and Experiment-211-Sampling [None]-VAR3*

Random Forest													
Variation 1	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	95,5	94,4	0,7	0,72	0,93	0,92	0,94	?	0,95	0,94	0,93	?
Backward Search	25	95,4	93,5	0,71	0,54	0,94	0,9	0,94	0,89	0,95	0,93	0,94	0,91
Random Forest													
Variation 2	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	95,5	94,1	0,71	0,57	0,93	0,9	0,95	0,89	0,95	0,94	0,93	0,91
Backward Search	18	95,5	93,5	0,59	0,47	0,92	0,89	0,95	0,89	0,95	0,93	0,93	0,91
Random Forest													
Variation 3	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	97	91,1	0,72	0,63	0,95	0,91	0,97	0,89	0,97	0,91	0,96	0,9
Backward Search	7	97,4	95,4	0,82	0,75	0,96	0,94	0,97	0,95	0,97	0,95	0,96	0,93

As seen in Table 5.38, the accuracies between training and validation are similar, as are the precision, recall, PRC and F-Measure values. The only exception is the forward search RF algorithm (applied to VAR1) where the precision and F-measure could not be calculated. As stated earlier, the ROC values cannot be relied upon in the case of these imbalanced variations.

5.4.5.2. *Experiment-211-Sampling [US]*

Both learning algorithms (DT and RF) performed poorly in generating valid models when undersampling was applied to any of the three variations. For VAR3, training accuracy achieved was 88.8% and 94.4% for the models generated using the DT and RF algorithms respectively. However, when either of these models were applied to the validation datasets, accuracy of less than 43% was obtained.

5.4.5.3. *Experiment-211-Sampling [OS]*

When oversampling is applied to the datasets, only the forward search DT algorithm when applied to VAR1 resulted in the generation of an acceptable model. In the case of the reverse search DT algorithm, the accuracy difference was greater than 10% (97.6% - 83.7%) and thus the model does not fit the acceptance criteria.

For VAR2 and VAR3, the training accuracies are above the acceptable range of 98% for the models generated by both the learning algorithms.

5.4.5.4. Experiment-211-Sampling [SMOTE]

The performance measures of the algorithms when SMOTE is applied to each of the dataset variations are similar to that of when oversampling was used. The accuracy obtained during training ranged above 96% for all variations. Table 5.39 shows the algorithms applied to VAR1 with SMOTE applied. For the experiments listed, the Percentage parameter value was set to 1900 to ensure the creation of sufficient instances for the minority class to match the number of instances from the majority class.

Table 5.39: Summary analysis for RF and DT generated models – Experiment-211-Sampling [SMOTE]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	96,9	92,6	0,97	0,5	0,96	0,89	0,96	0,9	0,96	0,92	0,96	0,91
Backward Search	22	95,6	93,2	0,97	0,63	0,96	0,91	0,95	0,91	0,95	0,93	0,95	0,92
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	96,9	91,7	0,98	0,52	0,98	0,89	0,97	0,89	0,97	0,91	0,97	0,9
Backward Search	24	97,1	92,3	0,98	0,53	0,98	0,89	0,97	0,89	0,97	0,92	0,97	0,9

The accuracy values for VAR1 were within the acceptable range for both training and validation. Further, the PRC, precision, recall and F-measure values also fall within the acceptable range targeted for this study. The only exception is the ROC value for validation, where it appears that, for the validation dataset, the model predictions are similar to that of guesswork.

Table 5.40 shows the analysis of the prediction models for VAR2. For all four experiments, a Percentage parameter value of 1750 was used and a training accuracy was in the range 96% to 97.2%. The forward search RF algorithm generated a model that did not fit for the validation data while the other algorithms yielded similar accuracy to training.

Table 5.40: Summary analysis for RF and DT generated models – Experiment-211-Sampling [SMOTE]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	97	93	0,97	0,46	0,96	0,89	0,97	0,89	0,97	0,93	0,97	0,91
Backward Search	12	97	90	0,97	0,69	0,97	0,91	0,97	0,89	0,97	0,9	0,97	0,89
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	96,7	56,8	0,97	0,46	0,97	0,89	0,96	0,89	0,96	0,56	0,96	0,68
Backward Search	27	97,2	92,6	0,99	0,62	0,99	0,92	0,97	0,89	0,97	0,92	0,97	0,9

As with the analysis for VAR1, the PRC, precision, recall and F-measure values all fall within an acceptable range for both training and validation data. In terms of ROC validation, the model generated using the backward search DT algorithm was closest to falling in an acceptable range (0.69). For VAR3, the analysis is listed in Table 5.41. The Percentage parameter was set to a value of 2500 for these experiments.

Table 5.41: Summary analysis for RF and DT generated models – Experiment-211-Sampling [SMOTE]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	97,6	91,7	0,98	0,38	0,97	0,88	0,97	0,89	0,97	0,91	0,97	0,9
Backward Search	4	97,8	90,2	0,97	0,45	0,96	0,88	0,97	0,9	0,97	0,9	0,97	0,9
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	11	98,4	92,6	0,98	0,57	0,98	0,9	0,98	0,89	0,98	0,92	0,98	0,9
Backward Search	36	98,4	92,3	0,99	0,64	0,99	0,92	0,98	0,89	0,98	0,92	0,98	0,9

For VAR3, the accuracy achieved for the training dataset is greater than 98% and thus these models are not considered valid for this study. The assessment measures for the models generated using the DT algorithm indicate that the models are reliable. However, as with the VAR1 and VAR2 variants, the ROC values for the model when applied to the validation dataset are less than 0.5, possibly indicating an unreliable model when applied to unseen data.

5.4.5.5. Analysis of experiments conducted

In this section, three (3) models, one from each of Sections 5.4.5.1 to 5.4.5.4 (excluding Section 5.4.5.2) are compared and discussed in terms of how well they have met the assessment criteria requirements for this study. For undersampling (Section 5.4.5.2), no models were generated that met the acceptance criteria. The assessment measures for the three models are shown in Table 5.42. The model named *VAR3-None* was generated using the RF algorithm while the *VAR1-OS* and the *VAR2-SMOTE* were generated using the DT algorithm.

Table 5.42: Best three models generated from Experiment-211

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR3-None	97,4	95,4	0,82	0,75	0,96	0,94	0,97	0,95	0,97	0,95	0,96	0,93
VAR1-OS	96,6	87,4	0,99	0,36	0,99	0,88	0,96	0,88	0,96	0,87	0,96	0,88
VAR2-SMOTE	97	90	0,97	0,69	0,97	0,91	0,97	0,89	0,97	0,9	0,97	0,89

The accuracy for each of the models are all of an acceptable value, with training accuracy in the range 96.6% to 97.4% while validation accuracy ranges from 87.4% to 95.4%. Figure 5.12 shows that the VAR3-None model has the closest accuracy difference.



Figure 5.12: Accuracy comparison for best three models

The remaining assessment measures (with the exception of ROC) all indicate that the models generated are reliable (Figure 5.13). The ROC values are particularly low for the validation data. This is to be expected as the ROC assessment metric behaves abnormally with imbalanced data.

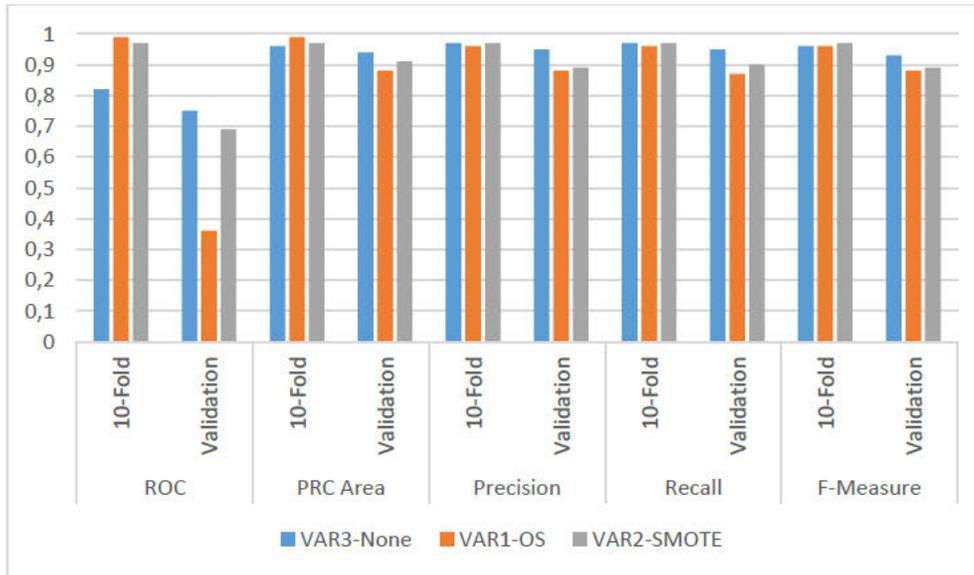


Figure 5.13: Performance measures of best three models

5.4.6. Experiments for the ISTN212 dataset

In this section, experiments are conducted for the ISTN212 dataset.

5.4.6.1. Experiment-212-Sampling [None]

When applying the learning algorithms to the VAR1 dataset with no sampling, acceptable performances were obtained for the DT algorithm (using backward search) and RF algorithm (using forward search). For these algorithms, the accuracies obtained are greater than 80% and the accuracy from applying the model to unseen data is 90.2% (DT) and 88.5% (RF), a difference of less than 10%. The PRC, precision, recall and F-measure for these models are also within the acceptable range, although there is an increase in difference between training and validation values when compared to previous experiments.

Table 5.43: Summary analysis for RF and DT generated models – Experiment-212-Sampling [None]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	80.7	95,6	0,49	0,5	0,68	0,91	?	?	0,8	0,95	?	?
Backward Search	22	81.9	90,2	0,69	0,68	0,77	0,93	0,79	0,93	0,82	0,9	0,8	0,91
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	82,4	88,5	0,75	0,61	0,81	0,92	0,81	0,92	0,82	0,88	0,81	0,9
Backward Search	23	80,6	94,6	0,71	0,76	0,8	0,94	0,77	0,92	0,8	0,94	0,78	0,93

For the VAR2 dataset, performance measures obtained for the learning algorithms are better than that obtained using VAR1 dataset (Table 5.44).

Table 5.44: Summary analysis for RF and DT generated models – Experiment-212-Sampling [None]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	89,4	96,6	0,64	0,61	0,83	0,93	0,89	0,96	0,89	0,96	0,86	0,95
Backward Search	7	89	96,6	0,63	0,61	0,82	0,93	0,88	0,96	0,89	0,96	0,86	0,95
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	12	91,2	95,6	0,85	0,89	0,92	0,97	0,9	0,94	0,91	0,95	0,9	0,94
Backward Search	18	89,8	95,9	0,86	0,94	0,92	0,97	0,88	0,95	0,89	0,96	0,88	0,94

As can be seen in Table 5.44, accuracy achieved using either of the algorithms was above 89%. Furthermore, when the generated models were applied to the validation dataset, accuracy of 95% to 96% was achieved. The other performance measures also indicated that the models generated were of a good quality in making predictions for any other unseen instances. The same can be said for VAR3 (see Table 5.45) where good accuracies were obtained by both algorithms for training as well as for unseen instances (validation dataset).

Table 5.45: Summary analysis for RF and DT generated models – Experiment-212-Sampling [None]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	97,9	94,2	0,68	0,73	0,95	0,94	0,98	0,95	0,98	0,94	0,97	0,94
Backward Search	16	97,1	89,2	0,69	0,73	0,95	0,93	0,96	0,93	0,97	0,89	0,96	0,91
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	97,1	96,6	0,58	0,82	0,94	0,96	0,97	0,96	0,97	0,96	0,96	0,95
Backward Search	9	97,1	95,9	0,73	0,8	0,95	0,96	0,97	0,96	0,97	0,96	0,96	0,96

5.4.6.2. Experiment-212-Sampling [US]

For the VAR1 and the VAR3 datasets, no models were generated that fit the acceptance criteria for the study. The analysis for VAR2 is shown in Table 5.46. Three out of the four experiments resulted in acceptable models with the accuracy of the RF generated models being better than that of the DT generated models. The remaining performance measures for the RF generated model were also in the acceptable range.

Table 5.46: Summary analysis for RF and DT generated models – Experiment-212-Sampling [US]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	80,1	87,2	0,79	0,71	0,77	0,93	0,82	0,94	0,8	0,87	0,79	0,9
Backward Search	14	79,5	91,2	0,79	0,85	0,75	0,95	0,79	0,93	0,79	0,91	0,79	0,92
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	10	84,1	92,5	0,9	0,89	0,9	0,96	0,84	0,95	0,84	0,92	0,84	0,93
Backward Search	10	83,6	92,2	0,89	0,92	0,89	0,96	0,83	0,95	0,83	0,92	0,83	0,93

5.4.6.3. Experiment-212-Sampling [OS]

For this section, oversampling was applied to the dataset variations, followed by the application of the learning algorithms. The performance measures for the resultant models when using the

VAR1 dataset are shown in Table 5.47. For these experiments, the `sampleSizePercent` parameter value was set to 161.

Table 5.47: Summary analysis for RF and DT generated models – Experiment-212-Sampling [OS]-VAR1

Decision Tree (148)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	17	87,6	76,7	0,91	0,67	0,89	0,93	0,88	0,93	0,87	0,76	0,87	0,83
Backward Search	17	87,6	72	0,92	0,54	0,9	0,91	0,87	0,93	0,87	0,72	0,87	0,8
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	94,5	77,4	0,98	0,68	0,98	0,94	0,94	0,93	0,94	0,77	0,94	0,83
Backward Search	21	94,5	82,1	0,98	0,65	0,98	0,93	0,94	0,92	0,94	0,82	0,94	0,86

As seen above, the accuracy for the oversampled VAR1 training dataset is 87.6% and 94.5% for the DT algorithms and RF algorithms respectively. However, it appears that the model overfits the training data as seen by the accuracy when the models are applied to the validation dataset. The models generated when the algorithms are applied to oversampled VAR2 are far closer in terms of accuracy, as shown in the performance measures in Table 5.48. For these experiments, the `sampleSizePercent` parameter value was 173.

In the case of the RF algorithm, the training accuracy falls outside the acceptable range for this study (greater than 98% accuracy). For the DT algorithm (backward search), the generated model accuracies are acceptable in terms of being in an acceptable range as well as the accuracy difference being less than 10%.

Table 5.48: Summary analysis for RF and DT generated models – *Experiment-212-Sampling [OS]-VAR2*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	94,9	84,8	0,97	0,54	0,97	0,92	0,95	0,92	0,94	0,84	0,94	0,88
Backward Search	17	96,2	90,9	0,97	0,66	0,97	0,93	0,96	0,93	0,96	0,9	0,96	0,92
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	98,3	96,6	0,99	0,88	0,99	0,97	0,98	0,96	0,98	0,96	0,98	0,95
Backward Search	23	98,8	95,6	1	0,93	1	0,97	0,98	0,94	0,98	0,95	0,98	0,94

For VAR3, the algorithms exhibited near perfect accuracy (above 98%) and thus above the acceptable range for this study. The generated models, when applied to the validation dataset, also resulted in high accuracy (in the range of 91% to 95%). It should be noted, however, that the pass rate for VAR3 was very high and thus further instances should be obtained for both training and testing datasets to better evaluate how the learning algorithms predict performance for this course.

5.4.6.4. *Experiment-212-Sampling [SMOTE]*

When SMOTE is applied to the VAR1 dataset (using a Percentage parameter value of 315), the learning algorithms produced accuracy in the range of 84% to 87%. The resultant prediction models were applied to the validation dataset and with the exception of the DT backward search model, the accuracy obtained was very close to the training dataset. The ROC, PRC and F-Measure values were also high, indicating that good models have been generated.

Table 5.49: Summary analysis for RF and DT generated models – *Experiment-212-Sampling [SMOTE]-VAR1*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	84,1	90,5	0,87	0,8	0,84	0,94	0,84	0,93	0,84	0,9	0,84	0,92
Backward Search	20	86,1	63,9	0,88	0,52	0,85	0,91	0,86	0,91	0,86	0,64	0,86	0,74
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	86,4	89,8	0,93	0,77	0,93	0,94	0,86	0,94	0,86	0,89	0,86	0,91
Backward Search	23	87,4	90,2	0,94	0,69	0,93	0,94	0,87	0,92	0,87	0,9	0,87	0,91

The performance of the algorithms when applied to VAR2 (Percentage parameter value was 550) appeared to be better than when compared to VAR1 as seen in Table 5.50. The accuracy achieved from training ranges from 89.4% to 94.1%, and when the models are applied to the unseen dataset, equivalent accuracy is achieved. The only exception is the forward search RF algorithm where overfitting was observed.

Table 5.50: Summary analysis for RF and DT generated models – *Experiment-212-Sampling [SMOTE]-VAR2*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	89,4	89,5	0,92	0,64	0,91	0,93	0,89	0,93	0,89	0,89	0,89	0,91
Backward Search	14	93,1	93,9	0,93	0,79	0,9	0,94	0,93	0,92	0,93	0,93	0,93	0,93
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	90,6	51,8	0,92	0,62	0,9	0,93	0,9	0,93	0,9	0,51	0,9	0,64
Backward Search	22	94,1	95,2	0,98	0,91	0,98	0,96	0,94	0,94	0,94	0,95	0,94	0,94

The application of the algorithms to VAR3 yielded almost perfect accuracy (above 98%). Similar to the conclusion found with oversampling, further instances should be collected to gain a better understanding of how the algorithms perform with the attributes of the dataset.

5.4.6.5. Analysis of experiments conducted

Unlike previous courses, four models were identified (one from each of the sections 5.4.6.1 to 5.4.6.4) as the best models and are listed in Table 5.51.

Table 5.51: Best four models generated from Experiment-212

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR3-None	97,1	95,9	0,73	0,8	0,95	0,96	0,97	0,96	0,97	0,96	0,96	0,96
VAR2-US	84,1	92,5	0,9	0,89	0,9	0,96	0,84	0,95	0,84	0,92	0,84	0,93
VAR2-OS	96,2	90,9	0,97	0,66	0,97	0,93	0,96	0,93	0,96	0,9	0,96	0,92
VAR2-SMOTE	94,1	95,2	0,98	0,91	0,98	0,96	0,94	0,94	0,94	0,95	0,94	0,94

All accuracies (training and validation) fall within the acceptable range for this study. A comparison of the accuracies is shown in Figure 5.14. The *VAR3-None* model appears to be the best model in terms having the highest accuracy for both training and validation. The *VAR2-SMOTE* model has the smallest accuracy difference of 1.1 between training and validation. The *VAR3-None* model has an accuracy difference of 1.2.

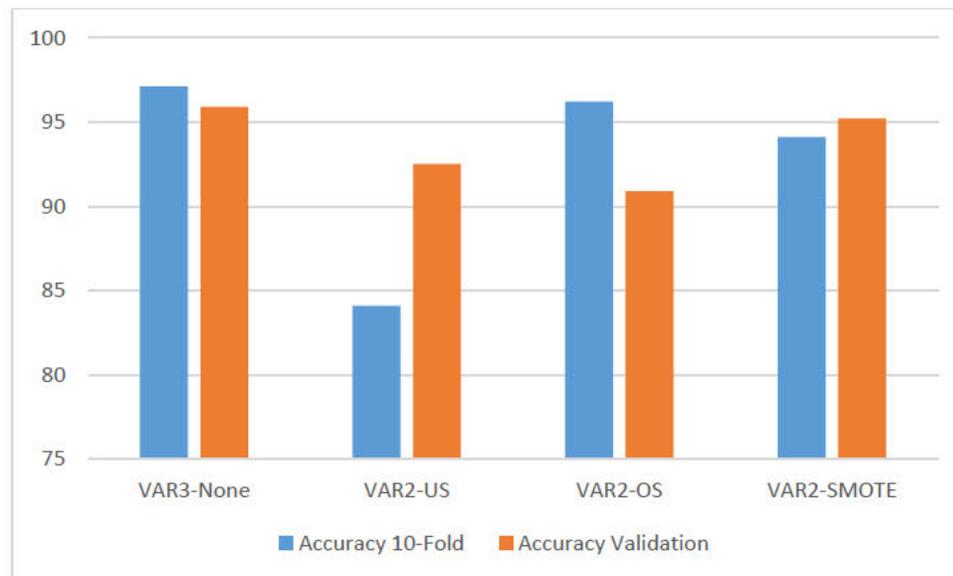


Figure 5.14: Accuracy comparison of four best models

A comparison of the remaining performance measures is shown in Figure 5.15. As with the other course experiments, the PRC, precision, recall and F-measure values are all within the acceptable range. For the ROC, only the *VAR2-OS* model has a ROC value less than the acceptable value of

0.7. It should be noted, however, that with the validation dataset being imbalanced (21.85 according to Table 5.2), it is expected that the ROC value would be adversely affected.

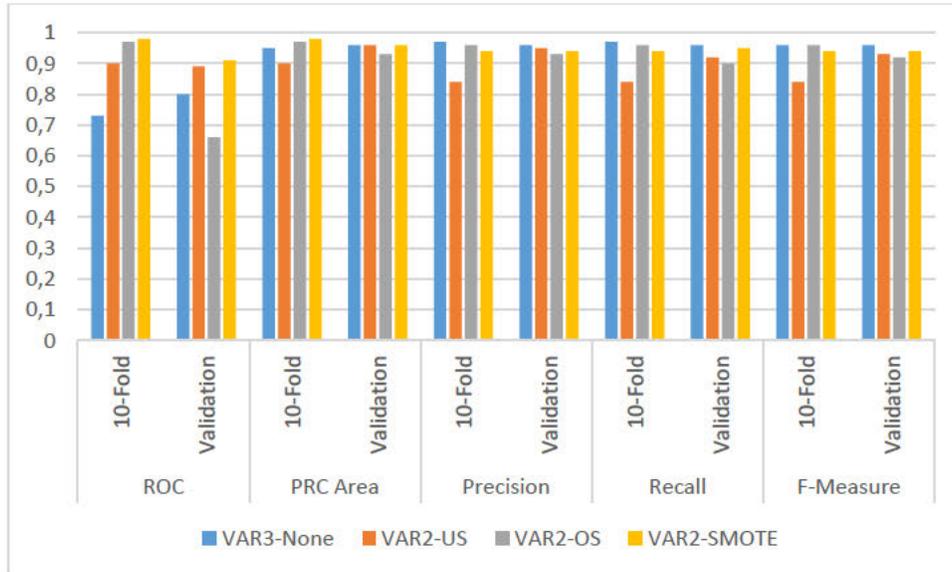


Figure 5.15: Performance measures of best four models

5.4.7. Experiments for the ISTN3SA dataset

This section covers the experiments for the learning algorithms applied to the ISTN3SA dataset.

5.4.7.1. Experiment-3SA-Sampling [None]

The DT algorithm could not produce viable models when applied to any of the variations, i.e., single leaf (“P”) models were generated. For VAR1, the RF algorithm was able to produce acceptable models (Table 5.52).

Table 5.52: Summary analysis for RF generated models – Experiment-3SA-Sampling [None]-VAR1

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	91,7	97,8	0,64	0,77	0,87	0,97	0,9	0,96	0,91	0,97	0,88	0,97
Backward Search	19	90,3	96,9	0,66	0,67	0,88	0,97	0,88	0,97	0,9	0,97	0,89	0,97

The RF algorithm could not produce viable models for VAR2 and VAR3. In the case of VAR2, out of the 57 failed class instances, the RF algorithm was only able to correctly predict 7 instances,

thus resulting in the poor performance of the model when applied to the validation dataset. In the case of VAR3, poor performance is due to the pass rate of 99% and only 2 failures in the course, respectively. With only two instances resulting in a fail class value, more instances are required to better study the inclusion of Moodle data, demographics and registration data for predictive purposes (for this course).

5.4.7.2. Experiment-3SA-Sampling [US]

When undersampling is applied to VAR1 and VAR2, poor training accuracy of less than 80% is obtained for both algorithms. During the year 2020 (online teaching due to COVID-19), only two (2) students had failed. Undersampling removes instances from the major class to match the number of instances of the minor class. As WEKA requires a minimum of 10 instances per class in order to train the algorithms, no analysis was performed for VAR3.

5.4.7.3. Experiment-3SA-Sampling [OS]

Table 5.53 shows the performance measures for the models generated by the DT algorithm and RF algorithm (the `sampleSizePercent` parameter value was set to 182) when each were applied to the oversampled VAR1 dataset. The training accuracy achieved was above 92% and both algorithms produced models that manage to achieve similar accuracy for the validation dataset. While the performance of the models can be verified by the PRC and F-measure values, only the forward search RF algorithm (forward search) produced acceptable ROC values.

Table 5.53: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [OS]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	92	91,7	0,97	0,49	0,97	0,96	0,92	0,97	0,92	0,91	0,92	0,94
Backward Search	18	94,8	73	0,96	0,54	0,96	0,96	0,95	0,96	0,94	0,73	0,94	0,82
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	97,6	90,8	0,99	0,73	0,99	0,98	0,97	0,97	0,97	0,9	0,97	0,93
Backward Search	24	98,2	96,5	0,99	0,58	0,99	0,97	0,98	0,97	0,98	0,96	0,98	0,96

For the performance measures of the models for Experiment-3SA-Sampling [OS]-VAR2, the RF algorithm produced about 98% training accuracy while the DT algorithm produced models with about 95.8% and 96.8% training accuracy (sampleSizePercent value is 178). The models, when applied to the validation dataset, also produced acceptable accuracy of 86% and 93% respectively (see Table 5.54).

Table 5.54: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [OS]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	95,8	93	0,97	0,82	0,96	0,98	0,96	0,97	0,95	0,93	0,95	0,94
Backward Search	12	96,8	86	0,97	0,57	0,97	0,96	0,97	0,96	0,96	0,86	0,96	0,91
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	13	98,8	96,9	1	0,71	1	0,97	0,98	0,96	0,98	0,97	0,98	0,96
Backward Search	17	98,7	96,5	1	0,79	1	0,98	0,98	0,97	0,98	0,96	0,98	0,96

Both algorithms' resultant models produced 100% prediction when training on oversampled VAR3. As stated earlier, with only 2 fail class instances, there is insufficient variety with number of failures. Acquiring more of this course data in the future will allow for more failing instances required in order for training to be more effective and understanding how well the algorithms perform on this 3rd year course dataset.

5.4.7.4. Experiment-3SA-Sampling [SMOTE]

For the VAR1 dataset with SMOTE applied (a Percentage parameter value of 900 was used), the algorithms were able to produce models with training accuracy of 92% to 94% (Table 5.55). With regard to the validation data, the accuracy of the models generated using the DT algorithms as well as the backward search RF were within 10% of the training accuracies. The ROC, PRC and F-measures also indicate that the models are reliable for making predictions for unseen instances.

Table 5.55: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [SMOTE]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	92	87,3	0,93	0,52	0,91	0,96	0,92	0,96	0,92	0,87	0,92	0,91
Backward Search	12	92,7	93,9	0,97	0,7	0,96	0,97	0,92	0,97	0,92	0,93	0,92	0,95
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	92	79,1	0,94	0,6	0,93	0,97	0,92	0,96	0,92	0,79	0,92	0,86
Backward Search	26	94	96	0,97	0,75	0,96	0,98	0,94	0,97	0,94	0,96	0,94	0,96

For VAR2, the performance of the algorithms appears better than that of VAR1. For all experiments in this case (Percentage parameter value was set to 725), training accuracy of above 91% was achieved (see Table 5.56). When the models are applied to the validation datasets, equivalent accuracy is produced.

Table 5.56: Summary analysis for RF and DT generated models – Experiment-3SA-Sampling [SMOTE]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	91,8	96	0,9	0,58	0,87	0,97	0,91	0,97	0,91	0,96	0,91	0,96
Backward Search	17	91,7	95,2	0,93	0,7	0,9	0,97	0,91	0,97	0,91	0,95	0,91	0,96
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	27	93,9	96,9	0,97	0,77	0,97	0,98	0,94	0,97	0,93	0,97	0,93	0,97
Backward Search	23	94,2	96,9	0,97	0,78	0,97	0,98	0,94	0,97	0,94	0,97	0,94	0,97

In the case of VAR3, accuracy neared 100%. As observed in Sections 5.4.7.1 (no sampling) and 5.4.7.3 (oversampling), an insufficient number of failures in this dataset contributes to the 100% accuracy. Acquiring more failing instances in future iterations of this course is required in order for training to be more effective.

5.4.7.5. Analysis of experiments conducted

From the experiments described in Sections 5.4.7.1 to 5.4.7.4, one model was respectively chosen from experiments relating to no sampling, oversampling and SMOTE. Due to poor performance of the models, no models were chosen from the undersampling experiments (Section 5.4.7.2). The models' assessment measures are shown in Table 5.57.

Table 5.57: Best three models from Experiment-3SA

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR1-None	91,7	97,8	0,64	0,77	0,87	0,97	0,9	0,96	0,91	0,97	0,88	0,97
VAR2-OS	95,8	93	0,97	0,82	0,96	0,98	0,96	0,97	0,95	0,93	0,95	0,94
VAR2-SMOTE	94,2	96,9	0,97	0,78	0,97	0,98	0,94	0,97	0,94	0,97	0,94	0,97

All training accuracy values obtained are greater than 90% with the *VAR1-None* model having the best accuracy when applied to the validation dataset. The *VAR2-OS* and *VAR2-SMOTE* models have the smallest accuracy difference between training and validation accuracy.

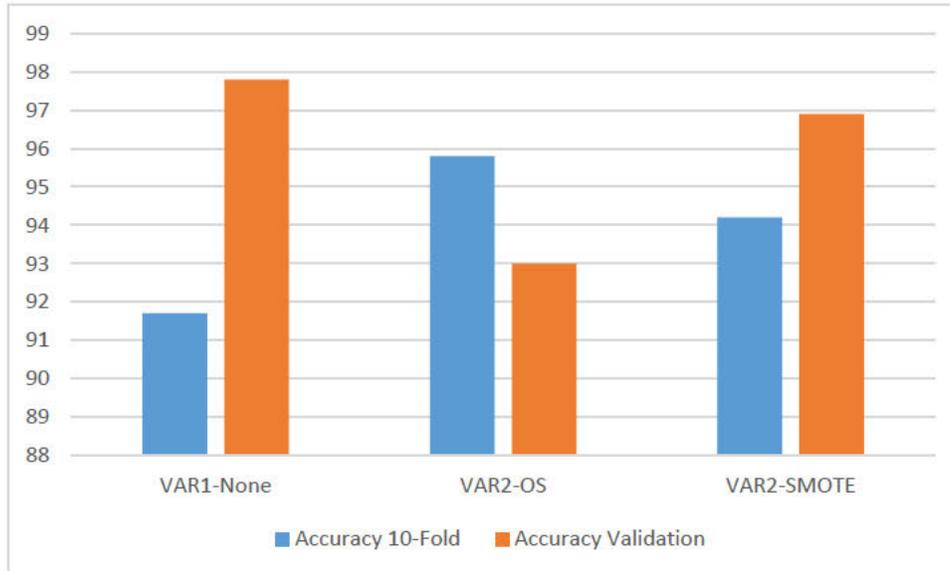


Figure 5.16: Accuracy comparison for best three models

The remaining assessment measures are illustrated in Figure 5.17. All the measures are within the acceptance criteria range for this study.

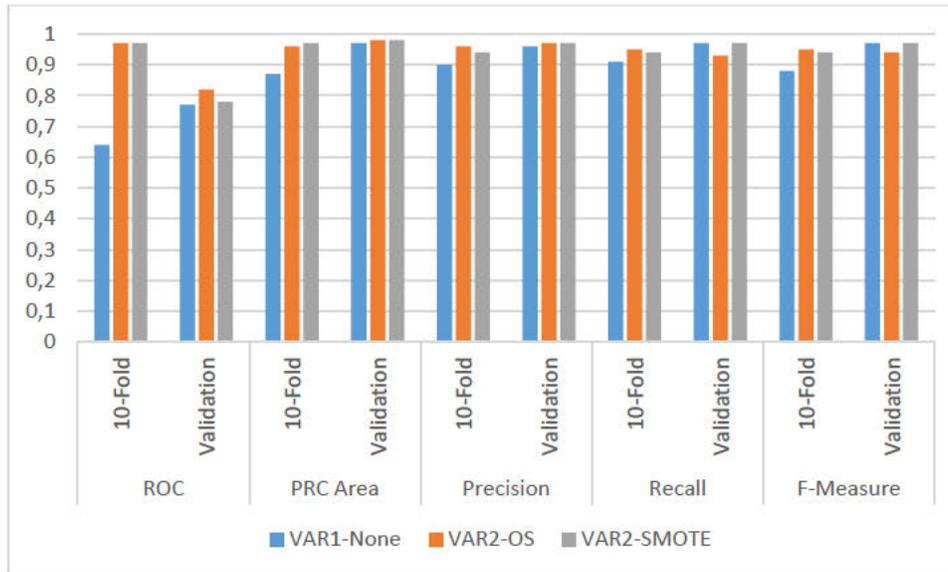


Figure 5.17: Assessment measure comparison of three best models

5.4.8. Experiments for the ISTN3AS dataset

This section covers experiments relating to the ISTN3AS dataset.

5.4.8.1. Experiment-3AS-Sampling [None]

The performance of the algorithms was poor when no sampling is applied. This is due to the imbalance of the dataset. In this case, an imbalance value of 45.83 for all instances (see Table 5.2) is the highest imbalance value for all courses covered in this study. With fewer failing instances, the algorithms overfit and show bias towards the major class, thus resulting in the poor performance of the generated models.

5.4.8.2. Experiment-3AS-Sampling [US]

For VAR1 and VAR2, training accuracy for each of the algorithms ranged from 65% to 90%. However, none of the models, when applied to the validation data, achieved an acceptable accuracy (55% to 78% range). For VAR3, the number of total instances after undersampling was insufficient for WEKA to conduct training and develop a prediction model.

5.4.8.3. Experiment-3AS-Sampling [OS] and Experiment-3AS-Sampling [SMOTE]

When oversampling was applied, only one acceptable model was generated, that being the model when the DT algorithm was applied to VAR1. In this case, training accuracy was found to be 96.8% with validation accuracy being 87.6% (sampleSizePercent value was 196). The PRC,

precision, recall and F-measure values were all above 0.87, indicating reliable models. However, the ROC validation value of 0.48 may indicate an element of guessing when the model is applied to unseen data and was thus deemed as unacceptable. However, the validation dataset is very imbalanced (31.43) with only seven (7) instances labelled as “Fail”. For the VAR2 and VAR3 dataset variations, training accuracy achieved was at least 99.2% for both algorithms.

In the case of SMOTE sampling, for all variations, both the learning algorithms were not able to produce acceptable models that satisfy the acceptance criteria for this study. All training accuracies achieved were in the range 98.7% to 100%. This indicates overfitting of the model to the training data.

5.4.8.4. Analysis of experiments conducted

For this course, no acceptable prediction models could be generated by either algorithm for any of the variations. The reason for this would be the high number of pass instances when compared to the number of fail instances. In the six years of data captured for this course, 23 students had failed while 1054 students passed.

The nature of the course must also be taken into consideration. As described in Table 5.1, the majority of the course and a large portion of the assessment focuses on a project and the development of a front-end (Windows-based) application. This differs from other IS&T courses where the assessments are predominantly individual-based. For this course, the presentations revolving around the group project is the predominant form of assessment with no examination. Further to this, all students in their group obtain the same mark (unless disputes are made), thus resulting in a large number of students passing this course as long as the group that they are in are organized and meets the minimum requirements of their project submissions. In addition, the activities for this course are group based and not recorded by the Moodle LMS.

5.4.9. Experiments for the ISTN3SI dataset

This section focuses on the ISTN3SI course dataset. As discussed in Table 5.1, this course is a continuation of the ISTN3AS course discussed in section 5.4.8.

5.4.9.1. Experiment-3SI-Sampling [None]

The DT algorithm could not produce a viable model when applied to the VAR1 dataset. For the RF algorithm, the performance measures when applied to the VAR1 dataset are shown in Table 5.58. The accuracy achieved through training is similar to that of when the models are applied to the validation datasets and is above 95%, indicating reliable predictions.

Table 5.58: Summary analysis for RF generated model – Experiment-3SI-Sampling [None]-VAR1

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	95,4	96,4	0,63	0,58	0,92	0,94	0,94	0,95	0,95	0,96	0,93	0,95
Backward Search	12	94,9	96,4	0,62	0,49	0,92	0,95	0,9	0,95	0,95	0,96	0,92	0,95

For the VAR2 dataset, the performance of the models generated by the learning algorithms are shown in Table 5.59. The accuracies for both training and validation are within the acceptable range for this study. The accuracy difference is also less than 10%, which is an acceptable range for this study. As no sampling is used in this case and the dataset is imbalanced, the ROC value is not considered when determining the suitability of the prediction model. The PRC, precision, recall and F-measure values are also greater than 0.8 and within the acceptable range for this study.

Table 5.59: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [None]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	95,8	97,7	0,7	0,64	0,92	0,96	0,95	0,97	0,95	0,97	0,95	0,97
Backward Search	6	94,8	96,4	0,7	0,72	0,91	0,95	0,94	0,95	0,94	0,96	0,93	0,96
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	96,2	94,6	0,81	0,69	0,95	0,95	0,96	0,95	0,96	0,94	0,95	0,95
Backward Search	12	95,4	97,3	0,79	0,88	0,94	0,97	0,95	0,96	0,95	0,97	0,94	0,96

The DT algorithm was not able to produce an acceptable model for the VAR3 dataset (the model generated was a single leaf “P”). For the RF algorithm, models were generated for both

experiments with accuracy of 95.1% and 96.1% respectively. The PRC and F-measure values were also good and accuracy of 94.6% and 98.2% (outside acceptable range) were respectively observed when the models were applied to the validation datasets (Table 5.60).

Table 5.60: Summary analysis for RF generated models – *Experiment-3SI-Sampling [None]-VAR3*

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	96,1	96,4	0,74	0,76	0,94	0,96	0,96	0,95	0,96	0,96	0,95	0,95
Backward Search	6	95,1	98,2	0,62	0,77	0,92	0,97	0,95	0,98	0,95	0,98	0,93	0,97

5.4.9.2. *Experiment-3SI-Sampling [US]*

As with most experiments on undersampling, the removal of the majority of instances from the pass class resulted in unacceptable accuracy (less than or equal to 80% training accuracy). Only a single experiment using the RF algorithm applied to the VAR3 dataset provided a 90% training accuracy and 82% validation dataset accuracy.

5.4.9.3. *Experiment-3SI-Sampling [OS]*

When oversampling is applied to the three variations, the algorithms accuracy exceeds 97% for all experiments. The models also exhibit high accuracy when applied to the validation dataset (in the range 86% to 98%). Only one model was deemed acceptable based on the acceptance criteria for this study, that being the DT algorithm model when applied to VAR1.

As with the ISTN3AS course, the high pass rate of the course results in potential overfitting of the model to the training data, resulting in the almost 100% accuracies. Future data acquisition for this course will allow for generation of better models with regard to this course.

5.4.9.4. *Experiment-3SI-Sampling [SMOTE]*

The performances for the learning algorithms when applied to VAR1 with SMOTE is shown in Table 5.61. For these experiments, the Percentage parameter value was set to 1900. The accuracy achieved through training ranges from 95.2% to 96.4% and similar accuracy is obtained when the resultant models are applied to the validation datasets.

Table 5.61: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [SMOTE]-VAR1

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	13	96,3	94,6	0,97	0,59	0,95	0,94	0,96	0,94	0,96	0,94	0,96	0,94
Backward Search	13	96,4	95,5	0,97	0,58	0,95	0,94	0,96	0,94	0,96	0,95	0,96	0,95
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	95,2	92,8	0,98	0,52	0,98	0,94	0,95	0,94	0,95	0,92	0,95	0,93
Backward Search	20	96,2	96,4	0,98	0,57	0,98	0,95	0,96	0,95	0,96	0,96	0,96	0,95

With the VAR2 dataset (Percentage parameter value equaled 1250), 94.1% to 97.1% accuracy is achieved for all experiments when training. For both forward searches, only two attributes are identified for features selection and the resultant models do not fit the validation dataset. The backward searches for both algorithms produced prediction models that result in similar accuracy to the training data (see Table 5.62).

Table 5.62: Summary analysis for RF and DT generated models – Experiment-3SI-Sampling [SMOTE]-VAR2

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	94,7	74,6	0,93	0,56	0,9	0,94	0,94	0,94	0,94	0,74	0,94	0,82
Backward Search	14	95,5	95,1	0,96	0,55	0,93	0,94	0,95	0,94	0,95	0,95	0,95	0,94
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	94,1	72,8	0,96	0,53	0,95	0,94	0,94	0,94	0,94	0,72	0,94	0,81
Backward Search	27	97,1	96	0,99	0,7	0,99	0,95	0,97	0,95	0,97	0,96	0,97	0,95

Similar to VAR1 and VAR2, high training accuracy as well as validation accuracy was achieved for all experiments for the VAR3 dataset (a Percentage value of 1700 was used). As seen in Table 5.63, only the RF generated models have ROC values that are above 0.7. The PRC and F-Measure values were also high, indicating acceptable models have been generated.

Table 5.63: Summary analysis for RF and DT generated models – *Experiment-3SI-Sampling [SMOTE]-VAR3*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	95,7	96,8	0,96	0,61	0,95	0,95	0,95	0,95	0,95	0,96	0,95	0,96
Backward Search	17	95,4	95,1	0,95	0,45	0,93	0,94	0,95	0,93	0,95	0,95	0,95	0,94
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	96,8	92,4	0,98	0,74	0,99	0,96	0,96	0,93	0,96	0,92	0,96	0,93
Backward Search	29	97,7	95,5	0,99	0,75	0,99	0,96	0,97	0,93	0,97	0,95	0,97	0,94

5.4.9.5. Analysis of experiments conducted

Unlike the ISTN3AS course dataset, the application of this course dataset did result in prediction models that met the acceptance criteria for this study. When comparing the imbalance level of this course to that of the ISTN3AS course, it was noted that the ISTN3SI course was less imbalanced with more fail instances in the dataset. The assessment form for this ISTN3SI course is similar to that of the ISTN3AS course in that the predominant mode of assessment is the group project presentations. The increased number of failures can be attributed to web-based programming and development that students have not previously experienced in other IS&T courses. The web-based programming and development is prone to a greater number of potential errors and mistakes when compared to windows-based development done in ISTN3AS.

Thus, four models were identified from the experiments described in Sections 5.4.9.1 to 5.4.9.4. The performance measures for these models are listed in Table 5.64. Three of the models (*VAR2-None*, *VAR3-US* and *VAR2-SMOTE*) were generated using the RF algorithm and one model (*VAR1-OS*) was generated using the DT algorithm.

Table 5.64: Best four models for *Experiment-3SI*

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR2-None	95,4	97,3	0,79	0,88	0,94	0,97	0,95	0,96	0,95	0,97	0,94	0,96
VAR3-US	90	82,2	0,83	0,68	0,81	0,95	0,91	0,95	0,9	0,82	0,89	0,87
VAR1-OS	97,4	88,4	0,98	0,41	0,98	0,93	0,97	0,94	0,97	0,88	0,97	0,91
VAR2-SMOTE	97,1	96	0,99	0,7	0,99	0,95	0,97	0,95	0,97	0,96	0,97	0,95

The VAR2-None and the VAR2-SMOTE algorithms produced the best accuracies for both training and validation. The accuracy difference for these two models were also the smallest as shown in Figure 5.18.

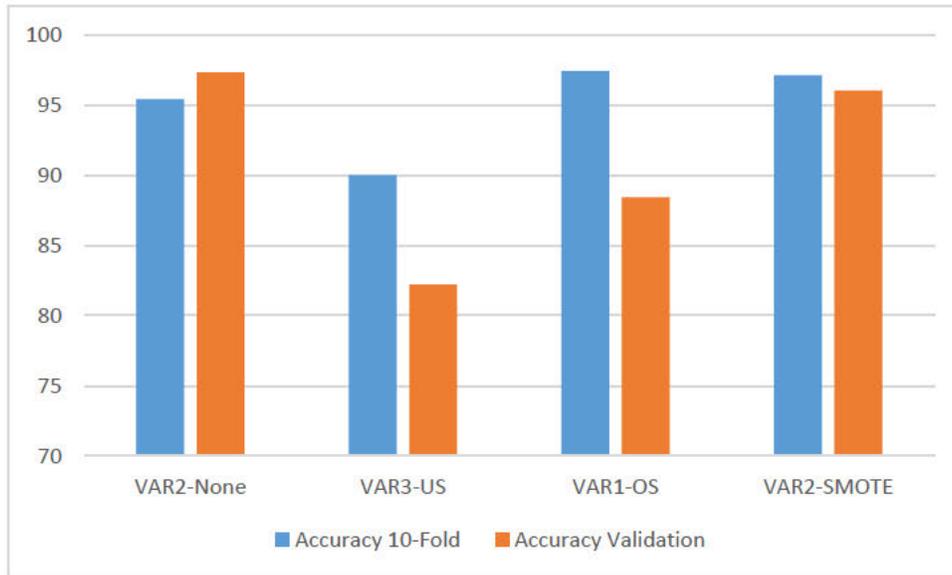


Figure 5.18: Accuracy comparison for four best models

As with the other models generated from previous courses, the PRC, precision, recall and F-measure values all indicate that the models are reliable in predicting unseen instances. The validation ROC values for the VAR1-OS (0.41) and VAR3-US (0.68) models were the only values less than the acceptable range for this study.

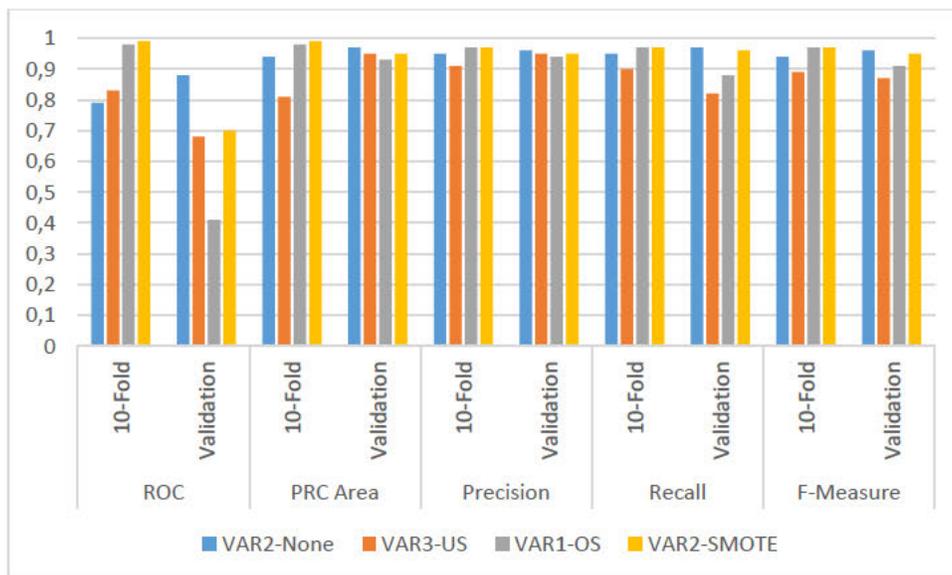


Figure 5.19: Assessment measure comparison for four best models

5.4.10. Experiments for the ISTN3ND dataset

This section covers the experiments related to the ISTN3ND dataset.

5.4.10.1. Experiment-3ND-Sampling [None]

The DT algorithm, when applied to VAR1 did not produce a viable model. The RF algorithm (forward search), however, did produce a viable model as shown in Table 5.65. The model generated by the backward search RF algorithm has an accuracy difference of 13.6%, which is outside the range for an acceptable model for this study.

Table 5.65: Summary analysis for RF generated model – Experiment-3ND-Sampling [None]-VAR1

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	82,3	88,8	0,64	0,46	0,77	0,95	0,79	0,95	0,82	0,88	0,77	0,92
Backward Search	22	80,3	93,9	0,68	0,71	0,79	0,97	0,78	0,96	0,8	0,94	0,79	0,95

For VAR2, neither of the algorithms were able to generate acceptable models. In the case of the forward search RF algorithm, the accuracy difference was greater than 10% while the DT algorithms and backward search RF algorithm resulted in validation accuracy of greater than 98%.

For VAR3, only 5 failing instances for the validation dataset resulted in near 100% accuracy achieved by both algorithms' models for training and validation datasets. Future dataset instances with a greater number of fail class instances would assist in better analysis of this dataset variation.

5.4.10.2. Experiment-3ND-Sampling [US]

For the VAR1 dataset with undersampling applied, the training accuracy for both algorithms was poor, ranging from 68% to 71% with poor accuracy for the validation dataset (40% to 64% range).

For the algorithms applied to the VAR2 dataset, the validation dataset did not fit the models generated during training of either of the algorithms. This is shown by the more than 10% differences between training accuracy and validation accuracy (Table 5.67).

Table 5.66: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [US]-VAR2

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	85,7	98,2	0,85	0,99	0,8	0,99	0,85	0,99	0,85	0,98	0,85	0,98
Backward Search	8	84,4	98,2	0,84	0,99	0,8	0,99	0,84	0,99	0,84	0,98	0,84	0,98
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	85,7	97,8	0,91	0,99	0,91	0,99	0,85	0,98	0,85	0,97	0,85	0,98
Backward Search	23	83,6	97,8	0,9	0,99	0,9	0,99	0,83	0,98	0,83	0,97	0,83	0,98

For VAR3, the difference in accuracies between training and validation is closer (see Table 5.67). However, when undersampling was applied, there were only 10 total instances to train on and thus future data acquisition will help better understand the effect of application of these learning algorithms in producing accurate predictions. Despite the small number of instances for training, the DT algorithms and the backward search RF algorithm were able to generate prediction models whose performance measures are acceptable for this study.

Table 5.67: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [US]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	90	89,2	0,8	0,84	0,86	0,97	0,91	0,97	0,9	0,89	0,89	0,92
Backward Search	1	90	96,1	0,9	0,98	0,86	0,98	0,91	0,98	0,91	0,96	0,9	0,97
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	100	95,2	1	0,99	1	0,99	1	0,98	1	0,95	1	0,96
Backward Search	8	90	94,8	1	0,98	1	0,99	0,91	0,98	0,9	0,94	0,89	0,96

5.4.10.3. Experiment-3ND-Sampling [OS]

For the VAR1 dataset, the RF algorithm produces high training accuracy (93.6%) but the model overfits the training data and similar accuracy cannot be obtained when the model is applied to the

validation data (sampleSizePercent was set to 163). The accuracies are closer when looking at the analysis of the DT algorithm (see Table 5.68)

Table 5.68: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [OS]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	88,9	80,2	0,94	0,44	0,92	0,95	0,89	0,95	0,89	0,8	0,88	0,87
Backward Search	8	88,9	80,2	0,94	0,44	0,92	0,95	0,89	0,95	0,89	0,8	0,88	0,87
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	93,6	64,8	0,96	0,58	0,96	0,96	0,94	0,96	0,93	0,64	0,93	0,76
Backward Search	21	93,6	66	0,97	0,58	0,96	0,96	0,94	0,96	0,93	0,66	0,93	0,77

The performance of both algorithms was much better when applied to VAR2 than when applied to VAR1 (for VAR2, a sampleSizePercent of 160 was used). Both accuracies obtained when training were in the range of 95.9% to 97.2%, with similar accuracy when the models are applied to the validation dataset (see Table 5.69). The good performance is confirmed by the ROC, PRC, precision, recall and F-Measure values.

Table 5.69: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [OS]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	12	96	93,9	0,96	0,88	0,95	0,98	0,96	0,97	0,96	0,94	0,96	0,95
Backward Search	17	95,9	94,4	0,97	0,78	0,96	0,97	0,96	0,97	0,95	0,94	0,95	0,95
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	97,2	98,7	0,98	0,97	0,98	0,99	0,97	0,98	0,97	0,98	0,97	0,98
Backward Search	24	96,4	97,8	0,99	0,98	0,99	0,99	0,96	0,96	0,96	0,97	0,96	0,98

The performances of the algorithms when applied to the VAR3 dataset resulted in models with at least 99.7% accuracy for the experiments. When the models were applied to the validation dataset,

accuracy of at least 97.4% was achieved. As with previous experiments, more instances are required to better understand the performance of the algorithms when applied to this dataset.

5.4.10.4. Experiment-3ND-Sampling [SMOTE]

The learning algorithms, when applied to VAR1 using SMOTE (with a Percentage value of 350), produced models with acceptable accuracy in the range 84% to 86.7%. The models, when applied to the validation dataset, achieved similar accuracy (with the exception of the forward search RF algorithm). The performance of algorithms when oversampling (described in the previous section) is used perform better than when the algorithms are applied using SMOTE for VAR1.

Table 5.70: Summary analysis for RF and DT generated models – Experiment-3ND-Sampling [SMOTE]-VAR1

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	5	86,7	87,5	0,85	0,46	0,81	0,95	0,86	0,96	0,86	0,87	0,86	0,91
Backward Search	15	84,2	83,6	0,86	0,79	0,82	0,97	0,84	0,97	0,84	0,83	0,84	0,89
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	84,5	68,2	0,92	0,56	0,91	0,96	0,84	0,96	0,84	0,68	0,84	0,79
Backward Search	20	85,7	82,8	0,93	0,76	0,93	0,97	0,85	0,97	0,85	0,82	0,85	0,88

For the VAR2 experiments (Percentage parameter value was set to 300), better performance was achieved for both algorithms with accuracy in the range of 85% to 90%. When applied to the validation dataset, near 100% accuracy was achieved for three of the four experiment variations (see Table 5.71). The difference in accuracies between training and validation were much closer when using the RF algorithm than when using the DT algorithm.

Table 5.71: Summary analysis for RF and DT generated models – *Experiment-3ND-Sampling [SMOTE]-VAR2*

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	14	88,2	98,7	0,9	0,89	0,87	0,98	0,88	0,98	0,88	0,98	0,88	0,98
Backward Search	11	88,2	99,1	0,89	0,89	0,86	0,99	0,88	0,99	0,88	0,99	0,88	0,99
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	85,2	94,8	0,89	0,42	0,88	0,95	0,85	0,96	0,85	0,94	0,85	0,95
Backward Search	30	90,7	98,7	0,96	1	0,96	1	0,9	0,98	0,9	0,98	0,9	0,98

The performances of the models when the algorithms are applied to VAR3 produce near 100% accuracy for both training and validation datasets. As with previous experiments using VAR3, more instances are required to better assess how these algorithms perform with this dataset.

5.4.10.5. Analysis of experiments conducted

Four models were identified from each of the experiments described from section 5.4.10.1 to 5.4.10.4. The assessment measures for each of these models are listed in Table 5.72.

Table 5.72: Performance measures for best four models for *Experiment-3ND*

Variation and Sampling	Accuracy		ROC		PRC Area		Precision		Recall		F-Measure	
	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
VAR1-None	82,3	88,8	0,64	0,46	0,77	0,95	0,79	0,95	0,82	0,88	0,77	0,92
VAR3-US	90	94,8	1	0,98	1	0,99	0,91	0,98	0,9	0,94	0,89	0,96
VAR2-OS	96,4	97,8	0,99	0,98	0,99	0,99	0,96	0,96	0,96	0,97	0,96	0,98
VAR1-SMOTE	84,2	83,6	0,86	0,79	0,82	0,97	0,84	0,97	0,84	0,83	0,84	0,89

Three of the models selected were generated using the RF algorithm, that being the *VAR1-None*, *VAR3-US* and *VAR2-OS* models while the *VAR1-SMOTE* model was generated using the DT algorithm. All models' accuracy falls within the acceptable range for this study and the accuracy differences fall to within 10%. Figure 5.20 (Accuracy comparison) shows that the *VAR2-OS* has the best training and validation accuracy while the *VAR1-SMOTE* had the closest accuracy difference.

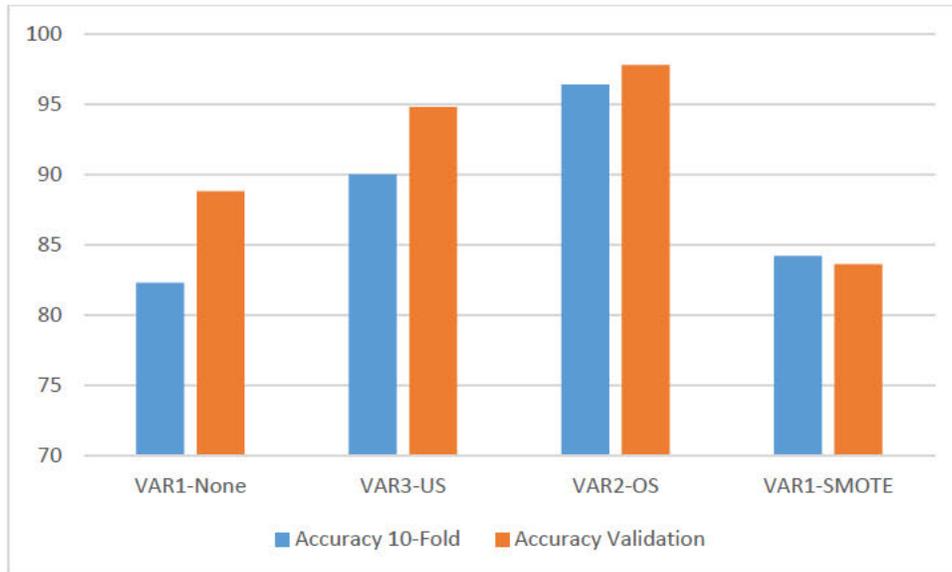


Figure 5.20: Accuracy comparison of four models

A comparison of the other performance measures for each of the models is illustrated in Figure 5.21. The best model when no sampling was used (*VAR1-None*) had a ROC value outside the range for the acceptance criteria for this study (less than 0.7). This was expected as the dataset is imbalanced. The remaining performance measures were greater than 0.8 and thus acceptable for this study.

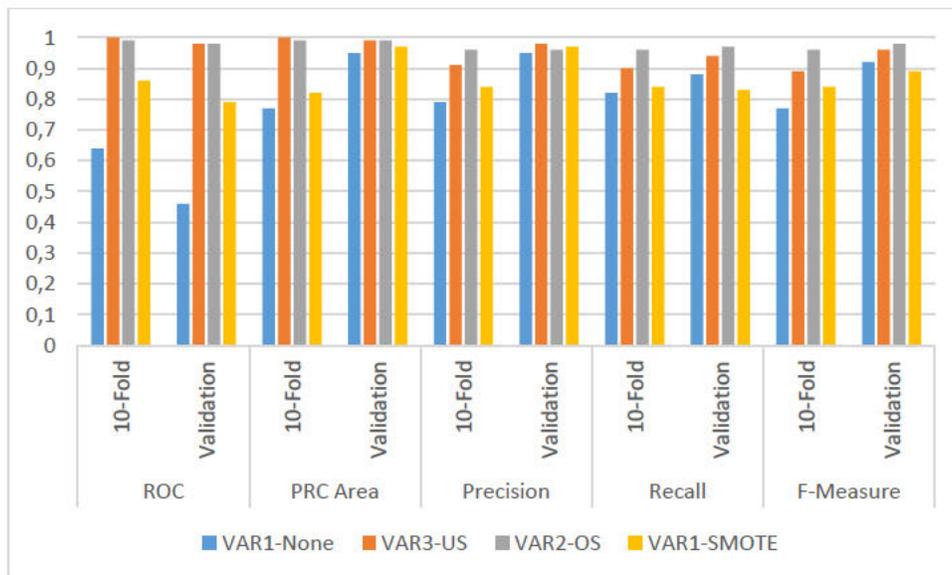


Figure 5.21: Assessment measure comparison for four models

5.5. Chapter summary

This chapter described experiments covering the application of two learning algorithms to the course datasets in the UKZN ISTN dataset. A summary of the best algorithms (and corresponding accuracies) are listed in Table 5.73. These results are based on a combination of closeness between validation and training accuracy as well as overall accuracy for the training and validation datasets (with the exception of ISTN100).

Table 5.73: Best performing algorithms for each course based on accuracy

Dataset	VAR	Sampling	Algorithm	Training accuracy %	Validation accuracy %	Accuracy difference
ISTN100	1	SMOTE	DT	89.3	84.6	4.7
ISTN101	1	OS	RF	94.5	92.6	1.9
ISTN103	2	OS	RF	96.5	95.8	0.7
ISTN2IP	2	SMOTE	RF	90.3	87.4	2.9
ISTN211	3	None	RF	97.4	95.4	2
ISTN212	3	None	RF	97.1	95.9	1.2
ISTN3SA	2	SMOTE	RF	94.2	96.9	2.7
ISTN3AS	No viable model found					
ISTN3SI	2	SMOTE	RF	97.1	96	1.1
ISTN3ND	2	OS	RF	96.4	97.8	1.4

Of the two learning algorithms used, the RF algorithm was noted to have performed better than the DT algorithm. From the models identified in Table 5.73, 8 of the 10 models were generated using the RF algorithm. This confirms the ability of the RF algorithm to better handle imbalanced data (Bekkar & Alitouche, 2013) in the case of the 2nd year courses (ISTN2IP, ISTN211 and ISTN212) where all three models listed were generated without sampling. Furthermore, the bagging procedure and random feature selection of the RF algorithm assisted in the development of viable, more generalizable models (Kovanović et al., 2018), unlike the DT algorithm which was prone to overfitting and development of unusable single node decision trees.

The ISTN100 course, while having acceptable accuracy and accuracy difference, did not meet the acceptance criteria for the ROC value when the model was applied to the validation dataset (0.6). The values for the other performance measures were acceptable for this study. Thus, a model with an improved ROC value would be preferred.

In the case of the ISTN2IP dataset, the accuracies for both training and validation are within the acceptable range for this study. However, it would be preferable if the validation accuracy was greater than 90% and thus an alternate approach was used (explained in Chapter 6) to find a better model.

No viable model was found for the ISTN3AS course. The difficulty in obtaining a model could be attributed to a very high imbalance characteristic. Furthermore, the course follows a different assessment format (the major project) that formed a large part of the student final mark. Students within each project group attained the same mark in most cases, possibly allowing for an increased number of student passes. As explained in section 5.4.9.5, while the ISTN3SI course follows a similar assessment format to that of ISTN3AS, the increase in the number of failures was due to increased difficulty in the course.

In terms of answering the research questions RQ3 and RQ4, standard learning algorithms were applied to the datasets. These algorithms were able to establish patterns to predict student academic performance. The performances of these algorithms were verified by applying the patterns to unseen data. The prediction accuracies as well as other performance measures achieved were within the ranges of the acceptance criteria specified in section 5.3.5. The standard learning algorithms performed well when applied to all but one dataset (ISTN3AS).

In the event that standard learning algorithms do not provide sufficient or acceptable solutions, the inclusion of artificial intelligence techniques are known to assist in this regard. Thus, in the next chapter, the area of Artificial Intelligence was investigated in an attempt to find better prediction models for the cases of ISTN100, ISTN2IP and ISTN3AS.

Chapter 6 – Prediction using genetic algorithms

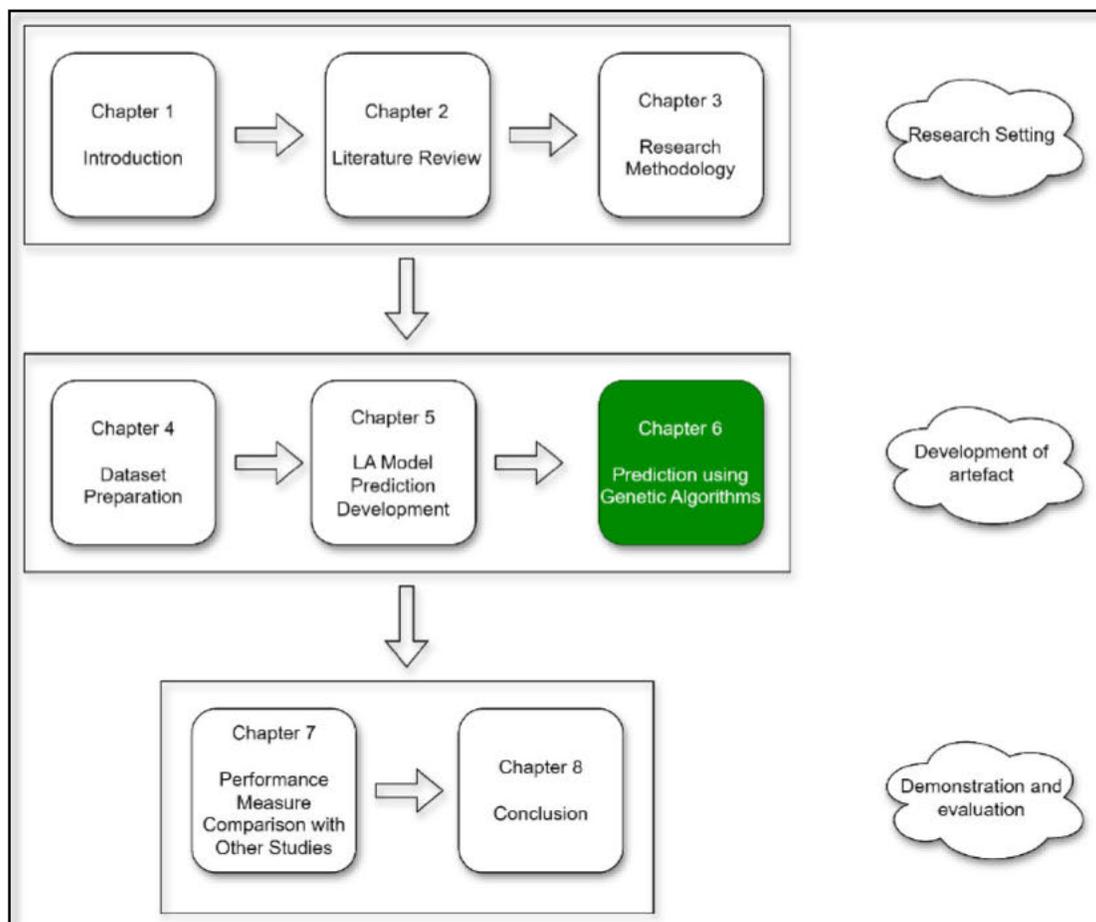


Figure 6.1: Thesis structure

6.1. Introduction

In section 5.2.3, it was stated that data imbalance can be addressed from a data perspective, usually through the use of sampling methods such as undersampling, oversampling and SMOTE. An alternate way of dealing with data imbalance is through an algorithmic approach. The area of Artificial Intelligence (AI) is important in any big data domain, including LA. The use of AI allows for difficult pattern recognition, learning and other tasks in the analytics process (O'Leary, 2013). Thus, the use of AI-based algorithms is beneficial to address data imbalance problems. Genetic Algorithms (GAs), also sometimes referred to as evolutionary algorithms, are one such

AI technique, and according to Minaei-Bidgoli and Punch (2003), GAs have the potential to improve accuracy by 10% to 12% when compared to non-GA classifiers.

For this study, two (2) GA-based approaches were identified, these being the use of the GA as part of the classifier as well as the use of the GA as an optimization tool with regard to feature selection for other classifiers. For the former approach, Romero, González, Ventura, Del Jesús and Herrera (2009) used an evolutionary algorithm to build rules for the discovery of relationships between student Moodle usage and academic performance. The latter, where GAs are used as a part of feature selection, was an approach followed by Lakshmi, Martin and Venkatesan (2013) and Preetha (2021), amongst others, to identify the most suitable set of features to use for different learning algorithms.

In Chapter 5, the experiments conducted yielded acceptable prediction models for all but three of the courses in the UKZN ISTN dataset. Where acceptable prediction models were found, the performance measures were all within the acceptance criteria and the accuracies were above 90%. As a pragmatic research paradigm is being followed, there is no need to use AI techniques to find prediction models for courses where acceptable models with at least 90% accuracy have already been found.

Thus, in this chapter, both GA-based approaches described above are applied to try to find better models for three courses in the UKZN ISTN dataset. Experiments are conducted in order to find better prediction models for the ISTN100 and ISTN2IP datasets, as well as finding an appropriate model for the ISTN3AS dataset. For the case of the ISTN100 and ISTN2IP datasets, the objective is to try to obtain an accuracy of above 90% (and less than 98% as per the acceptance criteria) while also having acceptable values for other performance measures. For the ISTN3AS dataset, no model could be found that meets the acceptance criteria in Chapter 5 so the GA is incorporated to attempt to find an acceptable model.

From the perspective of the DSRM described in Chapter 3, this chapter continues the focus on the design and development of the artefact. Chapter 5 focused on applying learning algorithms to the now prepared dataset with the objective of predicting student performance. Chapter 6 continues

the development of the artefact (see Figure 6.1) with the focus on predicting student performance and the use of an AI technique has been included with the objective to improve performance. Section 6.2 describes the courses that were chosen and the justification for using these courses. Section 6.3 describes the GA technique, the parameters used and how the GA is applied in this study. Section 6.4 describes the results of the experiments performed and how they compare to the best performance measures obtained for each course dataset. Section 6.4 provides a conclusion on the chapter. An outline of the chapter is provided in Figure 6.2.

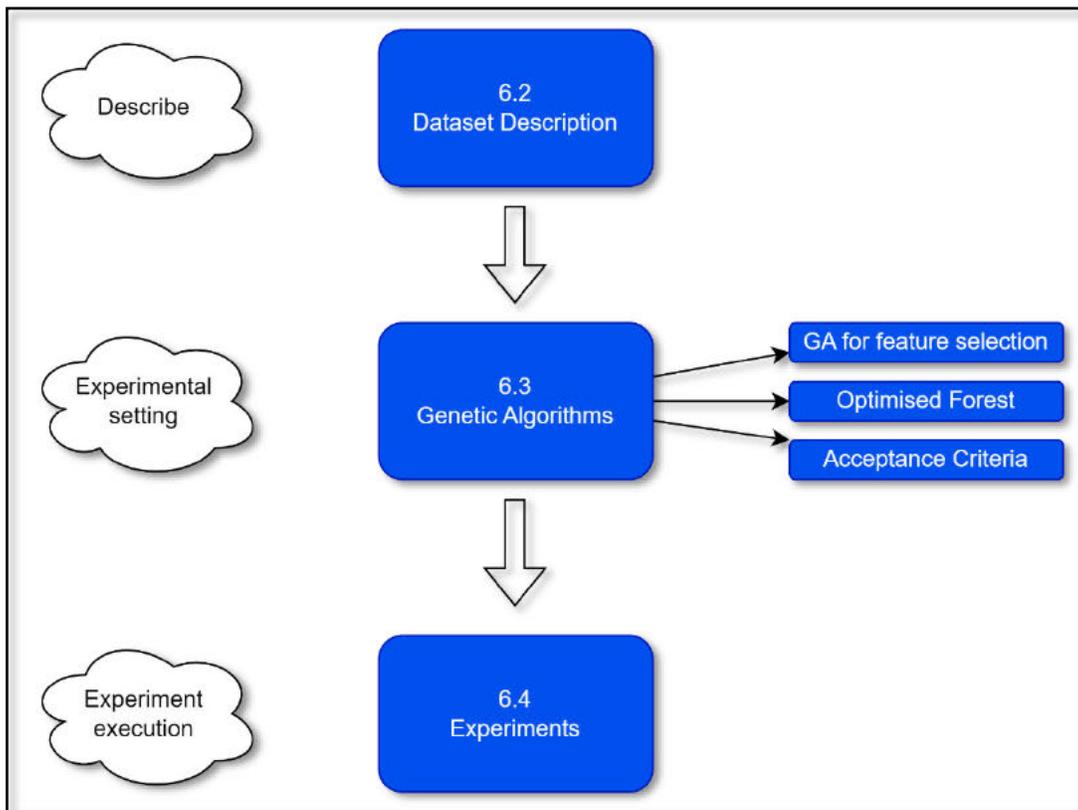


Figure 6.2: Map for Chapter 6 coverage

6.2. Course datasets to be applied to genetic algorithms

Chapter 5 describes experiments covering the application of the Random Forest (RF) and Decision Tree (DT) algorithms to the selected IS&T courses for this study. In the concluding Section 5.5, three courses were identified that either still required a prediction model (ISTN3AS) or courses where the prediction model could be further improved (ISTN100 and ISTN2IP).

The ISTN100 course was the only course with only one variation due to the Moodle data not being available for collection. Section 5.4.1 outlined the experiments conducted using the DT and RF algorithms for the ISTN100 course. The best model from these experiments was the DT algorithm. The performance measures for this algorithm were shown in Table 5.9, and for convenience, are shown in Table 6.1. As can be seen in Table 6.1, the model met all the acceptance criteria with the exception of the Receiver Operator Characteristic (ROC) value. The accuracy for training and validation were 89.3% and 84.6% respectively. As seven (7) of the ten (10) courses in this study achieved accuracy (training and validation) of above 90%, the objective was to then find an improved model through the use of artificial intelligence techniques.

Table 6.1: Prediction performance for DT algorithm extracted from Table 5.9 for ISTN100 course

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	11	89.3	84.6	0.9	0.6	0.86	0.78	0.89	0.82	0.89	0.84	0.89	0.82

Similarly, the assessment measures for the ISTN2IP course also met the acceptance criteria for this study (Table 6.2 shows the RF generated model performance extracted from Table 5.33). The best model using the RF algorithm obtained a training accuracy of 92.9% and a validation accuracy of 88.1%. As this model was generated on a dataset with no sampling, the ROC value of 0.66 was not considered as violating the acceptance criteria. The objective was to find a model where the validation accuracy is also above 90% (and less than 98%).

Table 6.2: Prediction performance for RF algorithm extracted from Table 5.33 for ISTN2IP course

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85

In the case of the ISTN3AS course, no suitable model was found using the DT or RF algorithms. Thus, the objective was to use artificial intelligence techniques to attempt to find a prediction model for this course dataset that meets the acceptance criteria for this study.

6.3. An overview of genetic algorithms

Genetic algorithms are a rapidly developing area within AI and is based on theories of biological evolution such as natural selection and genetic inheritance (Obitko, 1998). Genetic algorithms are commonly used to solve combinatorial optimization problems. For this study, two GA alternatives were used. When being used for feature selection, the GA used was similar to the initial GA proposed by Golberg (1989) and is described in Section 6.3.1. Similarly, when using the GA as part of classification, the GA proposed by Golberg (1989) was also used with the exception that an elitist approach was followed (discussed in Section 6.3.2).

6.3.1. Genetic algorithm used for feature selection

A GA begins with the creation of a population of individuals, referred to as the initial population. Each individual is a string containing a randomly chosen set of selected features from the integrated dataset. Each individual in the population is then evaluated based on a fitness function. In this case, the fitness function is the accuracy obtained when the selected algorithm has been applied to the dataset using only that individual's set of features. A selection process is then undertaken where individuals of this initial population are selected to become parents. Copies of the parents are made, followed by the application of genetic operators (in this case, crossover and mutation) to the copies. These copies are then referred to as the offspring and are added as individuals to a new population. This process of evaluation, selection and creation of a new generation of offspring continues until some termination criteria has been met. Termination criteria depends on the type of problem being addressed and may include reaching a generational limit or an ideal solution has been obtained (Golberg, 1989). In the case of this study, the GA terminates when the maximum specified number of generations is reached (see Section 6.3.3) and the set of features to be used is that of the individual with the best fitness function (accuracy). The GA used in this study is described in Algorithm 6.1.

Input: List of features	
Output: Individual with set of features that produced the best accuracy	
1	Function GeneticAlgorithm
2	{
3	Gen = 0
4	Create initial population consisting of different combinations of features from dataset
5	Evaluate initial population by applying chosen machine learning technique
6	Repeat
7	{
8	J = 1
9	Repeat
10	{
11	Select two (2) individuals, A and B , from population
12	If crossover occurs with probability P_c
13	Offspring C and D = Application of crossover operator to A and B
14	Else
15	Offspring C and D = Copy of A and B
16	If mutation occurs with probability P_m
17	Apply mutation to offspring C and D
18	End If
19	Evaluate C and D
20	Add C and D to new population
21	J = J + 2
22	}
23	Until J > Population Size
24	Old population replaced by new population
25	}
26	Until Termination criteria have been met
27	}

Algorithm 6.1: Genetic Algorithm to determine set of features with best accuracy

As stated in Algorithm 6.1, two (2) individuals from the population are randomly selected to be parents. Offspring are created by either duplicating the parents or by the application of genetic operators, which can occur based on a given probability. The processes of crossover and mutation, which will be executed based on probabilities P_c and P_m respectively, are explained in the examples shown in Figure 6.3 and Figure 6.4.

Parent A	<u>5, 7, 8, 12</u> , 17, 23, 25, 26, 28
Parent B	1, 4, 7, 12, <u>13, 16, 22, 24, 27, 30</u>
Resultant Offspring C	<u>5, 7, 8, 12, 13, 16, 22, 24, 27, 30</u>
Resultant Offspring D	1, 4, 7, 12, 17, 23, 25, 26, 28

Figure 6.3: Crossover example and resultant offspring

In Figure 6.3, assume that each number represents a selected feature. For example, 5 represents the ISTN212 mark, 8 is the student QUAL, 12 is the Age Category and so on. In the example shown in Figure 6.1, the crossover operator is applied where parts of each of the parents are taken and combined to form a resultant offspring. Here, the features 5, 7, 8 and 12 from Parent A are combined with the group of features 13, 16, 22, 24, 27 and 30 from Parent B, resulting in Offspring C. The other set of features from the parents are combined resulting in Offspring D.

For mutation, selected points (feature) in the individual are changed. In the case of Figure 6.4, feature 15 is replaced by feature 19.

Parent C	2, 4, 7, 8, 12, <u>15</u> , 18, 20
Resultant Offspring	2, 4, 7, 8, 12, 18, <u>19</u> , 20

Figure 6.4: Mutation Example

The newly created offspring are then added to the population of a new generation and evaluated. In the case of this study, the resultant offspring is a combination of features from both the parents (crossover) as well as after mutation has been applied. Once individuals of a new generation have all been created and evaluated, new generations of individuals are created continuously until the maximum number of generations have been met. The individual in the final generation with the set of attributes that produces the best predictive model (based on accuracy) is chosen and that model is applied to the validation dataset.

6.3.2. Optimized forest (OF) algorithm

The OF algorithm is an adapted Random Forest algorithm developed by Adnan and Islam (2016). In this case, rather than using an exhaustive search for the optimal decision tree, a genetic algorithm

is used to select the best performing decision trees with the objective of improving the final outcome or accuracy produced by a Random Forest algorithm. In terms of this algorithm, the GA described is also similar to that described in Algorithm 6.1. The only exception is that the genetic operators are applied with the concept of elitism, where the offspring is only accepted if it is evaluated as being better than the parent. If it is not, then the offspring is rejected and the parent becomes the offspring (Adnan & Islam, 2016).

6.3.3. Performance measures and parameters

The same performance measures from Section 5.3.5 were used, that being the accuracy, ROC, PRC (Precision Recall Curve), precision, recall and F-Measure values. The main acceptance criteria for the models were also the same as that described in Table 5.5., i.e., accuracy in the range between 80% and 98% and acceptable values for the PRC, ROC and F-measure values where required.

In WEKA, and as with Section 5.3.4, `WrapperSubsetEval` was used as the attribute evaluator with the classifier property being either the RF algorithm or the J48 DT algorithm. The number of folds was set to 10. Unlike the experiments conducted in the previous chapter, the search method that was used was the genetic search (described in Algorithm 6.1) rather than forward or backward searches.

With regard to the OF algorithm, the algorithm was tested using all attributes (no feature selection) as well as attributes obtained using genetic search RF, forward search RF and backward search RF respectively. The default parameter values were used.

In terms of GA parameters, the probability of the crossover operator occurring was set to 60%, mutation probability was 3.3%, the maximum number of generations was 20, and the population size was 20. These were the default parameters set by WEKA.

6.4. Results of genetic based experiments conducted

This section describes the results achieved when the learning algorithms were applied to the ISTN100, ISTN2IP and ISTN3AS course datasets. Here, GAs were used for feature selection.

The section also covers the performance when the OF algorithm was applied to the datasets. The experiments include whether the variations have had any sampling techniques applied or not.

6.4.1. Experiments for the ISTN100 dataset

This section covers the genetic algorithm-based experiments for the ISTN100 course dataset.

6.4.1.1. Experiment-100-FS [Genetic]

Table 6.3 presents the performance measures obtained for the ISTN100 prediction models generated when using genetic search for feature selection. The row labelled 1 (line 1), highlighted in green, indicates the performance measures for the best model for this course listed in Table 6.1.

Table 6.3: Summary analysis for RF and DT algorithms – Experiment-100-FS [Genetic]

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	SMOTE	Forward Search	DT	11	89,3	84,6	0,9	0,6	0,86	0,78	0,89	0,82	0,89	0,84	0,89	0,82
2	None	Genetic	RF	8	86,4	83	0,61	0,64	0,8	0,79	0,83	0,7	0,86	0,83	0,81	0,76
3		Genetic	DT	9	86,2	83,8	0,5	0,5	0,76	0,72	0,85	?	0,86	0,83	0,8	?
4	US	Genetic	RF	13	64,3	55,6	0,67	0,67	0,66	0,82	0,64	0,79	0,64	0,55	0,64	0,61
5		Genetic	DT	13	65,4	65,3	0,66	0,78	0,63	0,83	0,65	0,84	0,65	0,65	0,65	0,69
6	OS	Genetic	RF	25	95,5	74,1	0,99	0,69	0,99	0,82	0,95	0,78	0,95	0,74	0,95	0,75
7		Genetic	DT	16	91,6	80,6	0,96	0,61	0,94	0,78	0,92	0,79	0,91	0,8	0,91	0,8
8	SMOTE	Genetic	RF	26	89,6	79	0,95	0,65	0,94	0,82	0,89	0,77	0,89	0,79	0,89	0,78
9		Genetic	DT	15	89,2	82,2	0,89	0,59	0,86	0,77	0,89	0,78	0,89	0,82	0,89	0,79

In terms of the *None-Genetic-DT* (line 3) and *None-Genetic-RF* (line 2) algorithms, the models generated do produce acceptable accuracy but it was not as good as the best model achieved in Chapter 5. As with *Experiment-100-Sampling [US]* covered in Section 5.4.1.2, the use of undersampling results in models with unacceptable accuracy. For oversampling (OS), the training accuracy of the generated models was greater than 90% but the difference between the resultant validation accuracy is greater than 10% and thus the models are not acceptable. When SMOTE is applied, the training accuracy is similar to the best model from Chapter 5 (line 1) but the validation accuracy is lower.

6.4.1.2. Experiment-100-Algorithm [OF]

The OF algorithm is applied to the ISTN100 dataset using different sampling techniques and different attributes determined using different feature selection techniques. The performance measures obtained for each experiment are listed in Table 6.4.

Table 6.4: Summary analysis for OF algorithm – Experiment-100-Algorithm [OF]

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	SMOTE	Forward Search	DT	11	89,3	84,6	0,9	0,6	0,86	0,78	0,89	0,82	0,89	0,84	0,89	0,82
2	None	All attributes	OF	30	85,8	83,8	0,66	0,71	0,82	0,83	0,82	0,79	0,85	0,83	0,85	0,79
3		RF Forward Search	OF	6	86,3	83,8	0,51	0,5	0,76	0,72	0,84	?	0,86	0,83	0,8	?
4		RF Backward Search	OF	28	88,5	83,8	0,87	0,74	0,91	0,83	0,87	0,83	0,88	0,77	0,86	0,77
5		Genetic	OF	9	86,4	83,8	0,52	0,52	0,77	0,73	0,85	?	0,86	0,83	0,81	?
6	US	All attributes	OF	30	62,5	45,9	0,68	0,67	0,67	0,82	0,62	0,81	0,62	0,46	0,62	0,51
7		RF Forward Search	OF	16	62,8	50	0,67	0,71	0,67	0,83	0,62	0,8	0,62	0,5	0,62	0,55
8		RF Backward Search	OF	19	66,1	54,8	0,69	0,61	0,66	0,79	0,66	0,76	0,66	0,54	0,66	0,6
9		Genetic	OF	14	64,3	47,5	0,7	0,66	0,68	0,81	0,64	0,78	0,64	0,47	0,64	0,53
10	OS	All attributes	OF	30	89,6	68,5	0,98	0,68	0,98	0,81	0,9	0,8	0,89	0,68	0,89	0,72
11		RF Forward Search	OF	16	89,8	67,7	0,98	0,69	0,97	0,82	0,9	0,79	0,89	0,67	0,89	0,71
12		RF Backward Search	OF	28	91,7	61,2	0,98	0,62	0,97	0,79	0,92	0,76	0,91	0,61	0,91	0,62
13		Genetic	OF	26	89,7	64,5	0,98	0,69	0,98	0,82	0,9	0,78	0,89	0,64	0,89	0,69
14	SMOTE	All attributes	OF	30	88,5	75,8	0,95	0,65	0,94	0,81	0,88	0,74	0,88	0,75	0,88	0,75
15		RF Forward Search	OF	8	87,1	77,4	0,94	0,59	0,94	0,8	0,87	0,76	0,87	0,77	0,87	0,76
16		RF Backward Search	OF	29	88,9	75,8	0,95	0,62	0,94	0,81	0,89	0,74	0,89	0,75	0,89	0,75
17		Genetic	OF	26	90,2	77,4	0,94	0,59	0,92	0,76	0,9	0,77	0,9	0,77	0,9	0,77

The OF algorithm, when applied to the dataset, did not find any better models than that generated using the forward search DT algorithm (line 1 in Table 6.4). When no sampling is used, the models generated have assessment measure values similar to, but slightly lower than that of the best model generated by the forward search DT algorithm. The models, when undersampling was used, were not acceptable. In the case of oversampling and SMOTE, the models generated were noted to overfit onto the training data and similar accuracy for the validation dataset was not achieved.

Thus, the use of genetic algorithms, either for feature selection or as part of the OF algorithm, could not find a better model than the one generated by the forward search DT algorithm. The use of LMS interaction data was shown to improve accuracy for the other datasets, thus future research should look into the acquisition of LMS interaction data for the ISTN100 course.

6.4.2. Experiments for the ISTN2IP dataset

This section covers the genetic algorithm-based experiments for the ISTN2IP course dataset.

6.4.2.1. Experiment-2IP-FS [Genetic]

For VAR1, using a GA for feature selection did not yield any better models as shown in Table 6.5. The undersampling results have been removed as these results did not meet the acceptance criteria. The *OS-Genetic-RF* algorithm (line 5) produced the highest training accuracy (96%) but the model did not yield an equivalent accuracy when applied to the validation dataset (81.8%) where a difference of 14.2% was observed. The *SMOTE-Genetic-DT* (line 6) model produced similar

accuracy for both training and validation datasets, but the accuracy is lower when compared to the forward search RF algorithm obtained in Chapter 5 (line 1).

Table 6.5: Summary analysis for RF and DT algorithms – Experiment-2IP-FS [Genetic] – VARI

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Forward Search	RF	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
2	None	Genetic	DT	2	79,2	90,9	0,49	0,5	0,66	0,83	?	?	0,79	0,9	?	?
3		Genetic	RF	7	80,2	90,9	0,66	0,72	0,76	0,88	0,78	?	0,8	0,9	0,73	?
4	OS	Genetic	DT	15	87,8	71,3	0,94	0,56	0,93	0,85	0,88	0,85	0,87	0,71	0,87	0,76
5		Genetic	RF	19	96	81,8	0,99	0,66	0,98	0,89	0,96	0,87	0,96	0,81	0,96	0,84
6	SMOTE	Genetic	DT	16	84,9	88,1	0,86	0,74	0,82	0,89	0,84	0,88	0,84	0,88	0,84	0,88
7		Genetic	RF	15	85,7	79	0,91	0,59	0,9	0,85	0,85	0,84	0,85	0,79	0,85	0,81

With the VAR2 dataset, the *SMOTE-Genetic-RF* model (line 7 on Table 6.6) had the closest accuracy difference of 2.1% between training and validation datasets. However, the accuracy produced for both training and validation is less than that of the accuracies obtained for the forward search RF algorithm obtained in Chapter 5 (see line 1 on Table 6.6).

Table 6.6: Summary analysis for RF and DT algorithms – Experiment-2IP-FS [Genetic] – VAR2

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Forward Search	RF	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
2	None	Genetic	DT	11	85,7	90,9	0,66	0,81	0,78	0,9	0,84	?	0,85	0,9	0,83	?
3		Genetic	RF	18	85,7	90,9	0,82	0,88	0,88	0,94	0,84	0,87	0,85	0,9	0,83	0,87
4	OS	Genetic	DT	19	94,1	74,8	0,95	0,78	0,94	0,89	0,94	0,91	0,94	0,74	0,94	0,8
5		Genetic	RF	16	97,4	70,6	0,98	0,68	0,98	0,88	0,97	0,85	0,97	0,7	0,97	0,76
6	SMOTE	Genetic	DT	8	89,3	79,7	0,89	0,61	0,86	0,86	0,89	0,84	0,89	0,79	0,89	0,82
7		Genetic	RF	21	89,5	87,4	0,96	0,85	0,96	0,92	0,89	0,88	0,89	0,87	0,89	0,87

The performance measures when the genetic search was used for VAR3 is shown in Table 6.7. The use of the genetic search as part of feature selection did not result in models that were better than the model produced by the forward search RF algorithm from Chapter 5 (line 1). The accuracy differences for the experiments with no sampling and oversampling were not acceptable (greater than 10%). When SMOTE sampling was used, the accuracy difference is acceptable but the accuracies are not as good as the forward search RF algorithm model's accuracy.

Table 6.7: Summary analysis for RF and DT algorithms – Experiment-2IP-FS [Genetic] – VAR3

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Forward Search	RF	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
2	None	Genetic	DT	2	79,2	90,9	0,49	0,5	0,66	0,83	?	?	0,79	0,9	?	?
3		Genetic	RF	7	80,2	90,9	0,66	0,72	0,76	0,88	0,78	?	0,8	0,9	0,73	?
4	OS	Genetic	DT	15	87,8	71,3	0,94	0,56	0,93	0,85	0,88	0,85	0,87	0,71	0,87	0,76
5		Genetic	RF	19	96	81,8	0,99	0,66	0,98	0,89	0,96	0,87	0,96	0,81	0,96	0,84
6	SMOTE	Genetic	DT	16	84,9	88,1	0,86	0,74	0,82	0,89	0,84	0,88	0,84	0,88	0,84	0,88
7		Genetic	RF	15	85,7	79	0,91	0,59	0,9	0,85	0,85	0,84	0,85	0,79	0,85	0,81

6.4.2.2. Experiment-2IP-Algorithm [OF]

When the OF algorithm was applied to the VAR1 dataset, acceptable models were identified. However, the difference between training accuracy and validation accuracy were either greater than 10% or the accuracy achieved was not as good as the forward search RF algorithm’s performance found in Chapter 5 (see line 1 on Table 6.8). This was especially noted for the oversampled dataset where the accuracies for training were all in the range of 93.6% to 96.7%. The resultant models could not produce similar accuracy for the validation dataset.

Table 6.8: Summary analysis for OF algorithm – Experiment-2IP-Algorithm [OF] – VAR1

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Forward Search	RF	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
2	None	All	OF	29	78,1	88,1	0,6	0,66	0,73	0,87	0,72	0,85	0,78	0,88	0,72	0,86
3		RF Fwd Search	OF	4	80	91,6	0,52	0,47	0,68	0,83	0,78	0,92	0,8	0,91	0,72	0,88
4		RF Bkwd Search	OF	24	79,6	86,7	0,58	0,58	0,72	0,85	0,75	0,85	0,79	0,86	0,74	0,85
5		Genetic	OF	7	80,2	90,9	0,53	0,46	0,68	0,82	0,78	?	0,8	0,9	0,73	?
6	OS	All	OF	29	93,6	78,3	0,99	0,76	0,99	0,9	0,94	0,88	0,93	0,78	0,93	0,82
7		RF Fwd Search	OF	21	94,1	76,2	0,99	0,73	0,99	0,9	0,94	0,89	0,94	0,76	0,94	0,8
8		RF Bkwd Search	OF	24	96,7	80,4	0,99	0,72	0,99	0,9	0,96	0,88	0,96	0,8	0,96	0,83
9		Genetic	OF	19	95,5	80,4	0,98	0,68	0,98	0,89	0,95	0,86	0,95	0,8	0,95	0,83
10	SMOTE	All	OF	29	84,1	81,1	0,92	0,67	0,91	0,88	0,84	0,86	0,84	0,81	0,84	0,84
11		RF Fwd Search	OF	3	85,8	70	0,89	0,67	0,87	0,88	0,85	0,87	0,85	0,7	0,85	0,76
12		RF Bkwd Search	OF	24	84,5	83,9	0,92	0,72	0,92	0,89	0,84	0,87	0,84	0,83	0,84	0,84
13		Genetic	OF	17	83,7	89,5	0,9	0,81	0,89	0,92	0,83	0,9	0,83	0,89	0,83	0,9

When the OF algorithm was applied to the VAR2 dataset, four generated models were identified as shown in Table 6.9. The *OS-All-OF* model (line 6) had the same accuracy difference but both accuracy values were greater than that of the forward search RF algorithm. The *OS-Rf Bkwd Search-OF* model (line 8) also had both testing and validation accuracy above 90% with an accuracy difference of 5.8%. The *SMOTE-All-OF* model (line 10) produced an accuracy difference of 0.2% between training and validation datasets. However, the training accuracy was 4.3% lower than that of the forward search RF algorithm.

Table 6.9: Summary analysis for OF algorithm – Experiment-2IP-Algorithm [OF] – VAR2

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Forward Search	RF	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
2	None	All	OF	34	82,4	90,2	0,75	0,79	0,84	0,9	0,78	0,82	0,82	0,9	0,75	0,86
3		RF Fwd Search	OF	2	82,6	90,9	0,52	0,5	0,71	0,83	0,81	?	0,82	0,9	0,75	?
4		RF Bkwd Search	OF	29	85,4	90,9	0,8	0,87	0,86	0,93	0,84	0,87	0,85	0,9	0,82	0,87
5		Genetic	OF	18	86	90,9	0,81	0,87	0,87	0,93	0,85	0,87	0,86	0,9	0,83	0,87
6	OS	All	OF	34	95,7	90,9	0,99	0,9	0,99	0,93	0,95	0,9	0,95	0,9	0,95	0,9
7		RF Fwd Search	OF	13	97,6	89,5	0,99	0,75	0,98	0,89	0,97	0,89	0,97	0,89	0,97	0,89
8		RF Bkwd Search	OF	25	96,7	90,9	0,99	0,9	0,99	0,93	0,96	0,9	0,96	0,9	0,96	0,9
9		Genetic	OF	16	97,4	67,8	0,98	0,66	0,98	0,88	0,97	0,85	0,97	0,67	0,97	0,74
10	SMOTE	All	OF	34	88,6	88,8	0,96	0,84	0,96	0,92	0,88	0,91	0,88	0,88	0,88	0,89
11		RF Fwd Search	OF	3	88,8	49,6	0,91	0,53	0,89	0,85	0,89	0,86	0,88	0,49	0,88	0,59
12		RF Bkwd Search	OF	26	89	86,7	0,96	0,84	0,96	0,92	0,89	0,88	0,89	0,86	0,89	0,87
13		Genetic	OF	21	88,1	86	0,96	0,86	0,96	0,93	0,88	0,88	0,88	0,86	0,88	0,87

When the OF algorithm was applied to VAR3, two models were generated with the performances close to that of the forward search RF algorithm. These were the *None-RF Fwd Search-OF* model (line 3) as well as the *None-RF Bkwd Search-OF* model (line 4). In these cases, the accuracy differences were 4.2% and 5.6% respectively (see Table 6.10).

Table 6.10: Summary analysis for OF algorithm – Experiment-2IP-Algorithm [OF] – VAR3

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Forward Search	RF	4	92,9	88,1	0,66	0,78	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
2	None	All	OF	39	92,3	88,8	0,7	0,77	0,9	0,9	?	0,82	0,92	0,88	?	0,85
3		RF Fwd Search	OF	5	92,3	88,1	0,66	0,79	0,9	0,9	0,91	0,82	0,92	0,88	0,91	0,85
4		RF Bkwd Search	OF	15	94,4	88,8	0,79	0,67	0,93	0,87	0,94	0,84	0,94	0,88	0,92	0,86
5		Genetic	OF	11	93,7	86,7	0,83	0,68	0,94	0,87	0,92	0,85	0,93	0,86	0,92	0,85
6	OS	All	OF	39	98,8	85,3	1	0,63	1	0,86	0,98	0,85	0,98	0,85	0,98	0,85
7		RF Fwd Search	OF	5	99,2	82,5	1	0,57	1	0,85	0,99	0,83	0,99	0,82	0,99	0,82
8		RF Bkwd Search	OF	21	99,6	81,1	1	0,61	1	0,86	0,99	0,83	0,99	0,81	0,99	0,82
9		Genetic	OF	11	100	83,2	1	0,82	1	0,91	1	0,89	1	0,83	1	0,85
10	SMOTE	All	OF	39	96,9	84,6	0,99	0,58	0,99	0,85	0,97	0,84	0,97	0,84	0,97	0,84
11		RF Fwd Search	OF	9	96,5	81,8	0,98	0,63	0,98	0,87	0,96	0,84	0,96	0,81	0,96	0,82
12		RF Bkwd Search	OF	36	98,4	86	0,99	0,63	0,99	0,86	0,98	0,83	0,98	0,86	0,98	0,84
13		Genetic	OF	21	96,9	83,2	0,99	0,58	0,99	0,85	0,97	0,82	0,97	0,83	0,97	0,82

Thus, the use of genetic algorithms as part of the OF algorithm did result in predictive models that were better than or competitive with the best model obtained in Chapter 5 for this course (line 1 on Table 6.9). In this case, the models were the *OS-All-OF* model (line 6 on Table 6.9) and the *OS-RF Bkwd Search-OF* model for the VAR2 dataset (line 8 on Table 6.9). Both accuracies (training and validation) obtained were above 90% with the other performance measures also falling within an acceptable range. An accuracy comparison between these two models and the best model obtained in Chapter 5 is illustrated in the Figure 6.5 bar-chart.

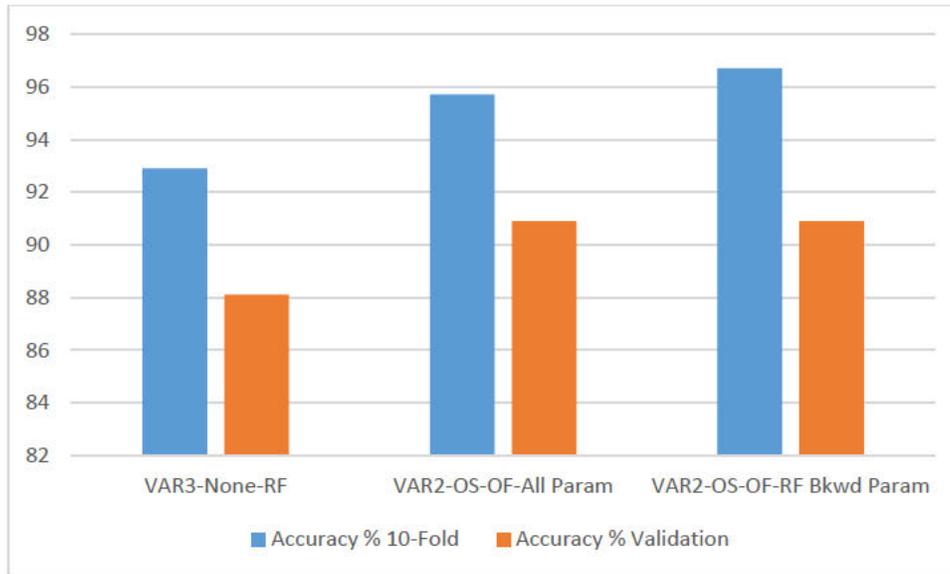


Figure 6.5: Accuracy comparison for two OF generated models with best model from Chapter 5 (VAR3-None-RF)

The ability of the GA component within the OF algorithm allowed for more variation to be considered when developing the model. This is due to the crossover and mutation operators that combined components of different trees, resulting in a greater variety of models developed within the population. Further to this, the concept of elitism allowed for the best trees to be kept in the population. The result is the development of two models with accuracy greater than 90% for both training and validation. The *OS-All-OF* model (using VAR2) generated the same accuracy difference as the model generated using the RF algorithm in Chapter 5 (4.8). The OF generated models are also better when the remaining performance measures are compared as shown in Figure 6.6.

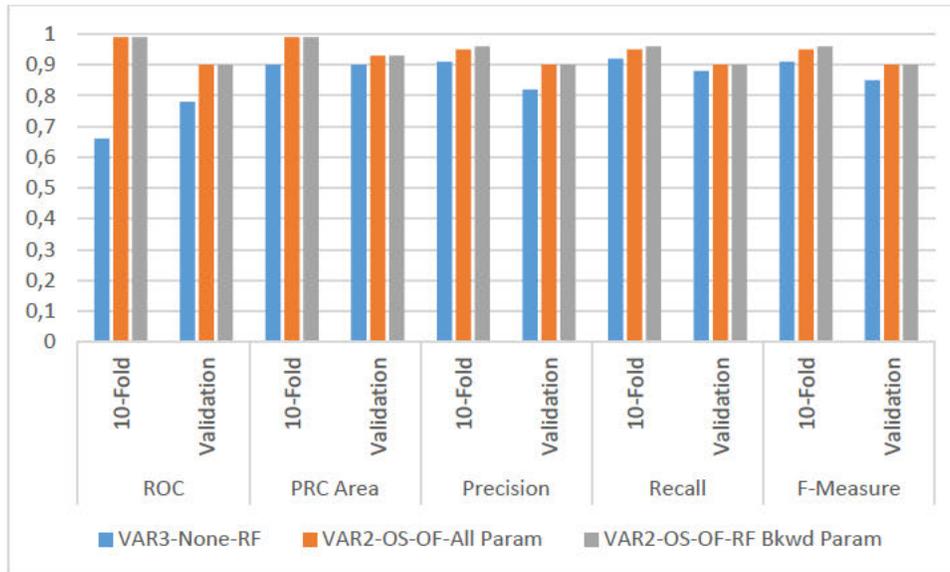


Figure 6.6: Comparison of performance measure of two OF generated models against best model from Chapter 5 (VAR3-None-RF)

The models generated by the OF algorithm are better than that of the best ISTN2IP model generated in Chapter 5 experiments. The ROC values are much better with values of 0.99 (training) and 0.9 (validation) compared to that of the Chapter 5 model (0.66 and 0.78). The PRC, precision, recall and F-measure values are also greater for the OF generated models.

6.4.3. Experiments for the ISTN3AS dataset

This section covers the genetic algorithm-based experiments for the ISTN3AS course dataset. For these experiments, the undersampling results have not been included as the performance measures did not meet the acceptance criteria for all experiments.

6.4.3.1. Experiment-3AS-FS [Genetic]

Table 6.11 shows the performance measures for the algorithms applied using GA-based feature selection (Genetic search). The accuracy achieved for each of the experiments listed in Table 6.11 is greater than 98.1% and thus outside the acceptable range for this study.

Table 6.11: Summary analysis for RF and DT algorithms – Experiment-3AS-FS [Genetic] – VAR1

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Genetic	DT	2	98,1	96,9	0,42	0,5	0,94	0,94	?	?	0,96	0,96	?	?
2		Genetic	RF	5	98,4	94,7	0,73	0,43	0,97	0,94	0,98	0,93	0,98	0,94	0,98	0,94
3	OS	Genetic	DT	18	99,1	92,5	0,99	0,54	0,99	0,94	0,99	0,94	0,99	0,92	0,99	0,93
4		Genetic	RF	21	99,8	95,5	1	0,69	1	0,95	0,99	0,93	0,99	0,95	0,99	0,94
5	SMOTE	Genetic	DT	3	99,3	94,7	0,99	0,41	0,99	0,93	0,99	0,93	0,99	0,94	0,99	0,94
6		Genetic	RF	20	99,2	95,5	0,99	0,68	0,99	0,95	0,99	0,93	0,99	0,95	0,99	0,94

Similarly, for VAR2, the training accuracy for each of the models was greater than 98.1% and hence outside the range for this study (Table 6.12).

Table 6.12: Summary analysis for RF and DT algorithms – Experiment-3AS-FS [Genetic] – VAR2

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Genetic	DT	2	98,1	97,3	0,58	0,74	0,96	0,96	0,98	0,97	0,98	0,97	0,97	0,96
2		Genetic	RF	8	98,1	97,3	0,76	0,48	0,97	0,94	0,98	0,97	0,98	0,97	0,97	0,96
3	OS	Genetic	DT	10	99,5	76,6	0,99	0,7	0,99	0,95	0,99	0,96	0,99	0,76	0,99	0,84
4		Genetic	RF	12	100	96,9	1	0,78	1	0,96	1	0,95	1	0,96	1	0,96
5	SMOTE	Genetic	DT	17	98,3	95,1	0,97	0,46	0,96	0,93	0,98	0,94	0,98	0,95	0,98	0,95
6		Genetic	RF	16	99	96,4	0,99	0,84	0,99	0,97	0,99	0,96	0,99	0,96	0,99	0,96

For VAR3, only one model (*None-Genetic-DT*) had training and validation accuracy within the acceptable range for this study (see Table 6.13). The PRC and Recall values were also acceptable but the precision and F-measure values could not be calculated. The reason for this is that the resultant decision tree consisted of a single leaf “P” and thus the model was not acceptable.

Table 6.13: Summary analysis for RF and DT algorithms – Experiment-3AS-FS [Genetic] – VAR3

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
1	None	Genetic	DT	2	97,8	96,9	0,19	0,5	0,94	0,94	?	?	0,97	0,96	?	?
2		Genetic	RF	8	99,4	96	0,99	0,55	0,99	0,94	0,99	0,93	0,99	0,96	0,99	0,95
3	OS	Genetic	DT	7	100	89,8	1	0,46	1	0,93	1	0,93	1	0,89	1	0,91
4		Genetic	RF	6	100	81,4	1	0,78	1	0,96	1	0,95	1	0,81	1	0,87
5	SMOTE	Genetic	DT	14	99,1	95,5	0,99	0,47	0,98	0,93	0,99	0,93	0,99	0,95	0,99	0,94
6		Genetic	RF	19	100	95,1	1	0,76	1	0,96	1	0,93	1	0,95	1	0,94

6.3.3.2. Experiment-3AS-Algorithm [OF]

For the OF algorithm, the experimental results were similar to that of the experiments described in Section 5.4.8. When the OF algorithm is applied to VAR1, the training accuracies were all

found to be greater than 98.1% and thus outside the acceptable range for this study. This was also the case for oversampling and SMOTE and when no sampling was applied.

For the experiments using VAR2 (see Table 6.14), the *None-All-OF* model (line 1) achieved the closest to an acceptable model. In this case, the training accuracy achieved is 97.5%, with the validation accuracy being 96.9%. The ROC values are above 0.7, with the PRC and Recall values also indicating a good model. In the case of the precision and F-Measure, the values are good for training but not determined for the validation dataset. This indicates that the prediction model was not able to predict any instances as a fail (correctly or incorrectly). Thus, based on the precision and F-Measure formulae (see formulae 5.3 and 5.5 in Chapter 5), these values cannot be determined as the denominator values are zero (formula 5.3) or undefined (formula 5.5). It should be noted that this course is dominated by assessments in the form of project work and thus further investigation is required in terms of better understanding the data requirements for conducting predictive analysis for this course.

Table 6.14: Summary analysis for OF algorithm – Experiment-3AS-Algorithm [OF] – VAR2

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation		
1	None	All	OF	35	97,5	96,9	0,76	0,71	0,97	0,95	0,97	?	0,97	0,96	0,97	?
2		RF Fwd Search	OF	4	98,1	96,9	0,8	0,69	0,97	0,95	0,97	0,95	0,98	0,96	0,98	0,96
3		RF Bkwd Search	OF	31	98,1	97,3	0,72	0,64	0,96	0,95	0,98	0,97	0,98	0,97	0,97	0,96
4		Genetic	OF	8	98,1	96,9	0,72	0,48	0,96	0,94	0,98	?	0,98	0,96	0,97	?

For VAR3, two models were noted to have acceptable accuracy, these being the *None-All-OF* model (line 1 on Table 6.15) and the *None-RF Bkwd Search-OF* model (line 3 on Table 6.15). However, similar to VAR2, the precision and F-measure values could not be calculated either for both training and validation (*None-All-OF*) or just for training (*None-RF Bkwd Search-OF*).

Table 6.15: Summary analysis for OF algorithm – Experiment-3AS-Algorithm [OF] – VAR3

	Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
					10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation		
1	None	All	OF	38	97,8	96,9	0,92	0,85	0,97	0,96	?	?	0,97	0,96	?	?
2		RF Fwd Search	OF	4	99,4	93,8	0,86	0,74	0,98	0,96	0,99	0,94	0,99	0,93	0,99	0,94
3		RF Bkwd Search	OF	6	97,8	90,7	0,79	0,41	0,97	0,93	?	0,93	0,97	0,9	?	0,92
4		Genetic	OF	8	99,4	96	0,99	0,54	0,99	0,94	0,99	0,93	0,99	0,96	0,99	0,95

6.5. Chapter summary

In the event that routine learning algorithms are not able to produce acceptable models, researchers and analysts often turn to the area of artificial intelligence as a means of improving current processes or finding better alternative processes. In Chapter 5, the DT and RF learning algorithms were able to develop acceptable prediction models for all but one of the courses in the UKZN ISTN dataset, i.e., the ISTN3AS dataset. It was also noted that better models may be possible for the ISTN100 and ISTN2IP datasets. In this chapter, the use of genetic algorithms was proposed as a means to find a prediction model for the ISTN3AS course and to find better models for the ISTN100 and ISTN2IP models.

Two approaches were used, these being the use of genetic algorithms as a part of feature selection and genetic algorithms as part of the classification process through the use of an optimized forest (OF) algorithm. For the ISTN100 course, neither approach was able to find a better model than that generated by the DT algorithm described in Chapter 5. As concluded in Chapter 5, the inclusion of LMS data may have assisted in the development of a prediction model as was shown with other courses.

With regard to ISTN2IP, an improved model was found using the OF algorithm whereby both the training and validation accuracy increased, with the accuracy difference remaining the same.

Finally, for the ISTN3AS course, the OF algorithm was able to create models with high accuracy for both training and validation datasets. However, the precision and F-measure values could not be calculated (“?” values as seen in Table 6.14 and Table 6.15) and thus the validity of the models was questioned. It was concluded that the nature of assessment in the course (presentations for group projects) led to extremely high pass rates where for the majority of cases, alternative attributes outside the LMS need to be considered and included when performing predictions for this course. These alternate attributes should revolve around the groupwork involved in the major project assessment such as the number of meetings between group members, interactions between members, individual activities or responsibilities within the group, group dynamics and presentation skills amongst others.

Table 6.16 provides an updated version of Table 5.73, listing the best models identified for each of the courses in the UKZN ISTN dataset. Only the ISTN2IP model was generated using the OF algorithm. The model achieved accuracy of greater than 90% for both training and validation while the other performance measures were also within an acceptable range.

Table 6.16: Best models from Chapter 5 and Chapter 6 experiments

Dataset	VAR	Sampling	Algorithm	Training Accuracy %	Validation Accuracy %	Accuracy Difference
ISTN100	1	SMOTE	DT	89.3	84.6	4.7
ISTN101	1	OS	RF	94.5	92.6	1.9
ISTN103	2	OS	RF	96.5	95.8	0.7
ISTN2IP	2	OS	OF	95.7	90.9	4.8
ISTN211	3	None	RF	97.4	95.4	2
ISTN212	3	None	RF	97.1	95.9	1.2
ISTN3SA	2	SMOTE	RF	94.2	96.9	2.7
ISTN3AS	No viable model found					
ISTN3SI	2	SMOTE	RF	97.1	96	1.1
ISTN3ND	2	OS	RF	96.4	97.8	1.4

Thus, in Chapter 5 and Chapter 6, learning algorithms and artificial intelligence techniques were applied to the UKZN ISTN dataset for the purpose of training and identifying learning patterns (as per research questions RQ3 and RQ4). These patterns were then applied to unseen course datasets and were able to predict student performance at an acceptable rate. In Chapter 7, the question of evaluating the predictive performance of the artefact will be answered by comparing the performance of the generated models against that of other LA/EDM studies in the literature.

Chapter 7 – Performance measure comparison with other studies

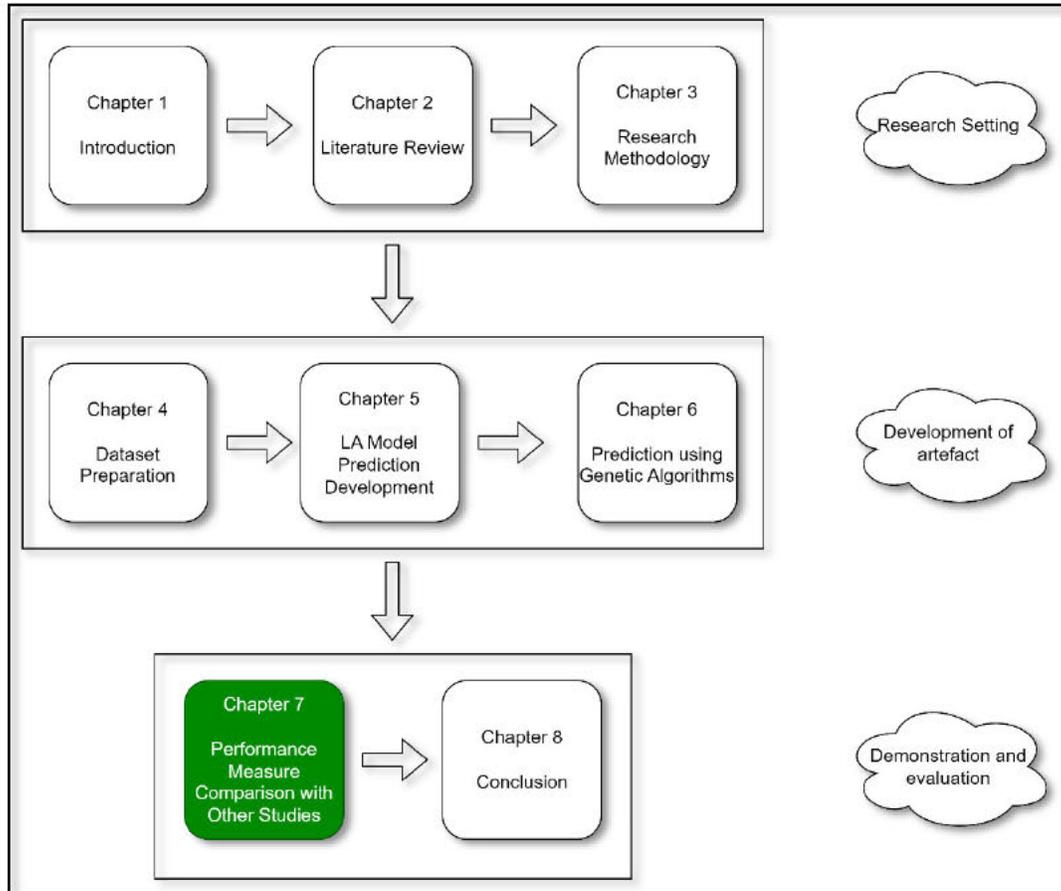


Figure 7.1: Thesis structure

7.1. Introduction

This chapter covers the demonstration and evaluation of the artefact (see Figure 7.1) and falls under the evaluation stage of the DSRM. The objective of this chapter is to address research question 5 (RQ5) stated in section 1.4, which is stated below:

RQ5 - How can the resultant information of student academic performance predictions be evaluated?

In this chapter, the performances of the prediction models generated from this study are compared to those of other studies identified in the literature. This is not an exact comparison due to a

number of variations in not only the algorithms and the parameters used but also the characteristics of the datasets and the categories of data used in these datasets. However, a comparison of this nature is useful in understanding where the study stands in terms of acceptable performance measures as well as how the dataset compares to others identified in the literature. Table 7.1 provides the comparison strategy that is followed in the chapter.

Table 7.1: Comparison strategy

Section	Description
7.2	This section provides an overview of the studies that are compared. This includes the algorithms used as well as the number of students or instances within the datasets for these studies.
7.3	This section provides a comparison of the performance measures between the prediction models generated for the UKZN ISTN 1 st year courses against other studies in the literature that focused on 1 st year courses.
7.4	A comparison is made of the performance measures between the prediction models generated for the UKZN ISTN 2 nd year courses against other studies in the literature covering 2 nd year courses.
7.5	This section covers comparisons between the prediction models generated for the UKZN ISTN 3 rd year courses against other studies in the literature covering 3 rd year courses.
7.6	In this section, the performance of the prediction models generated for the UKZN ISTN courses are compared against that of other studies that focused on technology related courses. In this context, technology related courses relate to the teaching of a variety of subjects related to technology such as programming, networking, computer engineering, IT literacy and e-commerce.
7.7	The performances of the prediction models generated from this study are compared to other studies that also applied decision tree, Random Forest and other algorithms.
7.8	This section focuses on a comparison of the accuracies generated using this study (both training and test/validation) against those of studies that also reported training and test/validation accuracies.

7.2. Studies for comparison

Forty-three (43) studies were identified as classification studies ranging from years 2013 to 2022. These studies focused on LA or EDM applications that were applied to a variety of student groups

and covered a number of courses, colleges and university instances. The studies are listed in Table 7.2 and are in terms of the algorithms that were used. The number next to the algorithm indicates the number of studies that used this algorithm.

Table 7.2: List of classification studies based on algorithms used

AdaBoost (AB) – 2
Eddin et al. (2018); Hooshyar et al. (2019)
Classification and Regression Trees (CART) – 1
Olaniyi et al. (2017)
Decision tree (DT) – 29
Ndou et al. (2020); Hasan et al. (2020); Nudelman et al. (2019); Tegegne and Alemu (2018); Bawah and Ussiph (2018); Taodzera et al. (2017); Yehuala (2015); Sunday et al. (2020); Akram et al. (2019); Al luhaybi et al. (2018); Olaniyi et al. (2017); Viloría et al. (2020); Ha et al. (2020); Hasan et al. (2018); Ribot et al. (2020); Sunday et al. (2020); Buenaño-Fernandez, Luján-Mora and Gil (2019); Hooshyar et al. (2019); Hamoud et al. (2018); Fynn and Adamiak (2018); Silva et al. (2022); Khakata et al. (2019); Adekitan and Salau (2019); Saheed et al. (2018); Eddin et al. (2018); Adejo and Connolly (2018); Wanjau and Muketha (2018); Taodzera et al. (2017); Vambe and Sibanda (2017)
Neural Network (NN) – 9
Hasan et al. (2020); Bawah and Ussiph (2018); Umar (2019); Olive et al. (2019); Jalota and Agrawal (2019); Adekitan and Salau (2019); Hooshyar et al. (2019); Adejo and Connolly (2018); Jia and Mareboyana (2013)
Partial Decision Tree algorithms (PART) – 2
Akram et al. (2019); Ha et al. (2020)
Regression (Reg) – 10
Ndou et al. (2020); Hasan et al. (2020); Jokhan et al. (2019); Hooshyar et al. (2019); Maraza-Quispe, Valderrama-Chauca, Cari-Mogrovejo, Apaza-Huanca and Sanchez-Ilabaca (2022); Adekitan and Salau (2019); Sandoval et al. (2018); Haggag et al. (2018); Eddin et al. (2018); Jayaprakash et al. (2014)
Random Forest (RF)/Random Tree (RT) – 15
Ndou et al. (2020); Hasan et al. (2020); Nudelman et al. (2019); Hooshyar et al. (2019); Akram et al. (2019); Ha et al. (2020); Hasan et al. (2018); Hamoud et al. (2018); Renò et al. (2022); Silva et al. (2022); Jalota and Agrawal (2019); Adekitan and Salau (2019); Wanjau and Muketha (2018); Eddin et al. (2018); Sandoval et al. (2018)
Sequential Minimal Optimization (SMO) – 3
Ndou et al. (2020); Ha et al. (2020); Hasan et al. (2018)
Best-First Trees (BFT) – 1
Olaniyi et al. (2017)
Decision Stump (DS) – 2
Akram et al. (2019); Hasan et al. (2018)
Feature Vector Analysis (FVA) – 1
Dorodchi et al. (2018)
Continued on next page...

Table 7.2 continued
Instance based k Algorithm (IBk) – 1
Fynn and Adamiak (2018)
Logistical Model Trees (LMT) – 1
Ndou et al. (2020)
Multilayer Perceptron (MLP) – 2
Bawah and Ussiph (2018); Ha et al. (2020)
Naïve Bayes (NB) – 13
Ndou et al. (2020); Hasan et al. (2020); Nudelman et al. (2019); Hooshyar et al. (2019); Akram et al. (2019); Al luhaybi et al. (2018); Viloría et al. (2020); Ha et al. (2020); Hasan et al. (2018); Fynn and Adamiak (2018); Jalota and Agrawal (2019); Adekitan and Salau (2019); Wanjau and Muketha (2018)
PRISM classifiers (PRISM) – 1
Akram et al. (2019)
Reduced Error Pruning Trees (RepT) – 2
Hasan et al. (2018); Hamoud et al. (2018)
Rule Induction (RI) – 1
Hasan et al. (2020)
Stochastic Gradient Descent (SGD) - 1
Eddin et al. (2018)
Simple Logistics (SL) – 1
Fynn and Adamiak (2018)
Tree Ensemble (TE) – 1
Adekitan and Salau (2019)
Support Vector Machine (SVM) - 7
Hasan et al. (2020); Jalota and Agrawal (2019); Hooshyar et al. (2019); Mahzoon et al. (2018); Eddin et al. (2018); Adejo and Connolly (2018); Oloruntoba and Akinode (2017)

Figure 7.2. reflects the number of times different algorithms from Table 7.2 were used in the various identified studies.

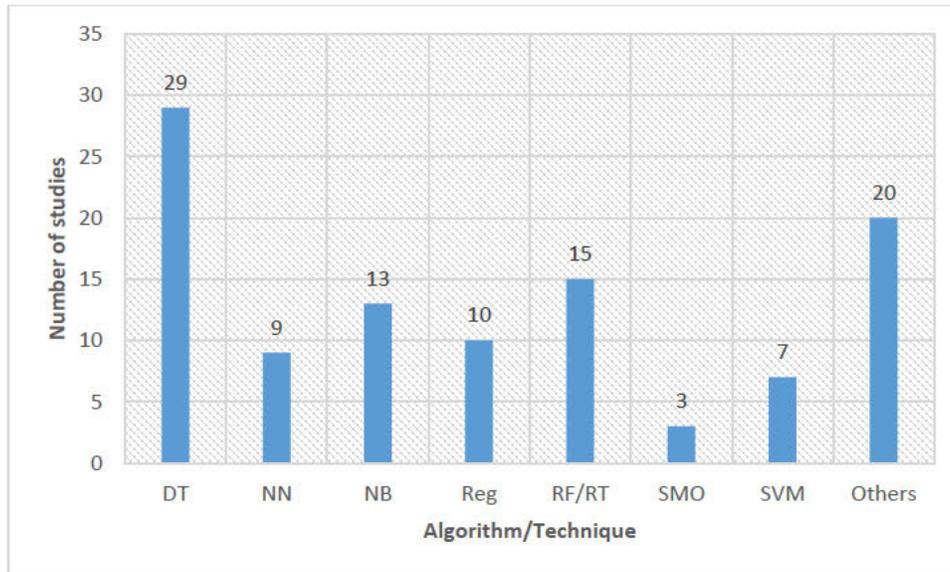


Figure 7.2: Bar chart showing distribution of studies using different algorithms/techniques

The bar representing “others” is a sum of the algorithms or techniques that appeared either once or twice (refer to Table 7.2). The histogram shows that the Decision Tree (DT) and Random Forest (RF/RT) algorithms were the most used algorithms from the listed studies. With design science following a pragmatic approach, the current study also used the DT and RF algorithms as these were not only the most commonly used algorithm but also was said to have performed well in many of the studies from the literature (see Table 2.8 and Table 2.9). Section 7.7 compares the performance measure values obtained in this study against those of other studies in terms of algorithms used.

Table 7.3 presents the above studies in terms of the number of students or instances in the dataset. The table excludes the studies by Fynn and Adamiak (2018) as well as Ogunde and Ajibade (2019) where the number of instances (student registrations) were reported rather than the number of students.

Table 7.3: Summary of studies and student numbers

Studies	No of students
Oloruntoba and Akinode (2017); Dorodchi et al. (2018); Hasan et al. (2018); Umar (2019);	Less than 100
Jia and Mareboyana (2013); Mwalumbwe and Mtebe (2017); Olaniyi et al. (2017); Vambe and Sibanda (2017); Bawah and Ussiph (2018); Saheed et al. (2018); Wanjau and Muketha (2018); Al luhaybi et al. (2018); Hamoud et al. (2018); Adejo and Connolly (2018); Khakata et al. (2019); Nudelman et al. (2019); Akram et al. (2019); Buenaño-Fernandez et al. (2019); Hooshyar et al. (2019); Jalota and Agrawal (2019); Hasan et al. (2020); Ha et al. (2020); Ribot et al. (2020); Sunday et al. (2020); Silva et al. (2022)	100 – 999
Taodzera et al. (2017); Kritzinger et al. (2018); Popoola et al. (2018); Eddin et al. (2018); Adekitan and Salau (2019); Jokhan et al. (2019); Ndou et al. (2020); Maraza-Quispe et al. (2022)	1 000 – 4 999
Tegege and Alemu (2018); Gulint and Adam (2019)	5 000 – 9 999
Jayaprakash et al. (2014); Yehuala (2015); Sandoval et al. (2018); Olive et al. (2019); Viloría et al. (2020); Renò et al. (2022)	10 000 or more

The above table is reflected as a histogram in Figure 7.3. The histogram shows that the majority of studies identified in the literature are studies whose datasets have between 100 to 999 students (inclusive).

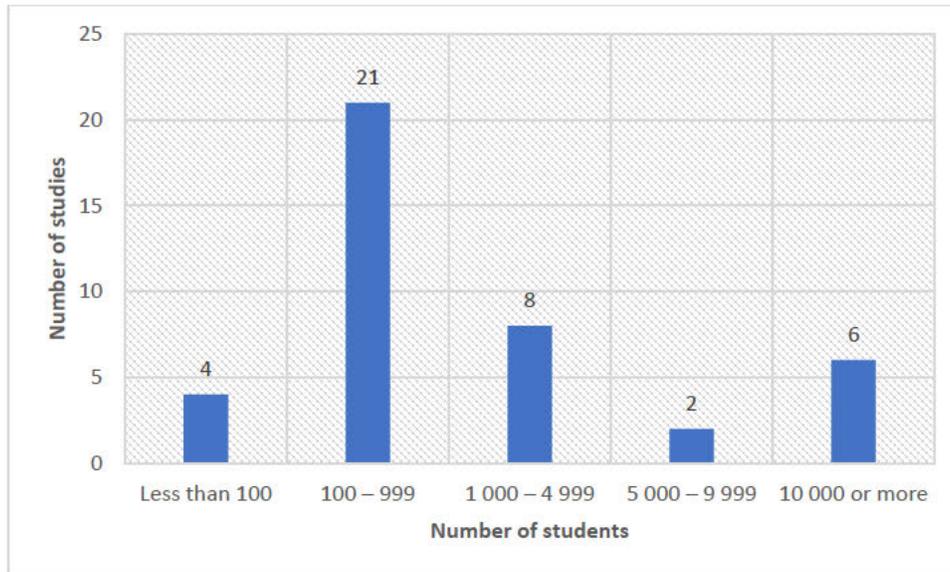


Figure 7.3: Histogram showing distribution of studies based on student numbers

For this study, the total number of students is 30 942 students over 10 courses, thus placing it in the 10 000 or more range. The breakdown of each of these courses is described in section 5.2.4. It should be noted that the other studies in the range of 10 000 or more classified students as a single group whereas this study classified students based on courses that students were registered for.

Thus, as this study focused on students within different courses, the comparisons made in Sections 7.3 to 7.5 also respectively compares 1st, 2nd and 3rd year courses covered in this study to that of studies identified in the literature.

In the following subsections, the tables listing the performance measure values are colour coded in terms of whether or not they meet the acceptance criteria for this study (Table 7.4).

Table 7.4: Colour coding for comparison performances

	Falls within the acceptable range for this study
	Does not fall within the acceptable range for this study
	Not considered due to dataset imbalance (i.e., no sampling was used)
	Value was not reported in the study

7.3. Comparison of performances related to first year courses

As there were some studies in the literature that focused on 1st year students and/or 1st year courses, it is logical to compare the performances of these studies to that of the performances of the models generated from the 1st year datasets in this study. Table 7.5 outlines the performance of the studies covering 1st year students/courses with that of the 1st year courses of this study (ISTN100, ISTN101 and ISTN103). For each of the performance measures for this study, the values for the training data are presented followed by the validation values within square brackets. Cells that are filled in black indicate that the performance measure value was not reported in the study. None of the studies listed in Table 7.5 reported on the PRC assessment metric and thus this has not been included in the table.

Table 7.5: Performance measures for studies on 1st year courses

Study	Algorithm	Accuracy %	Precision	Recall	F-Measure	ROC
ISTN100	DT	89.3[84.6]	0.89[0.82]	0.89[0.84]	0.89[0.82]	0.9[0.6]
ISTN101	RF	94.5[92.6]	0.94[0.96]	0.94[0.92]	0.94[0.93]	0.99[1]
ISTN103	RF	96.5[95.8]	0.96[0.95]	0.96[0.95]	0.96[0.95]	0.99[0.9]
Ndou et al. (2020)	RF	94.4	> 0.84	> 0.84	> 0.83	> 0.84
	LMT	91.9				
	DT	87.55				
	Reg	87.8				
	SMO	87.25				
	NB	83.9				
Hasan et al. (2020)	DT	87.4	0.74	0.99	0.93	
	Reg	87.1	0.74	0.99	0.92	
	NN	86.7	0.71	0.98	0.92	
	SVM	86.7	0.74	0.98	0.92	
	NB	84.7	0.68	0.95	0.91	
	RF	83.5	1	1	0.91	
	RI	83.4	0.88	0.92	0.9	
	NN	82.8	0.84	0.98	0.9	
Jokhan et al. (2019)	Reg	60.8				
Nudelman et al. (2019)	RF	92.4	0.92	0.66	0.84	
	DT	92.4	0.92	0.66	0.84	
	NB	74.2	0.74	0.7	0.69	
Dorodchi et al. (2018)	FVA	90				

Continued on next page...

Study	Algorithm	Accuracy %	Precision	Recall	F-Measure	ROC
Tegegne and Alemu (2018)	DT	81.4				
Bawah and Ussiph (2018)	DT	100				
	NN	96.3				
	MLP	84.7				
Taodzera et al. (2017)	DT	81.4				
Yehuala (2015)	DT	92.3	0.92			0.95

The accuracy for the 1st year IS&T courses ranged from 89% to 96.5% for training and 84.6% to 95.8% for validation. When compared to the other studies dealing with 1st year courses or students, the accuracies are acceptable when compared to that of the models generated in other studies. With the RF algorithm, the studies by Ndou et al. (2020), Nudelman et al. (2019) and Hasan et al. (2020) are 94.4%, 92.4% and 83.5%, respectively. In the case of Nudelman et al. (2019), while the accuracy, precision and F-measures are acceptable for this study, the recall value is 0.66, which would not be considered a good model by Han et al. (2012), who were aiming for recall values of 0.7 or more.

The six (6) DT algorithms have accuracies in the range 81.4%, in the case of Taodzera et al. (2017) and Tegegne and Alemu (2018), to 100%, in the case of Bawah and Ussiph (2018). Two other notable accuracies were 96.3% achieved with the Neural Network by Bawah and Ussiph (2018), as well as the LMT model by Ndou et al. (2020), where the accuracy was 91.9%. In the case of Bawah and Ussiph (2018), while the DT algorithm did produce a model with 100% accuracy, the split between training and test accuracy was not reported, and thus, how well this model applies to unseen data was not reported. Furthermore, the dataset used only consisted of 525 students and 7 attributes, making it not as complex as the dataset used in this study.

The precision, recall, F-measure and ROC values are also acceptable for this study when compared to the other studies.

7.4. Comparison of performance related to second year courses

The performances of the studies that covered 2nd year courses or students are compared to the performances of the 2nd year courses in this study. The performance measures are summarized in Table 7.6. As with Table 7.5, as no studies from the literature reported PRC values, this value was not included in the table.

Table 7.6: Performance measures for studies on 2nd year courses

Study	Algorithm	Accuracy %	Precision	Recall	F-Measure	ROC
ISTN2IP	OF	95.7[90.9]	0.95[0.9]	0.95[0.9]	0.95[0.9]	0.99[0.9]
ISTN211	RF	95.5[94.1]	0.95[0.89]	0.95[0.94]	0.93[0.91]	0.71[0.57]
ISTN212	RF	97.1[96.6]	0.97[0.96]	0.97[0.96]	0.96[0.95]	0.58[0.82]
Ndou et al. (2020)	RF	93.7	≥ 0.83	≥ 0.83	≥ 0.83	≥ 0.85
	LMT	91.7				
	DT	86.2				
	Reg	86.2				
	SMO	84.4				
	NB	83.4				
Sunday et al. (2020)	DT	87	0.8	0.87	0.82	
Akram et al. (2019)	ZeroR	80.7				
	OneR	90.8				
	DS	87.2				
	DT	94.5				
	NBT	93.6				
	PART	94.5				
	RF	95.4				
	PRISM	92.7				
Al luhaybi et al. (2018)	NB	<u>Course 1</u>				
		88.4				
	DT	<u>Course 2</u>				
		84.2				
	NB	<u>Course 3</u>				
		70.3				
	DT	<u>Course 4</u>				
		70.3				
	NB	<u>Course 1</u>				
		69.1				
DT	<u>Course 2</u>					
	75.3					
NB	<u>Course 3</u>					
	87.3					
DT	<u>Course 4</u>					
	84.1					
Olaniyi et al. (2017)	BFT					67
	DT	65.7				
	CART	66.8				

In terms of accuracy, the models generated for the study's 2nd year courses (ISTN2IP, ISTN211 and ISTN212) had the highest accuracy when compared to other studies focusing on 2nd year courses or students. This was also noted for the precision, recall and F-measure values where these performance measures were reported. For the ROC values, the ISTN211 and ISTN212 (where no sampling was used in the above cases), the values are less than that of the study by Ndou et al. (2020), where the dataset is balanced. However, as observed by Ma and He (2013), the ROC values can be adversely affected in the case of imbalanced datasets, which is the case for the ISTN211 and ISTN212 datasets. The PRC values, in this instance, do indicate that the model is good and the similar accuracy when applied to the validation dataset further indicate that the model is reliable.

7.5. Comparison of performance related to third year courses

In this section, the performances of the algorithms when applied to the 3rd year courses in the study are compared to similar 3rd year course or student studies identified in the literature. These performance measures are shown in Table 7.7. The F-measure and PRC performance measures were not included here as no studies listed in the table reported on these performance measures. The ISTN3AS performance measures are from the closest model that could meet the acceptance criteria.

Table 7.7: Performance measures for studies on 3rd year courses or students

Study	Algorithm	Accuracy %	Precision	Recall	ROC
ISTN3SA	RF	94[96]	0.94[0.97]	0.94[0.96]	0.97[0.75]
ISTN3AS	OF	97.5[96.9]	0.97[?]	0.97[0.96]	0.76[0.71]
ISTN3SI	RF	97.1[96]	0.97[0.95]	0.97[0.96]	0.99[0.7]
ISTN3ND	RF	96.4[97.8]	0.96[0.96]	0.96[0.97]	0.99[0.98]
Viloria et al. (2020)	DT	79.8	0.81		
	NB	77.9	0.83		
	OneR	75.8	0.83		
Ndou et al. (2020)	RF	95.4			>= 0.89
	LMT	93.1			
	DT	91.4			
	Reg	90.7			
	SMO	89.2			
	NB	84.4			
Continued on next page...					

Study	Algorithm	Accuracy %	Precision	Recall	ROC
Ha et al. (2020)	NB	86.1	0.64	0.61	
	MLP	86.1	0.56	0.77	
	SMO	85.6	0.57	0.6	
	DT	73.4	0.53	0.4	
	RT	77.9	0.54	0.65	
	RF	80.7	0.65	0.34	
	PART	74.5	0.41	0.4	
	OneR	76.2	0.46	0.37	
Hasan et al. (2018)	DT	63.6			
	RepT	45.4			
	RF	100			
	DS	50			
	NB	90.9			
	SMO	100			

The models generated in this study result in high accuracy values for the current study and these values are higher than accuracies identified in most of the other studies. In the case of Hasan et al. (2018), where 100% accuracy was achieved for Random Forest and Support Vector Machine, the dataset consisted of only 22 students and 11 attributes, compared to the ISTN 3rd year dataset, which consisted of over 1000 students and over 30 attributes.

Therefore, the artefact developed in this study fares very well when compared to other 3rd year related studies that were identified in the literature. The only exception is the ISTN3AS course where the model generated does not have acceptable precision values (the ? indicates that no value could be calculated for the validation dataset).

7.6. Comparison of studies dealing with technology-related courses

In this section, a comparison is made between the performance measures for the courses in this study and the performance measures of datasets related to technology-based learning. This includes courses covering computer programming, e-commerce, networking and engineering amongst others. The performance measures from this study and those of other identified studies are listed in Table 7.8.

Table 7.8: Comparison of studies relating to programming-based education

Study	Algorithm	Accuracy %	Precision	Recall	F-Measure	ROC	PRC
ISTN100	DT	89[84.6]	0.89[0.82]	0.89[0.84]	0.89[0.82]	0.9[0.6]	0.86[0.78]
ISTN101	RF	94.5[92.6]	0.94[0.96]	0.94[0.92]	0.94[0.93]	0.99[1]	0.99[1]
ISTN103	RF	96.5[95.8]	0.96[0.95]	0.96[0.95]	0.96[0.95]	0.99[0.9]	0.99[0.96]
ISTN2IP	OF	95.7[90.9]	0.95[0.9]	0.95[0.9]	0.95[0.9]	0.99[0.9]	0.99[0.93]
ISTN211	RF	95.5[94.1]	0.95[0.89]	0.95[0.94]	0.93[0.91]	0.71[0.57]	0.93[0.9]
ISTN212	RF	97.1[96.6]	0.97[0.96]	0.97[0.96]	0.96[0.95]	0.58[0.82]	0.94[0.96]
ISTN3SA	RF	94[96]	0.94[0.97]	0.94[0.96]	0.94[0.96]	0.97[0.75]	0.96[0.98]
ISTN3AS	OF	97.5[96.9]	0.97[?]	0.97[0.96]	0.97[?]	0.76[0.71]	0.97[0.95]
ISTN3SI	RF	97.1[96]	0.97[0.95]	0.97[0.96]	0.97[0.95]	0.99[0.7]	0.99[0.95]
ISTN3ND	RF	96.4[97.8]	0.96[0.96]	0.96[0.97]	0.96[0.98]	0.99[0.98]	0.99[0.99]
Ribot et al. (2020) – <u>BSc. IT</u>	DT	84.7	0.82	0.79	0.81		
Hasan et al. (2020) – <u>e-Commerce</u>	DT	87.4	0.74	0.99	0.93		
	Reg	87.1	0.74	0.99	0.92		
	NN	86.7	0.71	0.98	0.92		
	SVM	86.7	0.74	0.98	0.92		
	NB	84.7	0.68	0.95	0.91		
	RF	83.5	1	1	0.91		
	RI	83.4	0.88	0.92	0.9		
	NN	82.8	0.84	0.98	0.9		
Sunday et al. (2020) – <u>Intro to Programming</u>	DT	87	0.8	0.87	0.82		
Umar (2019) – <u>Computer Networking</u>	NN	73.6	0.81	0.73	0.73	0.8	
Buenaño-Fernandez et al. (2019) – <u>Computer Engineering</u>	DT	95.7	0.96	0.95	0.95	0.96	0.97
Jokhan et al. (2019) – <u>IT Literacy</u>	Reg	60.8					
Nudelman et al. (2019) – <u>Computer Science</u>	RF	92.4	0.92	0.66	0.84		
	DT	92.4	0.92	0.66	0.84		
	NB	74.2	0.74	0.7	0.69		

Continued on next page...

Table 7.8 continued							
Study	Algorithm	Accuracy %	Precision	Recall	F-Measure	ROC	PRC
Akram et al. (2019) - <u>Programming</u>	ZeroR	80.7					
	OneR	90.8					
	DS	87.2					
	DT	94.5					
	NBT	93.6					
	PART	94.5					
	RF	95.4					
	PRISM	92.7					
Hamoud et al. (2018) – <u>CS + IT</u>	DT		0.61	0.62			
	RT		0.61	0.61			
	RepT		0.58	0.6			
Fynn and Adamiak (2018) – <u>Science, engineering and technology</u>	ZeroR	90.2					
	OneR	90.2					
	NB	90.2					
	IBk	85.7					
	SL	90.2					
	DT	90.5					
Dorodchi et al. (2018) - <u>Programming</u>	FVA	90					
Bawah and Ussiph (2018) – <u>Computer Science</u>	DT	100					
	NN	96.3					
	MLP	84.7					
Hasan et al. (2018) – <u>e-Commerce</u>	DT	63.6					
	RepT	45.4					
	RF	100					
	DS	50					
	NB	90.9					
	SMO	100					
Olaniyi et al. (2017) – <u>Internet Technology and Programming</u>	BFT	67					
	DT	65.7					
	CART	66.8					

As with the course comparisons, the performances of the models generated from this study are very good when compared to the studies covering technology courses or technology studying students. In the case of the computer engineering study by Buenaño-Fernandez et al. (2019), while

the accuracy predicted is 95.7% with very good precision, recall, F-measure, ROC and PRC values, it should be noted that the dataset used in this study was made up of 164 students and 9 attributes and thus not as complex as the dataset used in this study. For the study by Nudelman et al. (2019), the dataset has 783 students and 17 attributes, while the study by Akram et al. (2019) used 109 students and 30 attributes. Finally, in the case of Hasan et al. (2018), the dataset consisted of only 22 students and 11 attributes.

Thus, from an educational dataset perspective, the UKZN ISTN dataset consists of a very large number of students or instances as well as between 30 and 40 attributes. This, in addition to the high-performance measure values, indicates that the artefact developed in this study has performed well when compared to other studies that also focused on educational datasets.

7.7. Comparison with other LA studies based on technique used

For this study, the Decision Tree and Random Forest algorithms were applied to the UKZN ISTN dataset. This section focuses on how the performance measures of the decision tree algorithm and the Random Forest algorithm compare against other studies in the literature that also applied either the decision tree or Random Forest algorithms. The optimized forest algorithm was covered in Chapter 6 but no studies in LA or EDM could be found that used this algorithm.

7.7.1. Decision tree algorithm

Table 7.9 shows a list of the best decision tree models produced per course from this study followed by studies identified that used decision tree algorithms.

Table 7.9: Comparison of studies that used decision tree algorithms

Study	Accuracy %	Precision	Recall	F-Measure	ROC	PRC
ISTN100	89 [84.6]	0.89[0.82]	0.89[0.84]	0.89[0.82]	0.9[0.6]	0.86[0.78]
ISTN101	89 [92]	0.87[0.92]	0.89[0.92]	0.85[0.88]	0.63[0.58]	0.84[0.86]
ISTN103	93.1 [94.5]	0.93[0.95]	0.93[0.95]	0.92[0.94]	0.72[0.74]	0.88[0.91]
ISTN2IP	94.1 [89.5]	0.94[0.89]	0.94[0.89]	0.94[0.89]	0.97[0.7]	0.96[0.88]
ISTN211	97 [90]	0.97[0.89]	0.97[0.9]	0.97[0.89]	0.97[0.69]	0.97[0.91]
ISTN212	93.1[93.9]	0.93[0.92]	0.93[0.93]	0.93[0.93]	0.92[0.79]	0.9[0.94]
ISTN3SA	92.7 [93.9]	0.92[0.97]	0.92[0.93]	0.92[0.95]	0.97[0.7]	0.96[0.97]
ISTN3AS	No viable model found using DT algorithm					
Continued on next page...						

Table 7.9 continued

Study	Accuracy %	Precision	Recall	F-Measure	ROC	PRC
ISTN3SI	94.8[96.4]	0.94[0.95]	0.94[0.96]	0.93[0.96]	0.7[0.72]	0.93[0.94]
ISTN3ND	96[93.9]	0.96[0.97]	0.96[0.94]	0.96[0.95]	0.96[0.88]	0.95[0.98]
Silva et al. (2022)	80	0.77	0.78	0.77		
Ndou et al. (2020)	87.55 86.2 91.4	> 0.84 ≥ 0.83 NR	> 0.84 ≥ 0.83 NR	> 0.83 ≥ 0.83 NR	> 0.84 ≥ 0.85 ≥ 0.89	
Hasan et al. (2020)	87.4	0.74	0.99	0.93		
Sunday et al. (2020)	87	0.8	0.87	0.82		
Viloria et al. (2020)	79.8	0.81				
Ribot et al. (2020)	84.7 (BSIT) 92.1 (Entre)	0.82 0.97	0.79 0.86	0.81 0.91		
Ha et al. (2020)	73.4	0.53	0.4			
Nudelman et al. (2019)	92.4	0.92	0.66	0.84		
Khakata et al. (2019)	84.6					
Adekitan and Salau (2019)	87.8					
Buenaño-Fernandez et al. (2019)	95.7	0.96	0.95	0.95	0.96	0.97
Akram et al. (2019)	94.5					
Tegegne and Alemu (2018)	81.4					
Bawah and Ussiph (2018)	100					
Saheed et al. (2018)	98.3	0.98	0.98	0.98		
Eddin et al. (2018)	72.8	0.9	0.8			
Adejo and Connolly (2018)	78					

Continued on next page...

Study	Accuracy %	Precision	Recall	F-Measure	ROC	PRC
Hasan et al. (2018)	63.6					
Al luhaybi et al. (2018)	84.1					
Wanjau and Muketha (2018)	84					
Fynn and Adamiak (2018)	90.5					
Taodzera et al. (2017)	81.4					
Olaniyi et al. (2017)	66.8					
Yehuala (2015)	92.3	0.92			0.95	

By viewing Table 7.9, it is evident that the performance measures of the courses in this study are higher than the majority of other studies that have used decision tree algorithms. For the study by Ribot et al. (2020), the Entrepreneurship dataset consists of 339 students and 11 attributes. The study by Saheed et al. (2018), where the decision tree algorithm produced 98% accuracy, had a dataset with 234 students and 13 attributes. The final study listed in Table 7.8, that being Yehuala (2015), obtained a model with 92.3% accuracy and high precision and ROC values. The dataset used by Yehuala (2015) consisted of a large number of students (11873 students) with 42 attributes. The performance measures from the current study are similar to that of Yehuala (2015) in terms of accuracy and ROC.

Thus, a conclusion can be made that for this study, the developed artefact that made use of the decision tree algorithm performed well when compared to other studies that also used the same or alternatives of this algorithm. This can be seen when comparing the performance measures against those from the decision tree studies in the literature. With the exception of ISTN3AS, all other performance measures for prediction models were higher than or equivalent to the performance measures reported in the literature.

7.7.2. Random forest algorithm

Table 7.10 lists the performance measures for the best Random Forest models for each of the courses in the UKZN ISTN dataset. Thereafter, the performance measures for studies from the literature that used Random Forest algorithms are listed.

Table 7.10: Comparison of studies that used Random Forest algorithms

Study	Accuracy %	Precision	Recall	F-Measure	ROC
ISTN100	85.8[83.8]	0.82[0.79]	0.85[0.83]	0.82[0.79]	0.6[0.7]
ISTN101	94.5 [92.6]	0.94[0.96]	0.94[0.92]	0.94[0.93]	0.99[1]
ISTN103	96.5 [95.8]	0.96[0.95]	0.96[0.95]	0.96[0.95]	0.99[0.9]
ISTN2IP	94.4[88.1]	0.94[0.82]	0.94[0.88]	0.92[0.85]	0.78[0.66]
ISTN211	95.5 [94.1]	0.95[0.89]	0.95[0.94]	0.93[0.91]	0.71[0.57]
ISTN212	97.1 [96.6]	0.97[0.96]	0.97[0.96]	0.96[0.95]	0.58[0.82]
ISTN3SA	94[96]	0.94[0.97]	0.94[0.96]	0.94[0.96]	0.97[0.75]
ISTN3AS	No acceptable model found using RF algorithm				
ISTN3SI	97.1 [96]	0.97[0.95]	0.97[0.96]	0.97[0.95]	0.99[0.7]
ISTN3ND	96.4 [97.8]	0.96[0.96]	0.96[0.97]	0.96[0.98]	0.99[0.98]
Renò et al. (2022)	95.2	0.92	0.97		
Silva et al. (2022)	97.5	0.9	0.9	0.89	
Jalota and Agrawal (2019)	67.4	0.67	0.67	0.67	0.86
Adekitan and Salau (2019)	87.7				
Wanjau and Muketha (2018)	82				
Eddin et al. (2018)	72.8	0.85	0.83		
Sandoval et al. (2018)	86.1			0.94	

The recent studies by Renò et al. (2022) and Silva et al. (2022) were the closest in terms of matching the performances of the Random Forest experiments conducted in this study. The study by Silva et al. (2022), similar to many other studies identified in earlier sections, had less than 500 students (in this case 200).

The dataset used by Renò et al. (2022) contained 32593 instances with a variety of students and was one of the larger datasets identified. Unlike the current study where instances were divided based on the courses in the degree, this study applied the Random Forest algorithm to the entire dataset as is. While the precision and recall values are similar for both the current study and that

by Renò et al. (2022), it was not reported as to how Reno’s model would perform when applied to unseen data.

It can be concluded that the Random Forest algorithm used in this study performed well when applied to the UKZN ISTN dataset. The performance measures obtained also indicate that the model will perform competently for future data collection within the IS&T discipline at UKZN.

7.7.3. Comparison with other techniques

The remaining studies listed in Table 7.11 lists the performance measures of studies not covered in sections 7.2 to 7.7.2. These are studies where the datasets focused on a variety of students and used techniques other than decision trees or Random Forest algorithms.

Table 7.11: LA or EDM Studies that have used other techniques

Study	Algorithm	Accuracy %	Precision	Recall	F-Measure	ROC
Silva et al. (2022)	NB	70	0.75	0.7	0.71	
	NN	88	0.88	0.88	0.88	
Maraza-Quispe et al. (2022)	Reg	89.7 – Course 1 94.2 – Course 2 93.8 – Course 3				
Olive et al. (2019)	NN	81.3			0.83	
Jalota and Agrawal (2019)	NB	64.4	0.64	0.64	0.63	0.84
	NN	76	0.77	0.76	0.76	0.89
	SVM	75.4	0.74	0.75	0.75	0.84
Adekitan and Salau (2019)	NN	85.8				
	NB	86.4				
	TE	87.8				
	Reg	89.1				
Wanjau and Muketha (2018)	NB	72				
Sandoval et al. (2018)	Reg	83.5			0.93	
Mahzoon et al. (2018)	SVM	97				
Haggag et al. (2018)	Reg	57.7				
Eddin et al. (2018)	SVM	49.2	0	0.99		
	SGD	50.7	0	1		
	Reg	69.1	0.74	0.66		
	AB	71.8	0.76	0.67		
Adejo and Connolly (2018)	NN	73.1				
	SVM	82.9				

As with the previous sections, it is evident that the performance measures obtained for this study are competitive with the studies by Mahzoon et al. (2018) and Maraza-Quispe et al. (2022). In the case of these studies as well as other studies discussed in previous sections where accuracy is greater than 90%, there is no indication of how the generated models would fare against unseen data instances. The next section discusses studies that have included performance measures for both training and test/validation datasets.

7.8. Performance comparison with studies that show training and testing accuracy

The majority of studies identified in the literature did not separately report the performance measures obtained for training and test datasets. Rather, a single value for accuracy, precision, recall or other performance measures were reported. There are studies that have reported different training performance measures and test or validation dataset measures when models are applied to the test or validation datasets. These studies, in comparison with the performance measures of this study, are discussed in this section and listed in Table 7.12. It should be noted that the model described for ISTN3AS was the closest model in terms of meeting the acceptance criteria for this study. For the ISTN3AS model, the precision and F-measure values could not be calculated and thus the model was not accepted.

Table 7.12: Studies that included both training and test/validation accuracy

	Algorithm	Training Accuracy %	Test Accuracy %	Accuracy difference	ROC Training	ROC Test
ISTN100	DT	89.3	84.6	4.7	0.9	0.6
ISTN101	RF	94.5	92.6	1.9	0.99	1
ISTN103	RF	96.5	95.8	0.7	0.99	0.9
ISTN2IP	OF	95.7	90.9	4.8	0.99	0.9
ISTN211	RF	95.5	94.1	1.4	0.71	0.57
ISTN212	RF	97.1	96.6	0.5	0.58	0.82
ISTN3SA	RF	94	96	2	0.97	0.75
ISTN3AS	OF	97.5	96.9	0.6	0.76	0.71
ISTN3SI	RF	97.1	96	1.1	0.99	0.7
ISTN3ND	RF	96.4	97.8	1.4	0.99	0.98
Vambe and Sibanda (2017)	DT	66	75	9	0.76 (average across degrees)	0.8 (average across degrees)
Oloruntoba and Akinode (2017)	SVM	94	97	3		
Jayaprakash et al. (2014)	Reg	94.2	61.9 to 85.6 (Applied to various colleges)	30.5 to 6.8		
Jia and Mareboyana (2013)	NN	94.4	93	1.3		

The performance measures for both training and validation datasets for this study compare well against those of other studies. In the case of Oloruntoba and Akinode (2017), the difference in accuracy between training and testing is 3%, with accuracy values of 94% (training) and 97% (testing). It should be noted that the dataset used was small, consisting of 89 students and 11 attributes. The study by Jia and Mareboyana (2013) used a larger dataset of 771 instances and 12 attributes. When the Neural Network was applied to this dataset, a 94.4% (training) and 93% (testing) accuracy was achieved (Jia & Mareboyana, 2013).

For the study by Jayaprakash et al. (2014), a total of 9938 instances were used for training while 5212 instances were used for testing. The dataset had fourteen (14) attributes. The study differed

from the current study in that the resultant regression model generated after training was applied to various colleges. The accuracy varied depending on the college, with the best college (Redwoods) best fitting the training model with 85.6% accuracy (Jayaprakash et al., 2014).

Based on the limited studies available, the artefact fares well in terms of the generated prediction models fitting to unseen data instances. All differences between training and validation accuracy are less than 10% and the accuracy achieved range from 84% to 97.8%.

7.9. Chapter summary

This chapter focused on addressing the fifth research question where a comparison is made between the performance of this artefact against the performances of other LA or EDM studies identified in the literature. An exact comparison cannot be made due to differing dataset characteristics and different techniques used. However, a comparison of the performance measures of this study against those of other studies is useful in understanding how the artefact model performs and whether these performance measures are acceptable or not.

The characteristics of the UKZN ISTN dataset was first compared in terms of the student and attribute count. The UKZN ISTN dataset was found to have one of the greatest number of students in terms of count, falling in the category of 10 000 or more students. With a maximum of 40 possible attributes, the UKZN ISTN dataset was also seen as complex dataset when compared to other datasets.

The respective 1st, 2nd and 3rd year courses in the UKZN ISTN dataset were compared to respective 1st year, 2nd year and 3rd year courses identified in the literature. From this year by year perspective, the performance measures were acceptable when compared to that of the literature.

The courses in the UKZN ISTN dataset were also compared to other courses that taught technology-based topics such as engineering, e-commerce, end user computing and programming. The performance measures, in most cases, were higher than the performance measures for other prediction models applied to datasets in this category (technology courses). The same was found

when the performance measures of models generated from the DT and RF algorithms were compared to the performance measures of other studies that used DT and RF algorithms.

Finally, a comparison was made against studies that reported both training and testing accuracies. The majority of best models for each of the courses in the UKZN ISTN dataset had accuracies of more than 90% for both training and validation accuracy, with an accuracy difference being less than 10%. The training accuracies, validation accuracies and accuracy differences were found to be better than most of the studies that reported training and test accuracies.

Thus, the prediction models generated by the artefact for this study performed well and were of a good standard when compared to other studies. The intention of this study is to report these performance measures and make the anonymized UKZN ISTN dataset available for other LA practitioners. These practitioners can then apply their own algorithms or research artefacts and compare their results against that of this study.

In the next and final chapter, a discussion is provided in terms of the objectives of the study, how the LA process can be improved within the discipline of IS&T at UKZN, and how this study can evolve in terms of future research.

Chapter 8 – Conclusion

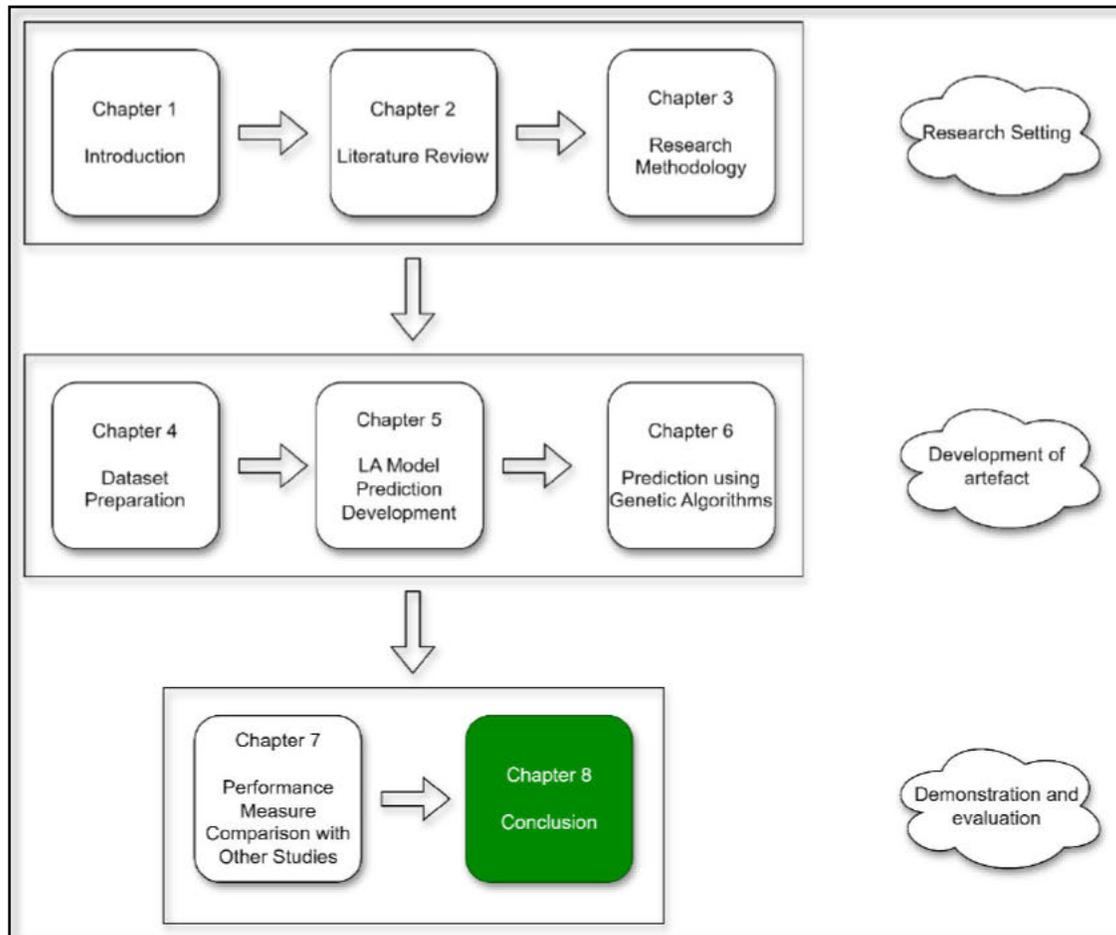


Figure 8.1: Thesis structure

8.1. Introduction

This final chapter (see Figure 8.1) provides a summary of the study, the major contributions to the body of knowledge, as well as recommendations, limitations and future work based on what has arisen from the study. The chapter forms part of the final communication stage of the DSRM.

From the researcher's perspective, an opportunity was presented to make better use of data collected within the discipline of IS&T at UKZN for the benefit of teaching and learning. Having been involved in teaching within the discipline for many years, the researcher often captured student data for no purpose other than administration, and this was the case throughout the

discipline. Learning Management Systems were used mainly for efficient lecture content dissemination with a few activities aimed toward summative assessment and student submissions.

Upon reviewing the literature, it was evident that this problem was not only at UKZN but also an emerging area of research known as Learning Analytics (LA). A small number of universities in developed countries had already implemented LA initiatives and most studies implemented small LA/EDM projects within disciplines or colleges. As an emerging area of research, it was noted that LA also has very few publicly accessible datasets to allow researchers to assist with the development and evolution of learning algorithms and artificial intelligence. Thus, the objective of this study was for the development of an artefact to outline the process of LA from a South African university context as well as the development of a dataset to allow for further research in the area of LA and/or EDM. Figure 8.2 illustrates the contents of this chapter along with each sections’s relation to other chapters.

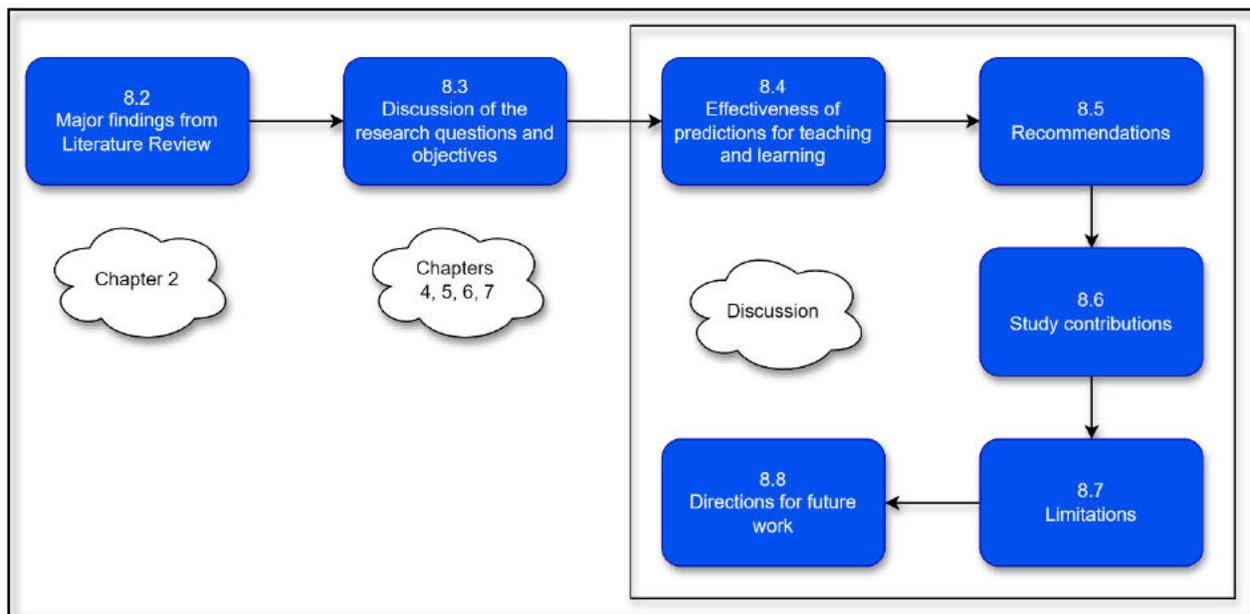


Figure 8.2: Map of Chapter 8 content

Section 8.2 provides a summary of the major findings from the literature review chapter. Section 8.3 provides a discussion relating to the five (5) research questions and objectives and how they were addressed in the research. Section 8.4 discusses the potential impact of performance prediction and how it may influence monitoring of student academic performance. Section 8.5

provides some recommendations to HEIs in order to further research on LA in the higher education domain. Section 8.6 covers the contributions of the study to the body of knowledge within LA and Section 8.7 describes the limitations of the study. Section 8.8 proposes future work that can arise from this study. Section 8.9 concludes the chapter and the research study.

8.2. Major findings of the literature review chapter

The literature review chapter provided an overview of LA, with the key findings being that Africa has been slow to undertake LA projects. As technology and infrastructure becomes available, it is imperative that African higher education institutions begin to better take advantage of the large amount of data being stored on a daily basis.

From a LA process perspective, the majority of studies focused on either data analysis (including prediction) or data preparation stages such as ethical and privacy issues, or data preparation techniques. No studies looked at the full coverage of LA from data acquisition to data analysis.

From the above findings, it was determined that the LA study must focus on data from an African university, (in this case, UKZN), and that the study cover the entire LA process from data acquisition to performance prediction.

8.3. Discussion of the research questions and objectives

By following a design science research methodology (described in Section 3.4), an artefact was developed in order to meet the objectives of the research discussed in Section 1.5. Sections 8.2.1 to 8.2.5 respectively discuss each research question and how it was addressed in the study, as well as the extent to which the objective was met.

8.3.1. How can the data from the relevant data sources (SMS, Moodle logs, registers etc.) be integrated?

This research question addressed the concept of data collection and integration of data sources. In order to meet the objectives to answer this question, ethical clearance was required, with the researcher having to gain a gatekeeper's letter from the registrar as well as sign a non-disclosure agreement. This process took longer than anticipated as all individuals were required to ensure

that the research complied with all aspects of POPIA, which had recently been introduced. The discussion on this can be found in Section 3.6.1.

Student demographics and performance data were acquired via a single MS-Excel file. Data from the Moodle LMS was acquired by downloading the relevant data files in .csv format. Section 4.2.1 describes the datasets and the attributes for each of these datasets.

The data was mapped to the files based on the student number. Activities of each student were recorded based on counting the number of times students performed different activities on the LMS (described in Section 4.2.3.2). Once integrated, student numbers were removed and replaced by a unique identifier to ensure anonymization (described in Section 4.2.2).

8.3.2. How can the integrated data be organized in preparation for data analysis?

The literature covered a number of techniques that can be used to integrate data and prepare it for data analysis (Section 2.5). Influenced by these techniques, the instances from the integrated dataset were grouped based on the courses that they originated from. Duplicate instances such as exemptions, instances with incomplete attribute values, and instances of student de-registrations were some of the instances that were removed (Section 4.2.3.4). Duplicate instances increase bias towards those instances during the learning process while instances with incomplete or missing data can adversely affect prediction capability.

To further assist with anonymization, some attributes were categorized (Section 4.2.3.3). Categorisation also improves comprehension and interpretation, as was stated by Khalil and Ebner (2016).

Thus, the first two research objectives have been addressed. Figure 8.3 is extracted from Figure 3.11 and represents the processes discussed in Chapter 4.

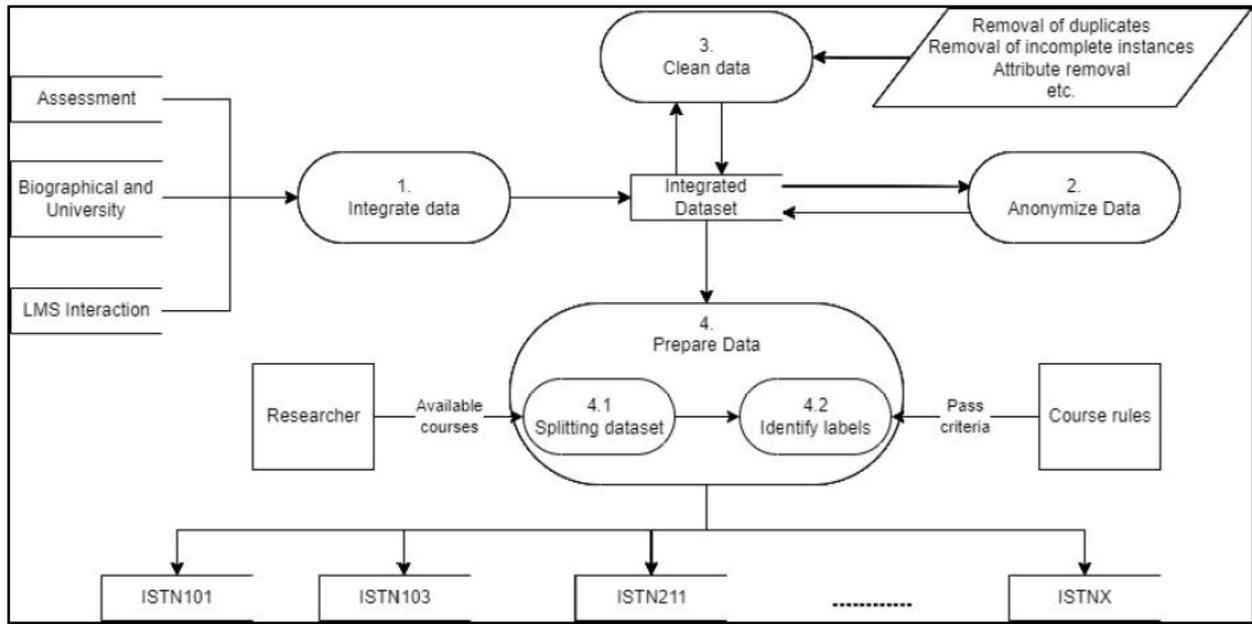


Figure 8.3: Data integration and preparation

8.3.3. How can the data be used for training towards identifying learning patterns?

Machine learning algorithms have previously been applied to datasets in order to identify learning patterns that can be used for prediction purposes. In Section 2.6, several learning algorithms that were commonly identified as being used for prediction were identified from the literature. An analysis of the literature showed that the decision tree and Random Forest algorithms were the most commonly used algorithms. By observing the accuracies in the studies described in Sections 2.6.1 to 2.6.6, these two algorithms also produced models with the best prediction accuracy in many of the studies. Thus, by following a pragmatic approach, the Decision Tree (Section 5.3.2) and Random Forest (Section 5.3.3) algorithms were applied to the UKZN ISTN dataset. Along with the process of feature selection (Section 5.3.1) and sampling techniques (Section 5.2.3), these algorithms were trained on variations of the courses within the UKZN ISTN dataset.

The experiments conducted when applying the machine learning algorithms to each of the course datasets are described in Chapter 5. From the ten (10) ISTN course datasets, a suitable prediction model could not be found for one course dataset (ISTN3AS – Section 5.4.8) and it was determined that better prediction models could be found for two courses (ISTN100 – Section 5.4.1 - and ISTN2IP – Section 5.4.4).

The literature had shown that artificial intelligence (AI) algorithms were becoming more popular to assist in establishing prediction models for complex learning situations. Chapter 6 covered the application of AI algorithms to the ISTN100, ISTN2IP and ISTN3AS models in order to find better prediction models than those found in the respective Chapter 5 experiments. Genetic algorithms (GAs) were previously used in LA/EDM studies with success by Minaei-Bidgoli and Punch (2003), Romero et al. (2009), Lakshmi et al. (2013), and Preetha (2021). Two approaches were followed, these being using GAs for feature selection (Section 6.3.1) and using a GA as part of the training process (Section 6.3.2). The genetic algorithm was able to find a better prediction model for the ISTN2IP course (Section 6.4.2). It was determined that the inclusion of Moodle data for the ISTN100 course (Section 6.4.1) would play a role in the development of a prediction model for the course. For the ISTN3AS course (Section 6.4.3), an alternate strategy for data collection would be required in the form of student activities and participation (not recorded by Moodle).

8.3.4. How can the trained data be used to predict student academic performance?

The prediction models acquired through training of the algorithms are applied to the validation datasets (unseen data instances). Based on the performance measures, models were found whose performance, when applied to the validation dataset, was similar to that of the performance obtained during training. The other performance measures such as precision, recall, F-measure, ROC and PRC also indicated that the models were of an acceptable quality. The performances of the prediction models against validation data was covered in Section 5.4 for each of the courses as well as in Section 6.4 for the ISTN100, ISTN2IP and ISTN3AS courses.

The aspects for meeting the requirements of the research questions stated in 8.3.3 and 8.3.4 are represented in the artefact. Figure 8.4 is an extraction of Figure 3.11 and represents the processes for training and validation of prediction models for the UKZN ISTN dataset.

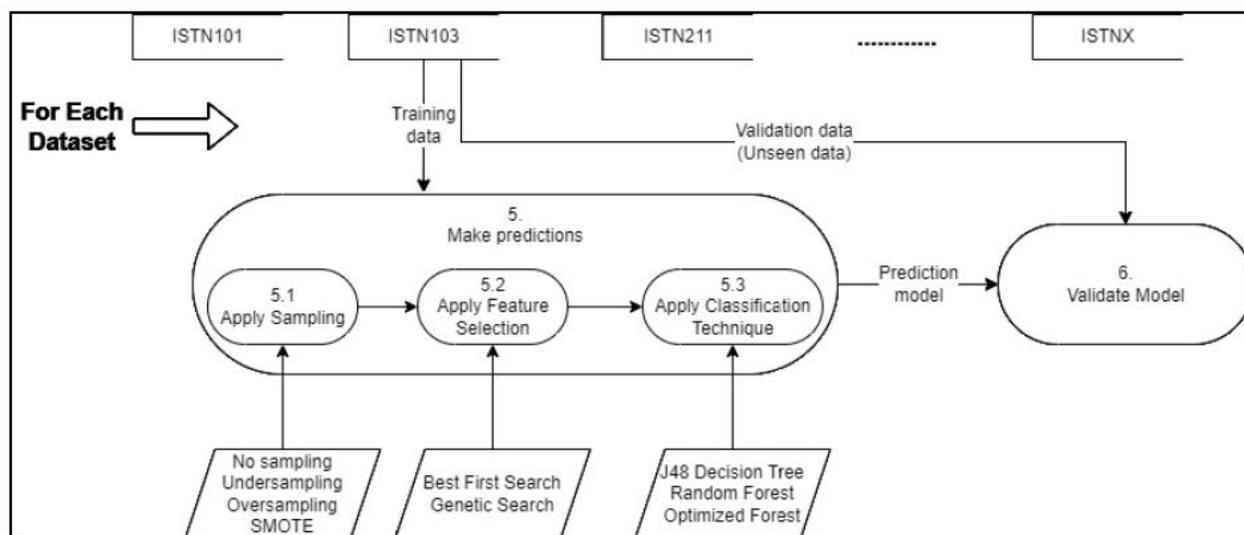


Figure 8.4: Data training and accuracy prediction

8.3.5. How can the resultant information of student academic performance predictions be evaluated?

As the UKZN ISTN dataset is a newly developed dataset in the LA field, there were no previous performances to compare the predictions to. Rather, the performances of the artefact, when applied to the UKZN ISTN datasets, were compared to that of similar EDM/LA studies. The objective of this comparison was not to determine which models were better between the literature studies and that produced in this study, but rather to understand how the models' performance measures compare to that of other studies. This question and objective were discussed in Chapter 7.

The best models for each course were identified (Table 6.16) and compared to performance measures identified in the literature from a variety of perspectives, i.e., students from 1st, 2nd and 3rd year levels of study (Sections 7.3, 7.4 and 7.5); learning algorithms used (Section 7.7); studies that separated training and test performance measures (Section 7.8); and studies that involved technology courses or degrees (Section 7.6).

From Sections 7.2 to 7.8, it was shown that the prediction models generated using the artefact for this course had acceptable accuracies when compared to accuracies seen in other LA/EDM studies. In terms of the other performance measures, the models from this study were also acceptable, although it was noted that not all studies included these performance measures when reporting on their results.

8.4. Discussion: How effective are the predictions in influencing or enabling monitoring of student academic behaviour?

Based on the five research objectives, the study has shown that using the artefact would result in the development of models that can predict whether a student is going to pass or fail a course from the IS&T discipline with a minimum of 89% accuracy, with the ISTN3AS course being the only exception. Thus, the predictions have great potential in enabling monitoring of student academic behaviour and addressing student issues before potential failure occurs. This was also seen in the ability of the prediction models to predict unseen data instances (via the validation dataset).

The resultant artefact created from the study falls under the LA application of early warning systems (discussed in Section 2.2.3.2). With a prediction accuracy that is comparable to that of other studies, the discipline of IS&T can use the prediction models to identify students that could potentially struggle at 1st, 2nd and 3rd year levels.

The inclusion of feature selection can also play a role in influencing monitoring of student academic behaviour. By identifying the most influential set of attributes (Romero et al., 2014), staff are able to identify groups of students that could potentially struggle based on attributes. For example, both the QUINTILE and CompTechSchoolYN were identified as predictive attributes for programming related courses (ISTN2IP, ISTN3AS, ISTN3SI). By analysing these attributes in more detail, students from lower quintile schools (where computer programming is not available) may require more assistance with programming. Furthermore, the use of feature selection reduces the complexity of prediction models (Kavipriya & Karthikeyan, 2019), and in the case of decision trees or optimized forest algorithms, allows for staff to better understand and evaluate these models.

Table 8.1 outlines the list of features identified in the best prediction models for each of the courses (Table 6.16). Column 2 to column 11 indicate the courses of the UKZN ISTN dataset. Rows A, B and C indicate the sampling, variation and algorithm, respectively, used to produce the best prediction model.

Rows 1 to 48 indicate each of the features identified in the UKZN ISTN dataset. A star (*) within the cell indicates that the feature was required to predict student performance in the prediction

model for that course. The blacked-out cells indicate features that were not considered, either due to the type of variation used, or the courses were not seen as prerequisites. The blue coloured features indicate course features that were seen as prerequisites. The features in the green section were obtained from student demographic and registration data, while the orange section indicates features obtained from the Moodle LMS. The last column is a count of the number of times each feature was included for the prediction models for each of the ten (10) undergraduate courses. The last row of the table indicates the number of features used to develop the prediction model for each course.

Table 8.1: Identified attributes per course

	Course --> ISTN	100	101	103	2IP	211	212	3SA	3AS	3SI	3ND	Attribute Count
A	Sampling	SMOTE	OS	OS	OS	None	None	SMOTE	None	SMOTE	OS	
B	Variation	1	1	2	2	3	3	2	2	2	2	
C	Algorithm	DT	RF	RF	OF	RF	RF	RF	OF	RF	RF	
1	ISTN101				*							1
2	ISTN102/3				*	*						2
3	ISTN2IP							*	*	*	*	4
4	ISTN211							*	*	*	*	2
5	ISTN212							*	*	*	*	3
6	BC	*									*	2
7	OT	*	*		*			*	*	*	*	7
8	QUAL		*		*			*	*		*	5
9	QUALCAT	*	*		*							3
10	SUBCAT		*					*	*	*	*	5
11	SELFFUNDED											0
12	Age Category	*	*	*	*				*	*	*	7
13	GENDER	*	*	*	*	*		*	*	*	*	9
14	RACE	*	*		*			*	*		*	6
15	RELIGIONDESC		*		*				*	*		4
16	ALIENYN		*		*			*	*		*	5
17	COUNTRYCITZDESC		*	*	*			*	*	*	*	7
18	HOMELANGDESC				*			*	*	*	*	5
19	MARITALSTATUS		*		*			*	*	*	*	5
20	MATRICTYPEDESC		*		*	*		*	*	*	*	6
21	MATRICRANGE		*		*				*	*	*	4
22	CompTechSchool?		*		*				*	*	*	5
23	QUINTILE		*	*	*			*	*	*	*	7
24	SECONDARYSCHOOL	*	*		*				*	*	*	4
25	AREA		*		*				*	*	*	4
26	RESYN		*	*	*			*	*	*	*	7
27	RESBLDOWNER		*	*	*			*	*	*	*	6
28	BURSARYYN	*	*		*			*	*	*	*	7
29	COUNCILLOANYN		*		*						*	1
30	NSFASBURSARYYN		*		*			*	*	*	*	6
31	NSFASLOANYN		*		*				*	*	*	4
32	SCHOLARSHIPYN		*		*	*			*	*	*	6
33	FUNDINGTOTALPAID	*	*		*		*	*	*	*	*	8
34	WEBREGYN	*	*	*	*			*	*	*	*	8
35	No of total clicks			*	*			*	*	*	*	5
36	File			*	*			*	*	*	*	6
37	Folder			*	*				*	*	*	5
38	Forum				*	*	*	*	*	*	*	6
39	Quiz			*			*					2
40	System			*	*		*	*	*	*	*	7
41	URL						*					1
42	Assignment				*		*		*			3
43	Kaltura Video Resource											0
44	Zoom meeting											0
45	H5P						*					1
46	Completed Activities						*					1
47	Activities Not Completed											0
48	% Completed						*					1
	Attributes per course	10	26	12	33	5	9	23	34	27	24	

Table 8.2 lists the number of times each of the attributes appeared in the best prediction models identified at the end of Chapter 6 (Table 6.16). The LMS attributes introduced during the COVID19 pandemic (Variation 3) appeared the least in the identified prediction models (along with other attributes such as COUNCILLOAN, ISTN101, SELFFUNDED and URL). It should be noted that the Kultura video resource and Zoom meetings were not automatically recorded in all cases as this option was not available on Moodle. At the end of the table, it was noted that demographic and registration data such as OT, Age Category, COUNTRYCITZDESC, QUINTILE and others appear the most times in the best prediction models. Thus, these factors should be studied in greater detail as they are able to play a significant role in identifying struggling students.

Table 8.2: Number of occurrences of parameters in best prediction models from Table 6.16

Attributes	Number of occurrences
SELFFUNDED, Kultura Video Resource, zoom meeting, Activities not completed	0
ISTN101, COUNCILLOAN, URL, H5P, Completed Activities, % Completed	1
ISTN102/3, ISTN211, BC, Quiz	2
ISTN212, QUALCAT, Assignment	3
ISTN2IP, RELIGIONDESC, MATRICRANGE, SECONDARYSCHOOL, AREA, NSFASLOANYN	4
QUAL, SUBCAT, ALIENYN, HOMELANGDESC, MARITALSTATUS, CompTechSchool?, No of Total Clicks, Folder	5
RACE, MATRICTYPEDESC, RESBLDOWNER, NSFASBURSARYYN, SCHOLARSHIPYN, File, Forum	6
OT, Age Category, COUNTRYCITZDESC, QUINTILE, RESYN, BURSARYYN, System	7
FUNDINGTOTALPAID, WEBREGYN	8
GENDER	9

Thus, a conclusion can be made that due to the high prediction accuracy of the models and the identification of attributes that play a role in prediction, the artefact has the potential to assist in

tracking or monitoring student academic performance. However, a number of challenges are still present that must be addressed, and these are discussed in the next section on recommendations.

8.5. Recommendations

As stated, Africa is fairly new to the area of LA and while the study has produced a useful artefact to develop prediction models, the following recommendations have been identified in order to further the knowledge and application base for LA. These recommendations are described below.

8.5.1. Effective data acquisition and management

A key challenge covered in Section 2.2.5 (Challenges of learning analytics) was that of data collection. Firstly, in terms of data acquisition, future researchers must be made aware that not only is the mandatory ethical clearance required, but also the gatekeeper letter that must be specific to the requirements of the study. In other words, a general statement outlining that data can be acquired from university servers is not sufficient, because of the identifying characteristics of the data that needs to be used in a study of this nature for effective training. This may require researchers to conduct meetings with university data custodians in regard to what exactly is required for the research. This is necessary in order for the data guardians to understand what data the researcher requires for the LA study (Hernández-de-Menéndez, Morales-Menendez, Escobar & Ramírez Mendoza, 2022).

Secondly, the addition of POPIA means that students' privacy is of the utmost importance and the data must be anonymized. In the case of this study, two separate sources of data were considered, these being the student biographical and registration data from UKZN Institutional Intelligence (II) as well as Moodle LMS interaction data. In order to effectively apply LA while ensuring that POPIA rules are maintained, both data sources must be anonymized in unison. This mean linking both sources of data, which inevitably increases the complexity of data entities and their relationships. Future researchers should be made aware of the requirements of POPIA from the outset of the research project and training or assistance must be provided where necessary. Efforts have been made in this regard by USAf (2020) but greater awareness is still required.

Once the areas of ethical clearance and POPIA were addressed, the data was provided by II to the researcher. However, as stated earlier (Section 4.2), the data sources were not associated or related to each other from a database perspective. Institutional Intelligence and ICS were not in the position to download and organize the Moodle LMS data and thus the researcher was required to request access from the relevant lecturers and manually download the Moodle logs and reports for each of the courses covered in this study. Thus, a recommendation would be for the availability of sufficient staff to assist researchers in data related research. The staff must be trained in areas of big data analytics in order to understand the requirements of the researcher and to assist with data management and knowledge discovery (Avella et al., 2016). This is a critical requirement if HEIs wish to take full advantage of the benefits of LA.

8.5.2. Better use of learning management systems

While collecting data from the Moodle LMS, it was observed that the course sites were mainly used as a content repository. Other features available on Moodle, such as activity completion and tracking, quizzes, assignment submission areas and other activities, were rarely utilized, resulting in reduced ability to better track student activities. The continued use of the LMS in this manner will reduce the effectiveness of any LA application. A recommendation is to train staff to better understand and apply the features of Moodle to better fit the needs and learning objectives of the course. Alternatively, the hiring of an LMS administrator is necessary for a course or for the discipline to advise and assist in management of LMS course sites. The administrator should be aware of the features of Moodle, and working with lecturing staff, should develop course sites to maximise the potential of Moodle and its ability to collect data related to students' Moodle interaction. The LMS administrator should also have data analytics experience in order to extract data requested for future LA research. This challenge of staff training and specialisation in data analytics was echoed by Prinsloo and Kaliisa (2022b).

In addition, from a data perspective, there should be a movement to align assessment data with student activities to better understand the impact of how student interaction with course content affects their academic performance with respect to particular topics covered in the course.

8.5.3. Improved communication of analytical findings between student and lecturer

Students and lecturers are key stakeholders in higher educational institutions and thus for LA to be effectively used for monitoring student academic performance, students and lecturers must be consulted on the manner in which data is presented to them. This could be in terms of frequency, type of data, as well as presentation of data. Better understanding and appreciation of the LA process by stakeholders will no doubt improve the probability of LA acceptance within UKZN. This was also noted by Guzmán-Valenzuela et al. (2021) but is seen as a challenge still to be addressed in the African context (Prinsloo & Kaliisa, 2022b).

8.6. Contributions of the study

There are several contributions that this study makes. However, the three main contributions are the development of a dataset that has a specific context (1st to 3rd year courses in the IS&T discipline), the LA process model, and the contribution to basic and applied research.

8.6.1. Development of a dataset

The literature review study by Romero and Ventura (2020) identified a total of 13 publicly available datasets. From the datasets listed, only eight (8) were still available for access. Of the eight datasets, four (4) were school-based and two (2) were MOOC-based. The research conducted in this study resulted in the development of an integrated dataset for a discipline from an HEI. The dataset is of a challenging complexity and can allow future researchers to continuously evolve the area of LA. The UKZN ISTN dataset consists of over 37000 instances with 65 attributes over ten (10) undergraduate courses. The dataset is divided into groups based on the courses, these being three (3) first year courses, three (3) second year courses and four (4) third year courses. When comparing this to 43 other datasets in the literature (Section 7.2), it was found that this dataset had one of the greatest number of students, attributes and registration instances.

The majority of the courses are imbalanced in terms of the number of passes and the number of failures, thus making the dataset more challenging in terms of finding a reliable prediction model (Ghorbani & Ghousi, 2020; Kaur et al., 2019). Unlike other datasets in the literature, the complexity of each of the courses based on imbalance level were also reported (Table 5.2) using the imbalance formula (see formula 5.1. in Section 5.2.3).

8.6.2. Learning Analytics process artefact

The study resulted in the development of an artefact in the form of an LA process model. The full process model was presented in Section 3.5.2, and for convenience, in Figure 8.5.

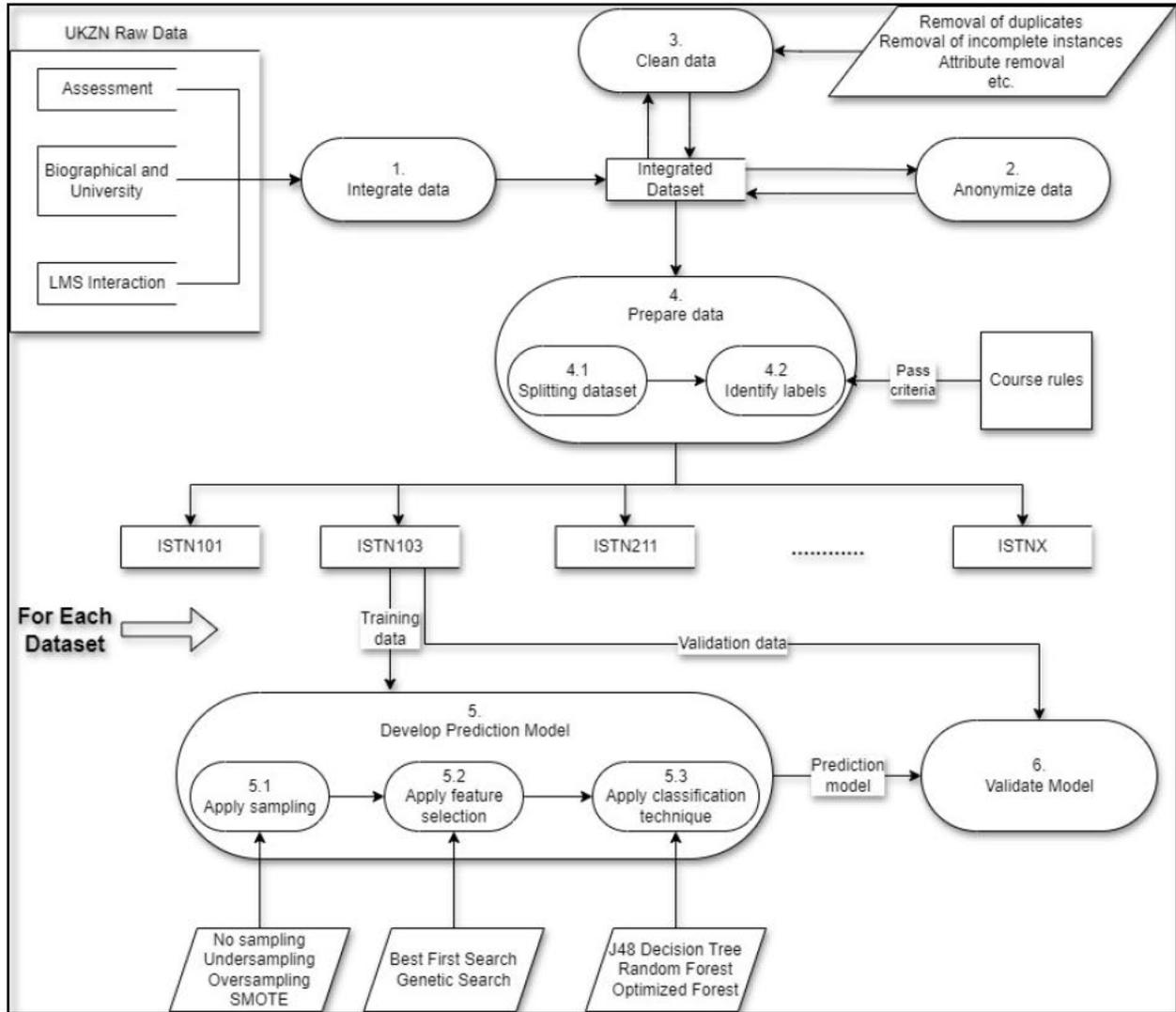


Figure 8.5: LA Process model developed from this study

The model is in the form of an adapted data flow diagram and covered each stage from data acquisition to learning algorithm application and finally, model validation. The identified data sources are integrated in process one, resulting in an integrated dataset. Process two involves the removal of data items to ensure anonymization of student information while Process three deals with cleaning of the integrated dataset. Process four is the data preparation phase where the data is split based on the current courses that form part of the degree structure (Process 4.1). This is

followed by label identification where the student final performance is converted to pass or fail class values (Process 4.2). For each of the individual courses, the data is divided into training and validation datasets (data obtained during 2021). The training data enters process five where sampling is applied (Process 5.1) to the dataset, feature selection is applied to determine the most effective attributes for prediction (Process 5.2), and a learning algorithm is applied to the training data (Process 5.3). The resultant prediction model generated from Process five is then applied to the validation dataset in Process six to validate the model.

The process model can be used as a guide for LA practitioners. Starting from data acquisition, a researcher can follow the steps of the model and understand how data moves from one process to another. Within each process, the researcher can then determine the ideal steps to clean, transform, train or validate the data, resulting in a prediction model. At each process, the researcher can choose techniques to apply based on what is available for them to implement.

8.6.3. Addition of learning analytics research within South Africa and the African continent

According to a literature survey study conducted by Guzmán-Valenzuela et al. (2021), the SciELO and WoScc research databases only identified 10 publications with 1st authors from African countries between 2013 and 2019. Waheed, Hassan, Aljohani and Wasif (2018) identified 19 publications from South Africa and 32 from Africa. Hooda and Rana (2020) identified active LA initiatives across the USA, Netherlands, UK and Australia with smaller case studies being conducted in South American, Asian and smaller European countries but did not indicate any Africa-based LA projects. Dhankhar and Solanki (2020) also did not identify any studies from the African continent. The most recent study related to Africa was an overview by Prinsloo and Kaliisa (2022b) who stated that African studies within the SoLAR community totalled to 15 studies. They further state that LA research in Africa is still in its infancy. Thus, this research, with the addition of the UKZN ISTN dataset and the LA process model artefact, improves the body of knowledge of LA for those interested in the subject for the continent of Africa.

In addition, the study serves as a valuable source of information that brings to light the changing nature of the data within the African continent. Also, it combines both basic and applied research. As applied research, it solves a problem of identifying students at risk early, with the potential to

assist and monitor performance. As basic research, the study reveals how datasets can be developed contextually, as well as the value of design science methodology in IS research, specifically in the learning analytics context. This suggests that the process model is generally applicable in this field.

8.7. Limitations

As with any research project, the limitations faced by the researcher must be acknowledged, and this is covered in this section. Three main limitations for this study are identified, these being the scope of the study, the limitations observed when working with Moodle data, and the ISTN100 as well as the ISTN3AS course. These three limitations are elaborated upon in the following subsections.

8.7.1. Scope of the study

In terms of the scope of the study, the initial objective was to obtain all student data from UKZN. However, the difficulty faced by II in providing this meant that the scope was narrowed to just the discipline of IS&T. Further to this, each course within UKZN follows different approaches in terms of teaching and learning; for example, some courses may be application-based while others are more theoretical. Furthermore, the teaching pedagogy and assessment methods will differ from one course to the next. Thus, the artefact generated cannot be generalised to all courses within UKZN and will require application to datasets in other disciplines in order to ascertain its effectiveness.

8.7.2. Working with Moodle data

From the perspective of the Moodle LMS, a number of limitations were acknowledged due to the inherent functionality of Moodle and the options available to UKZN. Firstly, Moodle allows for the tracking of whether students have opened files (such as lecture slides) or not. However, the LMS cannot determine the extent to which students have studied with that file and thus the assumption was made that students that completed this activity have, at the very least, read the content in the file. Secondly, it was noted that log data reflected the name of the student and which LMS action he or she has performed. Thus, students with the same name and surname could not be distinguished between each other, and thus the rows for both these students were removed (see Section 4.2.3.4). Thirdly, some courses did not make use of the activity tracing feature on Moodle

as they were not aware of its usefulness and how to incorporate it into their course. In these cases, it was uncertain if these activities were truly accomplished or not.

8.7.3. Availability of data sources for ISTN100 and ISTN3AS courses

The final limitation relates to the ISTN100 course and the ISTN3AS course. For ISTN100, Moodle data was not made available to the researcher and thus was not included in the analysis (see Section 5.2.4). In the case of ISTN3AS, students conduct interviews with possible clients, and are part of a team that undertakes tasks involved in project management, analysis, design and programming. While there are files that teach various programming concepts, the tasks focused around the development of the project are not recorded. Thus, this is seen as a limitation that should be addressed in the future.

8.8. Directions for future work

With the conclusion of the study, future research must be considered for the inevitable evolution of LA. This section looks at areas of future research with regard to this study.

Firstly, this research focused more on the development of a model and the use of machine learning techniques to predict student academic performance. However, the models, statistics and generated data or information may not be understood by users without required data analytics knowledge. Focus for future research must now move to the development of data visualisation to better inform students and/or staff about the analysis and prediction. The use of dashboards and summarized information will better inform individuals on progress so that intervention methods can be implemented, if required, to improve student progress (Sievert, 2020).

Secondly, for this study the data was stored within a relational database framework. In the research by Knight, Wise and Chen (2017) as well as Mahzoon et al. (2018), the use of a temporal model has the potential for improving personalised learning and to better understand a student's learning process over time. In the case of Mahzoon et al. (2018), the use of a temporal model was shown to improve accuracy by nine percent. In order to accomplish this, data must be available in the form of activities as well as time stamps of these activities. The Moodle LMS does offer activity

tracking but this must also be linked to activities outside the LMS environment, such as lecture attendance and test dates.

Thirdly, all the learning algorithms have a number of parameters that can be adjusted. For training, the number of folds was set to the default value of 10, but other values should be investigated for each course in the future. The decision trees allow for maximising the depth the tree can have as well as whether the tree should be pruned or not, amongst others. In the case of the Random Forest algorithm, parameters include the number of decision trees in the forest and the number of features (attributes) considered by each tree, amongst others. Parameter tuning is also important for artificial intelligence techniques such as genetic algorithms where parameter values such as population size, number of generations and genetic operator proportions need to be adjusted for improved algorithm performance. Future research requires that parameter tuning be implemented in order to improve the effectiveness of the learning algorithms and feature selection techniques.

Fourthly, as stated in Section 8.5, the ISTN3AS course, which focuses on project work in the field, needs to be studied further in terms of data requirements for predictive analysis. This includes resolving the issue of capturing of data for conducting interviews with possible clients, analysis and design tasks, as well as activities related to programming of the final solution for the project. These activities are not logged via the LMS and thus future research needs to delve further into capturing these activities for predictive purposes. Capturing this form of data would require questionnaires, observations and feedback techniques and the analytics would be qualitative in nature. Challenges associated with this include time taken to collect, capture and analyse the data sources.

8.9. Chapter summary

The objective of this study was to emphasise the potential of learning analytics, an area that is fairly new to the African continent. The research questions and objectives have been recalled and discussed in terms of what was covered in each of the chapters (Section 8.2). From the discussion in Section 8.2.6, the study has shown the potential of learning analytics and its capability in terms of predicting student performance. With high prediction accuracy and an indication of the models being reliable for both training and validation data, a conclusion can be made that the artefact has

the potential to predict student academic performance and potentially intercept students that could potentially fail courses in the discipline of IS&T.

Further to this, it was found that there is potential for improved prediction models when using artificial intelligence techniques. In this case, the use of genetic algorithms allowed for the generation of improved prediction models. From a perspective of LA, this provides an alternative avenue for practitioners to pursue in the event that the standard machine learning algorithms are not successful or if practitioners are searching for better solutions.

As discussed in Section 8.4.1, the dataset contributes to the LA community by allowing researchers to apply their LA or EDM technique against data from a real world higher educational institution. The dataset offers complexity in the form of a large number of student and registration instances as well being an imbalanced dataset. The dataset has already been anonymized, thereby allowing researchers to focus on the aspect of analytics.

Finally, the process model artefact (Figure 8.3) provides LA practitioners with a guide on conducting LA initiatives starting from data acquisition to data cleaning, preparation, training and finally, validation. The model allows for customisation by allowing the research to choose techniques for cleaning and preparation, machine learning techniques to apply, and division of data into training and validation.

The chapter concludes by outlining issues that need to be addressed in the form of recommendations (Section 8.3). This includes streamlining the data collection process through improved understanding of POPIA, ethical and privacy aspects, better training of staff to assist in data collection and use of the preferred LMS. Some of these recommendations are suggested as these were limitations experienced by the research when conducting this research (Section 8.5). Section 8.4 covers the key contributions of this research to the body of knowledge, these being the development of the process model and dataset as well as knowledge contribution to LA for the continent of Africa. Finally, the chapter covers future directions (Section 8.6) for this research in terms of visualisation, choice of data model to use, parameter tuning and finding of prediction models for the ISTN100 and ISTN3AS courses.

References

- Abaah Jnr, D. (2019). *Mining Educational Data to Predict Student Performance*. (B.Sc. Information and Communication Technology). University of Education, Winneba, Winneba.
- Abdullah, A., Malibari, A. & Alkhozai, M. (2014). Students' Performance Prediction System using Multi-Agent Data Mining Technique. *International Journal of Data Mining & Knowledge Management Process*, 4(5), 1.
- Adejo, O. & Connolly, T. (2017a). Learning analytics in a shared-network educational environment: Ethical issues and countermeasures. *International Journal of Advanced Computer Science and Applications*, 8(4), 156-163. doi:10.14569/IJACSA.2017.080404
- Adejo, O. & Connolly, T. (2017b). Learning Analytics in Higher Education Development: A Roadmap. *Journal of Education and Practice*, 8(15), 156-163.
- Adejo, O. & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, 10(1), 61-75.
- Adekitan, A. I. & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2), e01250.
- Adnan, M. N. & Islam, M. Z. (2016). Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm. *Knowledge-Based Systems*, 110, 86-97.
- Aggarwal, C. C. (2020). *Data Classification: Algorithms and Applications*: CRC Press.
- Agrawal, S., Vishwakarma, S. K. & Sharma, A. K. (2017). Using data mining classifier for predicting student's performance in UG level. *International Journal of Computer Applications*, 172(8), 39-44. doi:10.5120/ijca2017915201
- Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y. & Tang, Y. (2019). Predicting students' academic procrastination in blended learning course using homework submission data. *IEEE access*, 7, 102487-102498. doi:10.1109/ACCESS.2019.2930867
- Al luhaybi, M., Tucker, A. & Yousefi, L. (2018). *The Prediction of Student Failure using Classification Methods: A Case Study*. Paper presented at the Fourth International Conference on Image Processing and Pattern Recognition, Copenhagen, Denmark.
- Alasadi, S. A. & Bhaya, W. S. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 12(16), 4102-4107.
- Alexandropoulos, S.-A. N., Kotsiantis, S. B. & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, 34. doi:10.1017/S026988891800036X
- Algur, S. P., Bhat, P. & Ayachit, N. H. (2016). Educational data mining: RT and RF classification models for higher education professional courses. *International Journal of Information Engineering and Electronic Business*, 8(2), 59.
- Ali, S., Haider, Z., Munir, F., Khan, H. & Ahmed, A. (2013). Factors contributing to the students academic performance: A case study of Islamia University Sub-Campus. *American journal of educational research*, 1(8), 283-289. doi:10.12691/education-1-8-3
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In M. W. Berry, A. Mohamed & B. W. Hap (Eds.), *Supervised and unsupervised learning for data science* (1 ed., pp. 3-21). Switzerland: Springer.
- Alzahrani, A. S., Tsai, Y.-S., Iqbal, S., Marcos, P. M. M., Scheffel, M., Drachsler, H., . . . Gasevic, D. (2023). Untangling connections between challenges in the adoption of learning analytics in higher

- education. *Education and Information Technologies*, 28(4), 4563-4595. doi:10.1007/s10639-022-11323-x
- Amigud, A., Arnedo-Moreno, J., Daradoumis, T. & Guerrero-Roldan, A.-E. (2017). Using Learning Analytics for Preserving Academic Integrity. *The International Review of Research in Open and Distributed Learning*, 18(5). doi:<https://doi.org/10.19173/irrodl.v18i5.3103>
- Amrieh, E. A., Hamtini, T. & Aljarah, I. (2016). Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 9(8), 119-136. doi:10.14257/ijdta.2016.9.8.13
- Anuradha, C. & Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, 8(15), 1-12.
- Anuradha, C. & Velmurugan, T. (2016). Fast Boost Decision Tree Algorithm: A novel classifier for the assessment of student performance in Educational data. *Ciencia e Tecnica Vitivinicola*, 31(11), 139-155.
- Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach. *Information Systems Journal*, 16(1), 55-78.
- Asif, R., Hina, S. & Haque, S. I. (2017). Predicting Student Academic Performance using Data Mining Methods. *International Journal of Computer Science and Network Security*, 17(5), 187-191.
- Asif, R., Merceron, A., Ali, S. A. & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, 177-194.
- Avcı, Ü. & Ergün, E. (2019). Online students' LMS activities and their effect on engagement, information literacy and academic performance. *Interactive Learning Environments*, 30(3), 1-14. doi:<http://dx.doi.org/10.1080/10494820.2019.1636088>
- Avella, J. T., Kebritchi, M., Nunn, S. G. & Kanai, T. (2016). Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. *Online Learning*, 20(2), 13-29.
- Axelsen, M., Redmond, P., Heinrich, E. & Henderson, M. (2020). The evolving field of learning analytics research in higher education. *Australasian journal of educational technology*, 36(2), 1-7.
- Badat, S. (2016). Deciphering the Meanings and Explaining the South African Higher Education Student Protests of 2015–16. *Pax Academica*, 1, 71-106.
- Baek, C. & Doleck, T. (2023). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. *Interactive Learning Environments*, 31(6), 3828-3850.
- Balbay, S. & Kilis, S. (2018). Educational Analytics on an Opencourseware. *International Online Journal of Education and Teaching*, 5(3), 673-685.
- Baskerville, R., Pries-Heje, J. & Venable, J. (2009). *Soft design science methodology*. Paper presented at the Proceedings of the 4th international conference on design science research in information systems and technology.
- Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y. & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905-971. doi:10.1007/s10639-022-11152-y
- Bawah, F. U. & Ussiph, N. (2018). Appraisal of the Classification Technique in Data Mining of Student Performance using J48 Decision Tree, K-Nearest Neighbor and Multilayer Perceptron Algorithms. *International Journal of Computer Applications*, 179(33), 39-46. doi:<http://dx.doi.org/10.5120/ijca2018916751>
- Beale, M. H., Hagan, M. T. & Demuth, H. B. (2010). *Neural Network Toolbox 7 Users Guide* (7 ed.). Massachusetts, USA: MathWorks, Inc.
- Bekkar, M. & Alitouche, T. A. (2013). Imbalanced Data Learning Approaches Review. *International Journal of Data Mining & Knowledge Management Process*, 3(4), 15.

- Ben-David, A. (2008). About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence*, 21(6), 874-882.
- Berry, M. W., Mohamed, A. & Yap, B. W. (2019). *Supervised and Unsupervised Learning for Data Science* (1 ed.). Switzerland: Springer.
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., . . . Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1), 26-31.
- Bharati, S., Rahman, M. A. & Podder, P. (2018). *Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA*. Paper presented at the Fourth International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT).
- Bollier, D. & Firestone, C. M. (2010). *The promise and peril of big data*. Washington DC, USA: The Aspen Institute.
- Bonnin, G. & Boyer, A. (2017). *Higher Education and the Revolution of Learning Analytics*.
- Boyer, A. & Bonnin, G. (2016). Higher education and the revolution of learning analytics. *Report of the International Council for Open and Distance Education (ICDE)*.
- Buenaño-Fernandez, D., Luján-Mora, S. & Gil, D. (2019). *A Hybrid Machine Learning Approach for the Prediction of Grades in Computer Engineering Students*. Paper presented at the The International Research & Innovation Forum, Rome, Italy.
- Campbell, J. P., DeBlois, P. B. & Oblinger, D. G. (2007). Academic analytics: A New Tool for a New Era. *EDUCAUSE review*, 42(4), 40.
- Chatti, M. A., Dyckhoff, A. L., Schroeder, U. & Thüs, H. (2012). A Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning*, 4(5-6), 318-331.
- Chatti, M. A. & Muslim, A. (2019). The PERLA Framework: Blending Personalization and Learning Analytics. *International review of research in open and distributed learning*, 20(1), 244-261. doi:<http://dx.doi.org/10.19173/irrodl.v20i1.3936>
- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research*, 16, 321-357.
- CHE. (2013). The Aims of Higher Education. *Kagisano*, 9, v.
- Chen, J.-F., Hsieh, H.-N. & Do, Q. H. (2014). Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural Networks. *Algorithms*, 7(4), 538-553.
- Chiramba, O. & Ndofirepi, E. (2023). Access and success in higher education: Disadvantaged students' lived experiences beyond funding hurdles at a Metropolitan South African university. *South African journal of higher education*, 37(6), 56-75.
- Clow, D. (2012). *The learning analytics cycle: closing the loop effectively*. Paper presented at the Second International Conference on Learning Analytics and Knowledge, Vancouver British Columbia, Canada.
- Creswell, J. W. (2014). The Selection of a Research Approach. In *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed., pp. 3-24). California, USA: SAGE.
- Cronholm, S. & Göbel, H. (2016). Evaluation of the Information Systems Research Framework: Empirical Evidence from a Design Science Research Project. *Electronic Journal of Information Systems Evaluation*, 19(3), 158-168.
- Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5), 904-920.
- Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F. & Alowibdi, J. S. (2017). *Predicting Student Performance using Advanced Learning Analytics*. Paper presented at the 26th International Conference on World Wide Web Companion, Perth, Australia.

- Davis, J. & Goadrich, M. (2006). *The Relationship Between Precision-Recall and ROC Curves*. Paper presented at the 23rd international conference on Machine learning, Pittsburgh, USA.
- Dawkins, R. (2018). *Learning Analytics Classroom Hacks: Examples from an Australian University*. Paper presented at the Eighth International Conference on Learning Analytics and Knowledge (LAK'18), Sydney, Australia.
- Dawson, S., Joksimovic, S., Poquet, O. & Siemens, G. (2019). *Increasing the impact of learning analytics*. Paper presented at the Proceedings of the 9th International Conference on Learning Analytics & Knowledge.
- de Barba, P. G., Malekian, D., Oliveira, E. A., Bailey, J., Ryan, T. & Kennedy, G. (2020). The importance and meaning of session behaviour in a MOOC. *Computers & Education*, 146, 103772.
- Delgado, R. & Tibau, X.-A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PloS one*, 14(9), e0222916.
- Dennis, J. M., Phinney, J. S. & Chuateco, L. I. (2005). The Role of Motivation, Parental Support, and Peer Support in the Academic Success of Ethnic Minority First-Generation College Students. *Journal of college student development*, 46(3), 223-236.
- Devadoss, S. & Foltz, J. (1996). Evaluation of Factors Influencing Student Class Attendance and Performance. *American Journal of Agricultural Economics*, 78(3), 499-507.
- Dhankhar, A. & Solanki, K. (2020). State of the Art of Learning Analytics in Higher Education. *International Journal of Emerging Trends in Engineering Research*, 8(3), 868-877.
- Dorodchi, M., Benedict, A., Desai, D., Mahzoon, M. J., MacNeil, S. & Dehbozorgi, N. (2018). *Design and Implementation of an Activity-Based Introductory Computer Science Course (CS1) with Periodic Reflections Validated by Learning Analytics*. Paper presented at the IEEE Frontiers in Education Conference (FIE), California, USA.
- Eddin, M. M. Z., Khodeir, N. A. & Elnemr, H. A. (2018). A Comparative Study of Educational Data Mining Techniques for skill-based Predicting Student Performance. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(3).
- Ekubo, E. A. & Esiefarienrhe, M. B. (2019). *Attributes of low performing students in e-learning system using clustering technique*. Paper presented at the International Conference on Computational Science and Computational Intelligence (CSCI), Las Vega, USA.
- Ellaway, R. H., Pusic, M. V., Galbraith, R. M. & Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical teacher*, 36(3), 216-222.
- Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of educational research*, 83(1), 70-120.
- Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6), 304-317.
- Ferguson, R., Clow, D., Macfadyen, L., Essa, A., Dawson, S. & Alexander, S. (2014). Setting Learning Analytics in Context: Overcoming the Barriers to Large-Scale Adoption. *Journal of Learning Analytics*, 1(3), 120-144.
- Fernández, A., Garcia, S., Herrera, F. & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of artificial intelligence research*, 61, 863-905. doi:<https://doi.org/10.1613/jair.1.11192>
- Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staaldunin, J.-P. & Gašević, D. (2019). *Counting Clicks is not Enough: Validating a Theorized Model of Engagement in Learning Analytics*. Paper presented at the Ninth International Conference on Learning Analytics & Knowledge, Arizona, USA.
- Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., . . . Warschauer, M. (2020). Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*, 44(1), 130-160.

- Fong, S., Biuk-Aghai, R. P. & Millham, R. C. (2018). *Swarm search methods in weka for data mining*. Paper presented at the Proceedings of the 2018 10th international conference on machine learning and computing.
- Foreman, S. (2017). *The LMS guidebook: Learning management systems demystified*: American Society for Training and Development.
- Fraser, W. & Killen, R. (2005). The perceptions of students and lecturers of some factors influencing academic performance at two South African universities. *Perspectives in Education*, 23(1), 25-40.
- Fynn, A. & Adamiak, J. (2018). A comparison of the utility of data mining algorithms in an open distance learning context. *South African journal of higher education*, 32(4), 81-95.
- Gao, J., Xie, C. & Tao, C. (2016). *Big Data Validation and Quality Assurance -- Issues, Challenges, and Needs*. Paper presented at the IEEE Symposium on Service-Oriented System Engineering (SOSE), Oxford, United Kingdom.
- Gašević, D., Jovanović, J., Pardo, A. & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-Reported Measures and Academic Performance. *Journal of Learning Analytics*, 4(2), 113-128.
- George, G., Osinga, E. C., Lavie, D. & Scott, B. A. (2016). Big data and Data Science Methods for Management Research. *The Academy of Management Journal*, 59(5), 1493-1507.
- Ghorbani, R. & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance using Machine Learning Techniques. *IEEE access*, 8, 67899-67911.
- Gibson, A. & Lang, C. (2018). *The pragmatic maxim as learning analytics research method*. Paper presented at the Eighth International Conference on Learning Analytics and Knowledge, Sydney, Australia.
- Giddings, L. S. & Grant, B. M. (2006). Mixed methods research for the novice researcher. *Contemporary nurse*, 23(1), 3-11.
- Golberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning* (13th Ed. Vol. 1989). Boston, USA: Addison-Wesley.
- Gray, G., McGuinness, C., Owende, P. & Hofmann, M. (2016). Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education. *Journal of Learning Analytics*, 3(2), 330-372.
- Greenleaf, G. & Cottier, B. (2020). *2020 ends a decade of 62 new data privacy laws*. Retrieved from New South Wales, Australia: <https://ssrn.com/abstract=3572611>
- Greller, W. & Drachler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Educational technology & society*, 15(3).
- Gudivada, V., Apon, A. & Ding, J. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. *International Journal on Advances in Software*, 10(1), 1-20.
- Gulati, H. (2015). *Predictive Analytics Using Data Mining Technique*. Paper presented at the Second International Conference on Computing for Sustainable Global Development, New Delhi, India.
- Gulint, K. & Adam, T. (2019). Applying a Descriptive Model to Identify Determinant Factors on Quality of Higher Education: The Case of Ethiopian University. *IOSR Journal of Computer Engineering*, 21(3), 61-69.
- Gupta, A. & Saxena, N. (2021). Need for a robust debate around the ethics of Learning Analytics. *Academia Letters*, 2. doi:<https://doi.org/10.20935/AL907>
- Guzmán-Valenzuela, C., Gómez-González, C., Rojas-Murphy Tagle, A. & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: a preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education*, 18(1), 1-19.
- Ha, D. T., Loan, P. T. T., Giap, C. N. & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction using Machine Learning Techniques. *International Journal of Computer Science and Information Security (IJCSIS)*, 18(3), 21-28.

- Haggag, M. H., Latif, M. A. & Helal, D. M. (2018). A Learning Analytics Approach for Student Performance Assessment. *International Journal of Computer Science & Information Technology (IJCSIT)*, 10(4), 79-94.
- Hamoud, A., Hashim, A. S. & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), 26-31. doi:<http://dx.doi.org/10.9781/ijimai.2018.02.004>
- Hamoud, A., Humadi, A., Awadh, W. A. & Hashim, A. S. (2017). Students' Success Prediction Based on Bayes Algorithms. *International Journal of Computer Applications*, 178(7), 6-12.
- Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*.
- Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U. & Sattar, M. U. (2020). Predicting Student Performance in Higher Educational Institutions using Video Learning Analytics and Data Mining Techniques. *Applied Sciences*, 10(11), 3894.
- Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S. & Sarker, K. U. (2018). *Student Academic Performance Prediction by using Decision Tree Algorithm*. Paper presented at the Fourth International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur.
- Hashim, H., Talab, A. A., Satty, A. & Talab, S. A. (2015). Data Mining Methodologies to Study Student's Academic Performance using the C4. 5 Algorithm. *International Journal of Computer Science and Information Security*, 5(2), 104.
- Hedding, D. W., Greve, M., Breetzke, G. D., Nel, W. & Jansen van Vuuren, B. (2020). COVID-19 and the academe in South Africa: Not business as usual.
- Hernández-de-Menéndez, M., Morales-Menendez, R., Escobar, C. A. & Ramírez Mendoza, R. A. (2022). Learning analytics: state of the art. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 16, 1209-1230.
- Hershkovitz, A. & Alexandron, G. (2020). Understanding the potential and challenges of Big Data in schools and education. *Tendencias pedagógicas*, 35, 7-17.
- Hevner, A. & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice* (Vol. 22). New York, USA: Springer.
- Hevner, A., March, S., Park, J. & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 28(1), 75-105.
- Hooda, M. & Rana, C. (2020). Learning Analytics Lens: Improving Quality of Higher Education. *International Journal of Emerging Trends in Engineering Research*, 8(5), 1626-1646.
- Hooshyar, D., Pedaste, M. & Yang, Y. (2019). Mining Educational Data to Predict Students' Performance through Procrastination Behavior. *Entropy*, 22(1), 12.
- Iam-On, N. & Boongoen, T. (2017). Generating descriptive model for student dropout: a review of clustering approach. *Human-centric Computing and Information Sciences*, 7(1), 1-24.
- Ifenthaler, D. & Widanapathirana, C. (2014). Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines. *Technology, Knowledge and Learning*, 19(1), 221-240.
- Iraj, H., Fudge, A., Faulkner, M., Pardo, A. & Kovanović, V. (2020). *Understanding students' engagement with personalised feedback messages*. Paper presented at the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany.
- Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3), 264-323.
- Jalota, C. & Agrawal, R. (2019). *Analysis of educational data mining using classification*. Paper presented at the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India.

- Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R. & Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, 1(1), 6-47.
- Jia, J.-W. & Mareboyana, M. (2013). *Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention*. Paper presented at the World Congress on Engineering and Computer Science, San Francisco, USA.
- Jiang, Z., Pan, T., Zhang, C. & Yang, J. (2021). A New Oversampling Method Based on the Classification Contribution Degree. *Symmetry*, 13(2), 194.
- Jokhan, A., Sharma, B. & Singh, S. (2019). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education*, 44(11), 1900-1911.
- Joksimović, S., Kovanović, V. & Dawson, S. (2019). The Journey of Learning Analytics. *HERDSA Review of Higher Education*, 6, 37-63.
- Kaisler, S., Armour, F., Espinosa, J. A. & Money, W. (2013). *Big data: Issues and Challenges Moving Forward*. Paper presented at the 46th Hawaii International Conference on System Sciences, Maui, Hawaii.
- Kaliisa, R., Kluge, A. & Mørch, A. I. (2022). Overcoming Challenges to the Adoption of Learning Analytics at the Practitioner Level: A Critical Analysis of 18 Learning Analytics Frameworks. *Scandinavian Journal of Educational Research*, 66(3), 367-381.
- Kappe, R. & Van der Flier, H. (2012). Predicting academic success in higher education: what's more important than being smart? *European Journal of Psychology of Education*, 27(4), 605-619.
- Kaur, H., Pannu, H. S. & Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM computing surveys (CSUR)*, 52(4), 1-36.
- Kavipriya, P. & Karthikeyan, K. (2019). Comparative Analysis of Features Extraction Strategies for Classification in Educational Data Mining. *International Journal of Recent Technology and Engineering*, 7(5S3).
- Khakata, E., Omwenga, V. & Msanjila, S. (2019). Student Performance Prediction on Internet Mediated Environments using Decision Trees. *International Journal of Computer Applications*, 181(42).
- Khalil, M. & Ebner, M. (2016). De-identification in Learning Analytics. *Journal of Learning Analytics*, 3(1), 129-138.
- Khalil, M. & Ebner, M. (2017). Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. *Journal of computing in higher education*, 29(1), 114-132.
- Knight, S., Wise, A. F. & Chen, B. (2017). Time for Change: Why Learning Analytics Needs Temporal Analysis. *Journal of Learning Analytics*, 4(3), 7-17.
- Knowles, J. E. (2015). Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.
- Koç, M. (2017). Learning Analytics of Student Participation and Achievement in Online Distance Education: A Structural Equation Modeling. *Educational Sciences: Theory & Practice*, 17(6).
- Kohavi, R. & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2), 273-324.
- Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.
- Kovanovic, V., Gašević, D., Dawson, S., Joksimovic, S. & Baker, R. (2016). Does Time-on-task Estimation Matter? Implications on Validity of Learning Analytics Findings. *Journal of Learning Analytics*, 2(3), 81-110.
- Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G. & Dawson, S. (2018). *Understand students' self-reflections through learning analytics*. Paper presented at the Eighth International Conference on Learning Analytics and Knowledge, Sydney, Australia.
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221-232.

- Kritzinger, A., Lemmens, J.-C. & Potgieter, M. (2018). Learning Strategies for First-Year Biology: Toward Moving the “Murky Middle”. *CBE—Life Sciences Education*, 17(3), ar42.
- Kumar, M. & Salal, Y. K. (2019). Systematic Review of Predicting Student's Performance in Academics. *International Journal of Engineering and Advanced Technology*, 8(3), 54-61.
- Kumar, M. & Singh, A. (2017). Evaluation of Data Mining Techniques for Predicting Student's Performance. *International Journal of Modern Education & Computer Science*, 8(8), 25-31.
- Lakshmi, T. M., Martin, A. & Venkatesan, V. P. (2013). An Analysis of Students Performance Using Genetic Algorithm. *Journal of Computer Sciences and Applications*, 1(4), 75-79.
- Leitner, P., Ebner, M. & Ebner, M. (2019). Learning Analytics Challenges to Overcome in Higher Education Institutions. In D. Ifenthaler, D.-K. Mah & J. Y.-K. Yau (Eds.), *Utilizing Learning Analytics to Support Study Success* (pp. 91-104): Springer.
- Leitner, P., Khalil, M. & Ebner, M. (2017). Learning Analytics in Higher Education—A Literature Review. In *Learning analytics: Fundamentals, applications, and trends* (pp. 1-23): Springer.
- Limbu, J. & Sah, S. (2019). Prediction on Student Academic Performance Using Hybrid Clustering Algorithm. *LBEF Research Journal of Science, Technology and Management*, 1(1).
- Liu, Q. & Khalil, M. (2023). Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology*, 54(6), 1715-1747.
- Ma, Y. & He, H. (2013). *Imbalanced learning: foundations, algorithms, and applications*. New Jersey, USA: The Institute of Electrical and Electronics Engineers, Inc.
- Macfadyen, L. P. & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*, 54(2), 588-599.
- Madasamy, K. & Ramaswami, M. (2017). Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective. *International Journal of Computational Intelligence Research*, 13(9), 2267-2281.
- Mahroeian, H., Daniel, B. & Butson, R. (2017). The perceptions of the meaning and value of analytics in New Zealand higher education institutions. *International Journal of Educational Technology in Higher Education*, 14(1), 1-17.
- Mahzoon, M. J., Maher, M. L., Eltayeb, O., Dou, W. & Grace, K. (2018). A Sequence Data Model for Analyzing Temporal Patterns of Student Data. *Journal of Learning Analytics*, 5(1), 55–74.
- Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9), 1315-1316.
- Maphosa, M. & Maphosa, V. (2020). *Educational data mining in higher education in sub-Saharan Africa: A systematic literature review and research agenda*. Paper presented at the Second International Conference on Intelligent and Innovative Computing Applications, Plaine Magnien, Mauritius.
- Maraza-Quispe, B., Valderrama-Chauca, E. D., Cari-Mogrovejo, L. H., Apaza-Huanca, J. M. & Sanchez-Illabaca, J. (2022). A Predictive Model Implemented in KNIME Based on Learning Analytics for Timely Decision Making in Virtual Learning Environments. *International Journal of Information and Education Technology*, 12(2), 91-99.
- Marongwe, N., Mbodila, M. & Kariyana, I. (2020). Determinants of student dropout in Rural South African Universities. *Journal of Gender, Information and Development in Africa (JGIDA)*, 9(1), 109-130.
- McMillan, J. H. & Reed, D. F. (1994). At-Risk Students and Resiliency: Factors Contributing to Academic Success. *The Clearing House*, 67(3), 137-140.
- Mendez, G., Ochoa, X., Chiluzza, K. & De Wever, B. (2014). Curricular Design Analysis: A Data-Driven Perspective. *Journal of Learning Analytics*, 1(3), 84-119.
- Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G. & Punch, W. F. (2003). *Predicting student performance: an application of data mining methods with an educational Web-based system*. Paper presented at the 33rd Annual Frontiers in Education (FIE), Westminster, USA.

- Minaei-Bidgoli, B. & Punch, W. F. (2003). *Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System*. Paper presented at the Genetic and Evolutionary Computation Conference, Chicago, USA.
- Mlambo, V. H., Mlambo, D. N. & Adetiba, T. C. (2021). Expansion of higher education in South Africa: problems and possibilities. *J. Soc. Soc. Anthropol*, 12, 30-40.
- Moodley, P. & Singh, R. J. (2015). Addressing student dropout rates at South African universities. *Alternation (Durban)*.
- Muljana, P. S. & Placencia, G. (2018). Learning analytics: Translating Data into “Just-in-Time” Interventions. *Scholarship of Teaching and Learning, Innovative Pedagogy*, 1(1), 6.
- Munk, M., Drlík, M., Benko, L. u. & Reichel, J. (2017). Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. *IEEE access*, 5, 8989-9004.
- Mwalumbwe, I. & Mtebe, J. S. (2017). Using Learning Analytics to Predict Students’ Performance in Moodle Learning Management System: A Case of Mbeya University of Science and Technology. *The Electronic Journal of Information Systems in Developing Countries*, 79(1), 1-13.
- Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R. & Alajrami, E. (2015). Predicting Student Performance Using Artificial Neural Network: In the Faculty of Engineering and Information Technology. *International Journal of Hybrid Information Technology*, 8(2), 221-228.
- Ndou, N., Ajoodha, R. & Jadhav, A. (2020). *Educational Data-Mining to Determine Student Success at Higher Education Institutions*. Paper presented at the Second International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Kimberley, South Africa.
- Ngqulu, N. (2018). *Investigating the Adoption and the Application of Learning Analytics in South African Higher Education Institutions (Heis)*. Paper presented at the International Conference on e-Learning, Cape Town, South Africa.
- Nguyen, A., Gardner, L. & Sheridan, D. (2020). A design methodology for learning analytics information systems: Informing learning analytics development with learning design.
- Nguyen, A., Gardner, L. A. & Sheridan, D. (2017). *A Multi-Layered Taxonomy of Learning Analytics Applications*. Paper presented at the Pacific Asia Conference on Information Systems, Langkawi, Malaysia.
- Nguyen, A., Tuunanen, T., Gardner, L. & Sheridan, D. (2021). Design principles for learning analytics information systems in higher education. *European Journal of Information Systems*, 30(5), 541-568.
- Nudelman, Z., Moodley, D. & Berman, S. (2019). *Using Bayesian Networks and Machine Learning to Predict Computer Science Success*. Paper presented at the Annual Conference of the Southern African Computer Lecturers' Association (SACLA), Gordons Bay, South Africa.
- O'Leary, D. E. (2013). Artificial Intelligence and Big Data. *IEEE Intelligent Systems*, 28(2), 96-99.
- Obitko, M. (1998, 1998). Introduction to Genetic Algorithms. Retrieved from <http://www.obitko.com/tutorials/genetic-algorithms>
- Ogunde, A. O. & Ajibade, E. (2019). A K-nearest Neighbour Algorithm-Based Recommender System for the Dynamic Selection of Elective Undergraduate Courses. *International Journal of Data Science and Analysis*, 5(6), 128-135.
- Okewu, E. & Daramola, O. (2017). *Design of a learning analytics system for academic advising in Nigerian universities*. Paper presented at the International Conference on Computing Networking and Informatics (ICCNi), Lagos, Nigeria.
- Olaniyi, A. S., Kayode, S. Y., Abiola, H. M., Tosin, S.-I. T. & Babatunde, A. N. (2017). Student’s Performance Analysis using Decision Tree Algorithms. *Annals. Computer Science Series*, 15(1), 55-62.

- Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M. & Wiese, D. (2019). A Quest for a One-Size-Fits-All Neural Network: Early Prediction of Students at Risk in Online Courses. *IEEE Transactions on Learning Technologies*, 12(2), 171-183.
- Olivier, J. (2020). Research Ethics Guidelines for Personalized Learning and Teaching through Big Data. In D. Burgos (Ed.), *Radical Solutions and Learning Analytics* (pp. 37-55). Singapore: Springer.
- Oloruntoba, S. & Akinode, J. (2017). Student Academic Performance Prediction Using Support Vector Machine. *International Journal of Engineering Sciences and Research Technology*, 6(12), 588-597.
- Ortigosa-Hernández, J., Inza, I. & Lozano, J. A. (2017). Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, 98, 32-38.
- Oussous, A., Benjelloun, F.-Z., Lahcen, A. A. & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), 431-448.
- Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1), 217-222.
- Pal, S. (2012). Mining Educational Data Using Classification to Decrease Dropout Rate of Students. *International Journal of Multidisciplinary Sciences and Education*, 3(5).
- Patil, S. (2016). Big data analytics using R. *International Research Journal of Engineering and Technology*, 3(7), 78-81.
- Patwa, N., Seetharaman, A., Sreekumar, K. & Phani, S. (2018). Learning Analytics: Enhancing the Quality of Higher Education. *Research Journal of Economics*, 2(2), 1-7.
- Peffer, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.
- Pinheiro, R., Wangenge-Ouma, G., Balbachevsky, E. & Cai, Y. (2015). The Role of Higher Education in Society and the Changing Institutionalized Features in Higher Education. In J. Huisman, H. de Boer, D. Dill & M. Souto-Otero (Eds.), *The Palgrave International Handbook of Higher Education Policy and Governance* (pp. 225-242): Springer.
- Popoola, S. I., Atayero, A. A., Badejo, J. A., John, T. M., Odukoya, J. A. & Omole, D. O. (2018). Learning analytics for smart campus: Data on academic performances of engineering undergraduates in Nigerian private university. *Data in brief*, 17, 76-94.
- Prajapati, V. (2013). *Big data Analytics with R and Hadoop*. Birmingham, England: Packt Publishing Ltd.
- Preetha, V. (2021). *Data Analysis on Student's Performance based on Health status using Genetic Algorithm and Clustering algorithms*. Paper presented at the Fifth International Conference on Computing Methodologies and Communication (ICCMC).
- Prekopcsak, Z., Makraj, G., Henk, T. & Gaspar-Papanek, C. (2011). *Radoop: Analyzing big data with rapidminer and hadoop*. Paper presented at the Second RapidMiner Community Meeting and Conference (RCOMM), Dublin, Ireland.
- Prinsloo, P. (2018). Context Matters: An African Perspective on Institutionalizing Learning Analytics. *Responses from the Global South*, 1-35.
- Prinsloo, P. & Kaliisa, R. (2022a). Data privacy on the African continent: Opportunities, challenges and implications for learning analytics. *British Journal of Educational Technology*, 894-913.
- Prinsloo, P. & Kaliisa, R. (2022b). Learning Analytics on the African Continent: An Emerging Research Focus and Practice. *Journal of Learning Analytics*, 9(2), 218-235.
- Prinsloo, P. & Slade, S. (2017). *An elephant in the learning analytics room: The obligation to act*. Paper presented at the Seventh International Learning Analytics & Knowledge Conference, Vancouver British Columbia, Canada.
- Prinsloo, P., Slade, S. & Khalil, M. (2018). *Stuck in the Middle? Making Sense of the Impact of Micro, Meso and Macro Institutional, Structural and Organisational Factors on Implementing Learning*

- Analytics*. Paper presented at the European Distance and E-Learning Network Conference, Genova, Italy.
- Pritchard, M. E. & Wilson, G. S. (2003). Using Emotional and Social Factors to Predict Student Success. *Journal of college student development*, 44(1), 18-28.
- Raghavjee, R., Subramaniam, P. R. & Govender, I. (2021). Learning Analytics in Higher Education. In *Perspectives on ICT4D and Socio-Economic Growth Opportunities in Developing Countries* (pp. 398-431): IGI Global.
- Rajasekar, D. & Verma, R. (2013). *Research methodology*: Archers & Elevators Publishing House.
- Randles, B. M., Pasquetto, I. V., Golshan, M. S. & Borgman, C. L. (2017). *Using the Jupyter notebook as a tool for open science: An empirical study*. Paper presented at the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, Canada.
- Razaque, F., Soomro, N., Shaikh, S. A., Soomro, S., Samo, J. A., Kumar, N. & Dharejo, H. (2017). *Using naïve bayes algorithm to students' bachelor academic performances analysis*. Paper presented at the Fourth IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), Salmabad, Bahrain.
- Renò, V., Stella, E., Patruno, C., Capurso, A., Dimauro, G. & Maglietta, R. (2022). Learning Analytics: Analysis of Methods for Online Assessment. *Applied Sciences*, 12(18), 9296.
- Ribot, R. R., Ribot, V. M., Perez, J. G. & Cayabyab, G. T. (2020). A Prediction Model for Student Attrition Using J48 Classification. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 8(5), 329-337.
- Rish, I. (2001). *An empirical study of the naïve Bayes classifier*. Paper presented at the International Joint Conference on Artificial Intelligence, Seattle, USA.
- Rokach, L. & Maimon, O. (2005). Decision trees. In *Data mining and Knowledge Discovery Handbook* (pp. 165-192): Springer.
- Romero, C., González, P., Ventura, S., Del Jesús, M. J. & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems with Applications*, 36(2), 1632-1644.
- Romero, C., Romero, J. R. & Ventura, S. (2014). A survey on Pre-Processing Educational Data. In *Educational Data Mining: Application and Trends* (pp. 29-64): Springer.
- Romero, C. & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.
- Romero, C., Ventura, S. & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), 368-384.
- Saheed, Y., Oladele, T., Akanni, A. & Ibrahim, W. (2018). Student performance prediction based on data mining classification techniques. *Nigerian Journal of Technology*, 37(4), 1087-1091.
- Saito, T. & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS one*, 10(3), e0118432.
- Salal, Y., Abdullaev, S. & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Technology*, 8(4C), 54-59.
- Salihoun, M. (2020). State of Art of Data Mining and Learning Analytics Tools in Higher Education. *International Journal of Emerging Technologies in Learning (IJET)*, 15(21), 58-76.
- Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K. & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. *The Internet and Higher Education*, 37, 76-89.
- Satzinger, J. W., Jackson, R. B. & Burd, S. D. (2015). *Systems analysis and design in a changing world*. Boston, United States: Cengage learning.
- Saucerman, J., Ruis, A. & Shaffer, D. W. (2017). Automating the Detection of Reflection-on-Action. *Journal of Learning Analytics*, 4(2), 212-239.

- Saunders, M., Lewis, P. & Thornhill, A. (2019). *Research Methods for Business Students* (8th ed.). United Kingdom: Pearson Education Limited.
- Schütze, H., Manning, C. D. & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39): Cambridge University Press Cambridge.
- Slater, N., Peasgood, A. & Mullan, J. (2016). *Learning Analytics in Higher Education: A review of UK and International practice*. Retrieved from Bristol, England: https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v2_0.pdf
- Sghir, N., Adadi, A. & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, 28(7), 8299-8333.
- Shahiri, A. M. & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380-1400.
- Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., . . . Baker, R. (2011). *Open Learning Analytics: an integrated & modularized platform*. Open University Press Maidenhead,
- Siemens, G. & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5), 30.
- Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*: CRC Press.
- Silva, M., Rupasingha, R. & Kumara, B. (2022). *A Comparative Study of Predicting Students' Academic Performance Using Classification Algorithms*. Paper presented at the Second International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka.
- Slade, S. & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510-1529.
- Smith, T. C. & Frank, E. (2016). Introducing Machine Learning Concepts with WEKA. In E. Mathe & S. Davis (Eds.), *Statistical Genomics* (1st ed., pp. 353-378). New York, USA: Springer.
- Ştefan, L. (2017). *Big Data to Improve the Quality of Learning in Higher Education. Opportunities, Offerings and Challenges*. Paper presented at the 13th International Scientific Conference eLearning and Software for Education, Bucharest, Hungary.
- Sunday, K., Ocheja, P., Hussain, S., Oyelere, S., Samson, B. & Agbo, F. (2020). Analyzing Student Performance in Programming Education Using Classification Techniques. *International Journal of Emerging Technologies in Learning (IJET)*, 15(2), 127-144.
- Taodzera, T., Twala, B. & Carroll, J. (2017). *Predicting engineering student success using machine learning*. Paper presented at the Fourth Biennial Conference of the South African Society for Engineering Education, Cape Town, South Africa.
- Tegegne, A. K. & Alemu, T. A. (2018). Educational data mining for students' academic performance analysis in selected Ethiopian universities. *Information Impact: Journal of Information and Knowledge Management*, 9(2), 1-15.
- Thai-Nghe, N., Busche, A. & Schmidt-Thieme, L. (2009). *Improving Academic Performance Prediction by Dealing with Class Imbalance*. Paper presented at the Ninth International Conference on Intelligent Systems Design and Applications, Pisa, Italy.
- Thanasegaran, G. (2009). Reliability and Validity Issues in Research. *Integration & Dissemination*, 4.
- Thatcher, A., Fridjhon, P. & Cockcroft, K. (2007). The relationship between lecture attendance and academic performance in an undergraduate psychology class. *South African Journal of Psychology*, 37(3), 656-660.
- Tsai, C.-W., Lai, C.-F., Chao, H.-C. & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big data*, 2(1), 21.

- Umar, M. A. (2019). Student Academic Performance Prediction Using Artificial Neural Networks: A Case Study. *International Journal of Computer Applications*, 178(48), 24-29.
- USAf. (2020). *POPIA Industry Code of Conduct: Public Universities*. Retrieved from <https://www.usaf.ac.za/wp-content/uploads/2020/09/USAf-POPIA-Guideline-Final-version-1-September-2020.pdf>
- Ustun, A. B., Zhang, K., Karaođlan-Yilmaz, F. G. & Yilmaz, R. (2023). Learning analytics based feedback and recommendations in flipped classrooms: an experimental study in higher education. *Journal of Research on Technology in Education*, 55(5), 841-857. doi:10.1080/15391523.2022.2040401
- Vaishnavi, V., Kuechler, W. & Petter, S. (2004). Design science research in information systems. *January, 20*, 2004.
- Vambe, W. T. & Sibanda, K. (2017). *Using Data Mining Techniques for the Prediction of Student Dropouts from University Science Programs*. (Masters of Science in Computer Science). University of Fort Hare, University of Fort Hare.
- Viberg, O., Hatakka, M., Bälter, O. & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98-110.
- Viloria, A., Sierra, D. M., Samper, M. G., Basto, W. O. C., Pichón, A. R., Hernández-Palma, H., . . . Kamatkar, S. J. (2020). *Dropout-Permanence Analysis of University Students Using Data Mining*. Paper presented at the International Conference on Intelligent Computing, Information and Control Systems, Secundarabad, India.
- Waddington, R. J., Nam, S., Lonn, S. & Teasley, S. D. (2016). Improving Early Warning Systems with Categorized Course Resource Usage. *Journal of Learning Analytics*, 3(3), 263-290.
- Wadesango, N. & Machingambi, S. (2011). Causes and Structural Effects of Student Absenteeism: A Case Study of Three South African Universities. *Journal of Social Sciences*, 26(2), 89-97.
- Waheed, H., Hassan, S.-U., Aljohani, N. R. & Wasif, M. (2018). A bibliometric perspective of learning analytics research landscape. *Behaviour & Information Technology*, 37(10-11), 941-957.
- Wanjau, S. K. & Muketha, G. M. (2018). Improving Student Enrollment Prediction Using Ensemble Classifiers. *International Journal of Computer Applications Technology and Research*, 7(3), 122-128.
- Werner, L., McDowell, C. & Denner, J. (2013). A First Step in Learning Analytics: Pre-processing Low-Level Alice Logging Data of Middle School Students. *Journal of Educational Data Mining*, 5(2), 11-37.
- West, D., Heath, D. & Huijser, H. (2016). Let's Talk Learning Analytics: A Framework for Implementation in Relation to Student Retention. *Online Learning Journal*, 20(2), 1-21.
- Wise, A. F. (2019). Learning Analytics: Using Data-Informed Decision-Making to Improve Teaching and Learning: Maximizing Student Engagement, Motivation, and Learning. In O. Adesope (Ed.), *Contemporary Technologies in Education* (pp. 119-143): Springer.
- Xin, O. K. & Singh, D. (2021). Development of Learning Analytics Dashboard based on Moodle Learning Management System. *International Journal of Advanced Computer Science and Applications*, 12(7).
- Yehuala, M. A. (2015). Application of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre Markos University). *International journal of scientific & technology research*, 4(4), 91-94.
- Yusuf, A. & Lawan, A. (2018). *Prediction of students' academic performance using educational data mining technique: Literature review*. Federal University Dutse, Jigawa State, Nigeria.
- Zaffar, M., Hashmani, M. A. & Savita, K. (2017, November 2017). *Performance analysis of feature selection algorithm for educational data mining*. Paper presented at the IEEE Conference on Big Data and Analytics (ICBDA), Kuching, Malaysia.
- Žukauskas, P., Vveinhardt, J. & Andriukaitienė, R. (2018). Philosophy and paradigm of scientific research. *Management culture and corporate social responsibility*, 121(13), 506-518.

Appendix A – Experimental results not listed in Chapter 5

Experiment-101-Sampling [None]-VARI – Decision tree (J48)

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	85.2	91.2	0.49	0.5	0.74	0.83	?	?	0.85	0.91	?	?
Backward Search	None	85.2	91.2	0.49	0.5	0.74	0.83	?	?	0.85	0.91	?	?

Experiment-103-Sampling [None]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	83.4	91,6	0,49	0,5	0,72	0,84	?	?	0,83	0,91	?	?
Backward Search	None	83.4	91,6	0,49	0,5	0,72	0,84	?	?	0,83	0,91	?	?

Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	83,5	91,6	0,63	0,59	0,78	0,87	0,8	?	0,83	0,91	0,76	?
Backward Search	27	82,2	88,1	0,65	0,74	0,79	0,9	0,78	0,87	0,82	0,88	0,79	0,87

Experiment-103-Sampling [US]-All Variations

Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	9	66,9	40,8	0,66	0,6	0,63	0,87	0,67	0,9	0,66	0,4	0,66	0,5
	Backward Search		14	66,8	46,9	0,67	0,67	0,63	0,88	0,66	0,89	0,66	0,47	0,66
	Random Forest													
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
10-Fold			Validation	10-Fold	Validation									
Forward Search	6	66,6	48,9	0,68	0,7	0,67	0,89	0,67	0,9	0,66	0,49	0,66	0,58	
		Backward Search	22	63,2	51,1	0,68	0,72	0,67	0,9	0,63	0,9	0,63	0,51	0,63
Variation 2	Decision Tree (J48)													
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
			10-Fold	Validation										
	Forward Search	5	73,7	95,5	0,72	0,78	0,67	0,92	0,74	0,95	0,73	0,95	0,73	0,95
	Backward Search		15	73,2	95	0,73	0,78	0,69	0,92	0,73	0,94	0,73	0,95	0,73
	Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	10	72,8	95,5	0,71	0,78	0,66	0,92	0,73	0,95	0,72	0,95	0,72	0,95	
Backward Search		25	72,5	91,6	0,8	0,9	0,8	0,96	0,72	0,93	0,72	0,91	0,72	0,92
Variation 3	Decision Tree (J48)													
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
			10-Fold	Validation										
	Forward Search	4	78,1	78,2	0,75	0,82	0,73	0,93	0,78	0,91	0,78	0,78	0,78	0,82
	Backward Search		6	76,5	83,7	0,76	0,84	0,72	0,93	0,77	0,92	0,76	0,83	0,76
	Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	10	80,8	79,3	0,82	0,82	0,8	0,94	0,81	0,9	0,8	0,79	0,8	0,83	
Backward Search		35	78,1	80,8	0,85	0,9	0,86	0,96	0,78	0,92	0,78	0,8	0,78	0,84

Experiment-2IP-Sampling [None]-VARI – Decision tree (J48)

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	79,2	90,9	0,49	0,5	0,66	0,83	?	?	0,79	0,9	?	?
Backward Search	None	79,2	90,9	0,49	0,5	0,66	0,83	?	?	0,79	0,9	?	?

Experiment-2IP-Sampling [US]-All

Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	5	70,7	78,3	0,69	0,8	0,66	0,9	0,71	0,9	0,7	0,78	0,7	0,82
	Backward Search	12	70,2	78,3	0,69	0,87	0,69	0,92	0,71	0,91	0,7	0,78	0,69	0,82
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	7	74,7	76,2	0,72	0,66	0,68	0,87	0,75	0,86	0,74	0,76	0,74	0,8
	Backward Search	20	73,2	82,5	0,79	0,81	0,78	0,9	0,73	0,91	0,73	0,82	0,73	0,85
Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 2	Forward Search	3	71,8	90,9	0,65	0,5	0,61	0,83	0,72	?	0,71	0,9	0,71	?
	Backward Search	12	66,4	92,3	0,65	0,57	0,63	0,86	0,68	0,91	0,66	0,92	0,65	0,9
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 2	Forward Search	7	78,1	82,5	0,8	0,69	0,78	0,86	0,78	0,89	0,78	0,82	0,78	0,85
	Backward Search	13	78,9	79	0,82	0,76	0,8	0,9	0,78	0,88	0,78	0,79	0,78	0,82
Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 3	Forward Search	2	90,9	60,1	0,86	0,59	0,83	0,85	0,9	0,85	0,9	0,6	0,9	0,68
	Backward Search	2	90,9	59,4	0,85	0,6	0,83	0,85	0,9	0,86	0,9	0,59	0,9	0,68
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 3	Forward Search	3	90,9	58,7	0,88	0,61	0,86	0,85	0,9	0,86	0,9	0,58	0,9	0,67
	Backward Search	30	72,7	48,2	0,79	0,66	0,8	0,87	0,73	0,89	0,72	0,48	0,72	0,57

Experiment-2IP-Sampling [OS]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	97,3	86,7	0,99	0,4	0,99	0,81	0,97	0,82	0,97	0,86	0,97	0,84
Backward Search	6	99,2	90,2	0,99	0,49	0,98	0,83	0,99	0,82	0,99	0,9	0,99	0,86
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	99,2	85,5	1	0,57	1	0,85	0,99	0,83	0,99	0,82	0,99	0,82
Backward Search	20	100	81,1	1	0,62	1	0,86	1	0,83	1	0,81	1	0,82

Experiment-2IP-Sampling [SMOTE]-VAR2

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	89,1	68,5	0,89	0,61	0,86	0,86	0,89	0,84	0,89	0,68	0,89	0,74
Backward Search	17	85,9	79	0,88	0,74	0,85	0,89	0,86	0,89	0,85	0,79	0,85	0,82
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	88,8	48,9	0,91	0,64	0,88	0,87	0,88	0,87	0,88	0,49	0,88	0,58
Backward Search	25	90,3	87,4	0,96	0,82	0,96	0,91	0,9	0,87	0,9	0,87	0,9	0,87

Experiment-2IP-Sampling [SMOTE]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	95,4	82,5	0,94	0,5	0,91	0,83	0,95	0,83	0,95	0,82	0,95	0,82
Backward Search	12	94,3	72	0,97	0,59	0,95	0,85	0,94	0,87	0,94	0,72	0,94	0,77
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	96,5	81,8	0,98	0,63	0,98	0,87	0,96	0,84	0,96	0,81	0,96	0,82
Backward Search	35	98,1	86,7	0,99	0,63	0,99	0,86	0,98	0,82	0,98	0,86	0,98	0,84

Experiment-211-Sampling [None]-All – Decision tree (J48)

Decision Tree (J48)													
Variation 1	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
	Forward Search	None	95,2	94,4	0,48	0,5	0,9	0,89	?	?	0,95	0,94	?
Backward Search	None	95,2	94,4	0,48	0,5	0,9	0,89	?	?	0,95	0,94	?	?
Decision Tree (J48)													
Variation 2	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
	Forward Search	None	95	94,4	0,47	0,5	0,9	0,89	?	?	0,95	0,94	?
Backward Search	None	95	94,4	0,47	0,5	0,9	0,89	?	?	0,95	0,94	?	?
Decision Tree (J48)													
Variation 3	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
	Forward Search	None	95,3	94,4	0,47	0,5	0,92	0,89	0,92	?	0,95	0,94	0,93
Backward Search	None	95,3	94,4	0,47	0,5	0,92	0,89	0,92	?	0,95	0,94	0,93	?

Experiment-211-Sampling [US]-All

Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	3	69,3	60,2	0,65	0,63	0,72	0,91	0,72	0,92	0,69	0,6	0,68	0,7
	Backward Search	14	68,6	56,2	0,64	0,65	0,65	0,92	0,71	0,93	0,68	0,56	0,67	0,67
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	7	72,6	59	0,72	0,67	0,68	0,92	0,72	0,92	0,72	0,59	0,72	0,69
	Backward Search	25	72	56,5	0,74	0,51	0,73	0,9	0,72	0,9	0,72	0,56	0,72	0,68
Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 2	Forward Search	1	73,6	78,2	0,63	0,75	0,62	0,92	0,73	0,93	0,73	0,78	0,73	0,83
	Backward Search	3	76,3	78,2	0,69	0,71	0,66	0,92	0,76	0,93	0,76	0,78	0,76	0,83
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 2	Forward Search	3	75	74,9	0,66	0,75	0,64	0,92	0,75	0,92	0,75	0,74	0,75	0,81
	Backward Search	25	67,1	74	0,69	0,67	0,69	0,92	0,67	0,91	0,67	0,74	0,67	0,8
Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 3	Forward Search	1	88,8	42,8	0,83	0,64	0,85	0,91	0,9	0,93	0,88	0,42	0,88	0,54
	Backward Search	1	88,8	42,8	0,83	0,64	0,85	0,91	0,9	0,93	0,88	0,42	0,88	0,54
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 3	Forward Search	2	94,4	37	0,91	0,71	0,91	0,92	0,95	0,93	0,94	0,37	0,94	0,48
	Backward Search	2	94,4	37	0,91	0,71	0,91	0,92	0,95	0,93	0,94	0,37	0,94	0,48

Experiment-211-Sampling [OS]-VARI

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	96,6	87,4	0,99	0,36	0,99	0,88	0,96	0,88	0,96	0,87	0,96	0,88
Backward Search	21	97,6	83,7	0,98	0,57	0,97	0,9	0,97	0,9	0,97	0,83	0,97	0,86
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	98	86,5	0,99	0,4	0,99	0,88	0,98	0,88	0,98	0,86	0,98	0,87
Backward Search	25	99,5	90,2	1	0,51	1	0,89	0,99	0,89	0,99	0,9	0,99	0,89

Experiment-211-Sampling [OS]-VAR2

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	98,7	88,3	0,99	0,47	0,99	0,89	0,98	0,89	0,98	0,88	0,98	0,88
Backward Search	14	98	88,3	0,99	0,54	0,98	0,9	0,98	0,9	0,98	0,88	0,98	0,89
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	15	99,9	93,2	1	0,48	1	0,88	0,99	0,89	0,99	0,93	0,99	0,91
Backward Search	24	99,7	92,6	1	0,54	1	0,9	0,99	0,9	0,99	0,92	0,99	0,91

Experiment-211-Sampling [OS]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	100	82,2	1	0,56	1	0,9	1	0,9	1	0,82	1	0,86
Backward Search	6	99,5	93,5	0,99	0,49	0,99	0,89	0,99	0,89	0,99	0,93	0,99	0,91
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	6	100	91,4	1	0,68	1	0,92	1	0,9	1	0,91	1	0,9
Backward Search	5	100	78,2	1	0,75	1	0,93	1	0,91	1	0,78	1	0,83

Experiment-212-Sampling [US]-VARI

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	8	76,3	67,3	0,73	0,66	0,69	0,93	0,76	0,93	0,76	0,67	0,76	0,77
Backward Search	11	75,4	71	0,74	0,67	0,69	0,93	0,75	0,93	0,75	0,71	0,75	0,79
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	9	75,6	79,7	0,78	0,75	0,75	0,94	0,75	0,93	0,75	0,79	0,75	0,85
Backward Search	9	75,9	62,2	0,79	0,61	0,75	0,93	0,76	0,91	0,76	0,62	0,76	0,73

Experiment-212-Sampling [US]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	95	77,7	0,95	0,75	0,93	0,94	0,95	0,94	0,95	0,77	0,95	0,84
Backward Search	1	75	76,4	0,73	0,39	0,75	0,9	0,83	0,91	0,75	0,76	0,73	0,83
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	85	48,8	0,76	0,42	0,76	0,91	0,88	0,9	0,85	0,48	0,84	0,62
Backward Search	19	80	64,9	0,75	0,7	0,77	0,94	0,85	0,93	0,8	0,65	0,79	0,75

Experiment-212-Sampling [OS]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	99,3	91,58	1	0,66	1	0,93	0,99	0,93	0,99	0,91	0,99	0,92
Backward Search	8	99,3	91,9	0,99	0,74	0,99	0,94	0,99	0,94	0,99	0,91	0,99	0,93
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	100	92,2	1	0,74	1	0,94	1	0,93	1	0,92	1	0,92
Backward Search	12	99,7	95,9	1	0,81	1	0,96	0,99	0,96	0,99	0,96	0,99	0,96

Experiment-212-Sampling [SMOTE]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	98,3	94,9	0,98	0,73	0,98	0,94	0,98	0,95	0,98	0,94	0,98	0,95
Backward Search	5	97,4	92,2	0,97	0,6	0,96	0,92	0,97	0,91	0,97	0,92	0,97	0,91
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	98	54,5	0,99	0,55	0,99	0,91	0,98	0,92	0,98	0,54	0,98	0,66
Backward Search	23	98,7	94,6	0,99	0,82	0,99	0,95	0,98	0,93	0,98	0,94	0,98	0,93

Experiment-3SA-Sampling [None]-VAR1 – Decision tree (J48)

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	91,06	98,2	0,49	0,5	0,83	0,96	?	?	0,91	0,98	?	?
Backward Search	None	91,06	98,2	0,49	0,5	0,83	0,96	?	?	0,91	0,98	?	?

Experiment-3SA-Sampling [US]-All

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	73,4	63,4	0,73	0,56	0,72	0,96	0,75	0,97	0,73	0,63	0,72	0,76
	Backward Search	9	78,4	72,1	0,81	0,64	0,79	0,97	0,78	0,97	0,78	0,72	0,78
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	74	69,5	0,75	0,68	0,71	0,97	0,74	0,97	0,74	0,69	0,73	0,8
Backward Search	7	76,5	68,6	0,79	0,75	0,77	0,97	0,76	0,97	0,76	0,68	0,76	0,79
Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	73,4	63,4	0,73	0,56	0,72	0,96	0,75	0,97	0,73	0,63	0,72	0,76
Backward Search	9	78,4	72,1	0,81	0,64	0,79	0,97	0,78	0,97	0,78	0,72	0,78	0,82
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	74	69,5	0,75	0,68	0,71	0,97	0,74	0,97	0,74	0,69	0,73	0,8
Backward Search	7	76,5	68,6	0,79	0,75	0,77	0,97	0,76	0,97	0,76	0,68	0,76	0,79
Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	73,4	63,4	0,73	0,56	0,72	0,96	0,75	0,97	0,73	0,63	0,72	0,76
Backward Search	9	78,4	72,1	0,81	0,64	0,79	0,97	0,78	0,97	0,78	0,72	0,78	0,82
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	7	74	69,5	0,75	0,68	0,71	0,97	0,74	0,97	0,74	0,69	0,73	0,8
Backward Search	7	76,5	68,6	0,79	0,75	0,77	0,97	0,76	0,97	0,76	0,68	0,76	0,79

Experiment-3SA-Sampling [OS]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	100	97,3	1	0,49	1	0,96	1	0,96	1	0,97	1	0,97
Backward Search	2	100	98,2	1	0,5	1	0,96	1	?	1	0,98	1	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	100	97,3	1	0,47	1	0,96	1	0,96	1	0,97	1	0,97
Backward Search	2	100	98,2	1	0,5	1	0,96	1	?	1	0,98	1	?

Experiment-3SA-Sampling [SMOTE]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	99,4	97,3	0,99	0,74	0,99	0,97	0,99	0,97	0,99	0,97	0,99	0,97
Backward Search	1	99,4	98,2	0,99	0,5	0,98	0,96	0,99	?	0,99	0,98	0,99	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	100	98,2	1	0,5	1	0,96	1	?	1	0,98	1	?
Backward Search	2	100	98,2	1	0,49	1	0,96	1	?	1	0,98	1	?

Experiment-3AS-Sampling [None]-VAR1

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	98,1	96,9	0,42	0,5	0,96	0,94	?	?	0,98	0,96	?	?
Backward Search	None	98,1	96,9	0,42	0,5	0,96	0,94	?	?	0,98	0,96	?	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	98,5	94,7	0,74	0,46	0,97	0,94	0,98	0,93	0,98	0,94	0,98	0,94
Backward Search	13	98,5	96	0,82	0,73	0,98	0,96	0,98	0,93	0,98	0,96	0,98	0,95

Experiment-3AS-Sampling [None]-VAR2

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	98,1	97,3	0,58	0,74	0,96	0,96	0,98	0,97	0,98	0,97	0,97	0,96
Backward Search	1	98,1	97,3	0,58	0,74	0,96	0,96	0,98	0,97	0,98	0,97	0,97	0,96
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	98,1	96,9	0,79	0,69	0,97	0,95	0,97	0,95	0,98	0,96	0,98	0,96
Backward Search	30	98,1	97,3	0,71	0,63	0,96	0,95	0,98	0,97	0,98	0,97	0,97	0,96

Experiment-3AS-Sampling [None]-VAR3

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	97,2	96,9	0,29	0,5	0,95	0,94	0,95	?	0,97	0,96	0,96	?
Backward Search	None	97,2	96,9	0,29	0,5	0,95	0,94	0,95	?	0,97	0,96	0,96	?
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	99,4	93,3	0,97	0,71	0,99	0,95	0,99	0,94	0,99	0,93	0,99	0,94
Backward Search	5	97,8	90,7	0,8	0,46	0,97	0,94	?	0,93	0,97	0,9	?	0,92

Experiment-3AS-Sampling [US]-VAR1 and VAR2 (VAR3 – Too few instances)

		Decision Tree (J48)												
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
			10-Fold	Validation										
Variation 1	Forward Search	1	84,3	72,6	0,73	0,73	0,72	0,95	0,84	0,96	0,84	0,72	0,84	0,81
	Backward Search	1	84,3	72,6	0,73	0,73	0,72	0,95	0,84	0,96	0,84	0,72	0,84	0,81
	Random Forest													
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	3	84,3	78,4	0,76	0,73	0,71	0,96	0,84	0,94	0,84	0,78	0,84	0,85	
	Backward Search	22	90,6	67,4	0,89	0,73	0,89	0,96	0,9	0,95	0,9	0,67	0,9	0,77
Variation 2	Decision Tree (J48)													
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
			10-Fold	Validation										
	Forward Search	1	76,9	72,6	0,72	0,73	0,71	0,95	0,76	0,96	0,76	0,72	0,76	0,81
	Backward Search	5	73	78,8	0,73	0,82	0,71	0,96	0,74	0,96	0,73	0,78	0,72	0,85
	Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Forward Search	3	84,6	67,4	0,82	0,64	0,82	0,94	0,88	0,95	0,84	0,67	0,84	0,77	
Backward Search	31	65,3	55	0,79	0,89	0,82	0,97	0,65	0,97	0,65	0,55	0,65	0,68	

Experiment-3AS-Sampling [OS]-All

Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	5	96,8	87,6	0,99	0,48	0,99	0,93	0,97	0,93	0,96	0,87	0,96	0,9
	Backward Search	13	99,2	91,1	0,99	0,53	0,99	0,94	0,99	0,94	0,99	0,91	0,99	0,92
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	7	99,7	94,2	0,99	0,5	0,99	0,94	0,99	0,93	0,99	0,94	0,99	0,94
	Backward Search	24	100	96	1	0,61	1	0,94	1	0,93	1	0,96	1	0,95
Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 2	Forward Search	2	99,6	96,9	0,99	0,56	0,99	0,94	0,99	0,95	0,99	0,96	0,99	0,96
	Backward Search	8	99,5	89,4	0,99	0,52	0,99	0,94	0,99	0,94	0,99	0,89	0,99	0,91
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 2	Forward Search	3	100	97,3	1	0,56	1	0,94	1	0,97	1	0,97	1	0,96
	Backward Search	7	100	96,4	1	0,78	1	0,95	1	0,93	1	0,96	1	0,95
Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 3	Forward Search	3	100	92	1	0,54	1	0,94	1	0,94	1	0,92	1	0,93
	Backward Search	7	100	94,7	1	0,55	1	0,94	1	0,94	1	0,94	1	0,94
Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 3	Forward Search	3	100	92,5	1	0,58	1	0,94	1	0,94	1	0,92	1	0,93
	Backward Search	5	100	94,2	1	0,67	1	0,95	1	0,93	1	0,94	1	0,94

Experiment-3AS-Sampling [SMOTE]-All

Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	1	99,3	94,7	0,99	0,41	0,99	0,93	0,99	0,93	0,99	0,94	0,99	0,94
	Backward Search		23	99,2	94,2	0,99	0,68	0,99	0,95	0,99	0,93	0,99	0,94	0,99
	Random Forest													
	Forward Search	2	99,3	94,7	0,99	0,48	0,99	0,94	0,99	0,93	0,99	0,94	0,99	0,94
Backward Search	23		99,2	94,2	0,99	0,68	0,99	0,95	0,99	0,93	0,99	0,94	0,99	0,94
Variation 2	Decision Tree (J48)													
	Forward Search	1	98,7	94,2	0,98	0,44	0,98	0,93	0,98	0,93	0,98	0,94	0,98	0,94
	Backward Search		8	98,5	95,5	0,98	0,5	0,97	0,94	0,98	0,94	0,98	0,95	0,98
	Random Forest													
Forward Search	2	98,7	94,2	0,99	0,49	0,98	0,94	0,98	0,93	0,98	0,94	0,98	0,94	
Backward Search		23	99,1	95,5	0,99	0,65	0,99	0,95	0,99	0,93	0,99	0,95	0,99	0,94
Variation 3	Decision Tree (J48)													
	Forward Search	1	99,1	95,5	0,99	0,45	0,99	0,93	0,99	0,93	0,99	0,95	0,99	0,94
	Backward Search		4	99,1	94,7	0,99	0,48	0,99	0,94	0,99	0,93	0,99	0,94	0,99
	Random Forest													
Forward Search	1	99,1	95,5	0,99	0,49	0,99	0,94	0,99	0,93	0,99	0,95	0,99	0,94	
Backward Search		23	100	96,9	1	0,71	1	0,96	1	?	1	0,96	1	?

Experiment-3SI-Sampling [None]-VARI – Decision tree (J48)

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	95,1	96,8	0,48	0,5	0,9	0,94	?	?	0,95	0,96	?	?
Backward Search	None	95,1	96,8	0,48	0,5	0,9	0,94	?	?	0,95	0,96	?	?

Experiment-3SI-Sampling [None]-VAR3 – Decision tree (J48)

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	93,4	96,8	0,47	0,5	0,89	0,94	0,89	?	0,93	0,96	0,91	?
Backward Search	None	93,4	96,8	0,47	0,5	0,89	0,94	0,89	?	0,93	0,96	0,91	?

Experiment-3SI-Sampling [US]-All

Decision Tree (J48)														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	
Variation 1	Forward Search	3	74,3	78,6	0,71	0,45	0,68	0,93	0,75	0,94	0,74	0,78	0,74	0,85
	Backward Search	1	73,1	72	0,62	0,54	0,6	0,93	0,75	0,95	0,72	0,72	0,72	0,81
	Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
Forward Search	5	75,6	71,1	0,68	0,52	0,67	0,94	0,78	0,95	0,75	0,71	0,75	0,8	
Backward Search	26	69,5	66,6	0,77	0,49	0,78	0,93	0,69	0,94	0,69	0,66	0,69	0,77	
Variation 2	Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	Forward Search	1	76,4	72	0,67	0,54	0,64	0,93	0,78	0,95	0,76	0,72	0,76	0,81
	Backward Search	1	76,4	72	0,67	0,54	0,64	0,93	0,78	0,95	0,76	0,72	0,76	0,81
Variation 3	Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	Forward Search	3	77,9	48,4	0,76	0,58	0,72	0,94	0,8	0,94	0,77	0,48	0,77	0,62
	Backward Search	32	73,5	73,3	0,8	0,73	0,79	0,95	0,73	0,95	0,73	0,73	0,73	0,82
Variation 3	Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	Forward Search	1	80	72	0,68	0,57	0,68	0,94	0,81	0,94	0,8	0,72	0,79	0,81
	Backward Search	1	80	72	0,68	0,57	0,68	0,94	0,81	0,94	0,8	0,72	0,79	0,81
Variation 3	Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
	Forward Search	5	90	82,2	0,83	0,68	0,81	0,95	0,91	0,95	0,9	0,82	0,89	0,87
	Backward Search	27	75	84	0,83	0,64	0,83	0,95	0,77	0,94	0,75	0,84	0,74	0,88

Experiment-3SI-Sampling [OS]-All

Decision Tree (J48)															
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure			
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation		
Variation 1	Forward Search	7	97,4	88,4	0,98	0,41	0,98	0,93	0,97	0,94	0,97	0,88	0,97	0,91	
	Backward Search	21	97,1	86,2	0,98	0,51	0,97	0,94	0,97	0,94	0,97	0,86	0,97	0,89	
	Random Forest														
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
10-Fold			Validation	10-Fold	Validation										
Variation 1	Forward Search	15	98,8	89,7	0,99	0,48	0,99	0,94	0,98	0,94	0,98	0,89	0,98	0,91	
	Backward Search	23	99,5	93,3	0,99	0,63	0,99	0,95	0,99	0,94	0,99	0,93	0,99	0,93	
Variation 2	Decision Tree (J48)														
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
			10-Fold	Validation											
	Variation 2	Forward Search	4	98,7	91,5	0,93	0,59	0,99	0,94	0,98	0,95	0,98	0,91	0,98	0,93
		Backward Search	10	99	96,8	0,99	0,85	0,98	0,97	0,99	0,97	0,99	0,96	0,99	0,97
	Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure			
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation		
Variation 2	Forward Search	9	99,5	96,8	1	0,67	1	0,96	0,99	0,96	0,99	0,96	0,99	0,96	
	Backward Search	15	99,7	96	1	0,83	1	0,97	0,99	0,95	0,99	0,96	0,99	0,95	
Variation 3	Decision Tree (J48)														
		Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure		
			10-Fold	Validation											
	Variation 3	Forward Search	3	99,4	92	0,99	0,74	0,99	0,95	0,99	0,96	0,99	0,92	0,99	0,93
		Backward Search	6	100	96,4	1	0,77	1	0,96	1	0,96	1	0,96	1	0,96
	Random Forest														
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure			
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation		
Variation 3	Forward Search	3	100	90,2	1	0,76	1	0,96	1	0,95	1	0,9	1	0,92	
	Backward Search	6	100	96,4	1	0,73	1	0,96	1	0,96	1	0,96	1	0,96	

Experiment-3ND-Sampling [None]-VARI – Decision tree (J48)

Decision Tree (J48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	None	81,5	97,8	0,49	0,5	0,69	0,95	?	?	0,81	0,97	?	?
Backward Search	None	81,5	97,8	0,49	0,5	0,69	0,95	?	?	0,81	0,97	?	?

Experiment-3ND-Sampling [US]-VAR1

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	2	71	40,7	0,71	0,69	0,68	0,96	0,72	0,96	0,71	0,4	0,7	0,55
Backward Search	16	68,8	48	0,71	0,54	0,67	0,95	0,69	0,96	0,68	0,48	0,68	0,63
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	4	70,7	64,8	0,72	0,66	0,71	0,96	0,7	0,96	0,7	0,64	0,7	0,76
Backward Search	17	71,6	56,6	0,75	0,67	0,73	0,97	0,71	0,96	0,71	0,56	0,71	0,56

Experiment-3ND-Sampling [OS]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	100	97,4	1	0,69	1	0,96	1	0,97	1	0,97	1	0,97
Backward Search	4	100	98,7	1	0,7	1	0,97	1	0,98	1	0,98	1	0,98
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	1	100	97,4	1	0,78	1	0,97	1	0,97	1	0,97	1	0,97
Backward Search	4	99,7	99,1	1	0,99	1	0,99	0,99	0,99	0,99	0,99	0,99	0,99

Experiment-3ND-Sampling [SMOTE]-VAR3

Decision Tree (I48)													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	98,7	98,2	0,98	0,69	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98
Backward Search	3	99	95,2	0,98	0,87	0,98	0,97	0,99	0,97	0,99	0,95	0,99	0,96
Random Forest													
	Attribute Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
		10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
Forward Search	3	99	98,7	0,99	0,89	0,99	0,99	0,99	0,98	0,99	0,98	0,99	0,98
Backward Search	30	99,5	98,2	1	0,98	1	0,99	0,99	0,98	0,99	0,98	0,99	0,97

Appendix B – Experimental results not listed in Chapter 6

Experiment-2IP-FS [Genetic] – VAR1

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	Genetic	DT	10	74,75	68,5	0,73	0,62	0,69	0,86	0,74	0,86	0,74	0,68	0,74	0,75
	Genetic	RF	17	75,24	73,42	0,77	0,77	0,76	0,89	0,75	0,89	0,75	0,73	0,75	0,78

Experiment-2IP-Algorithm [OF] – VAR1

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	All	OF	29	63,86	69,2	0,66	0,77	0,64	0,9	0,64	0,9	0,63	0,69	0,63	0,75
	RF Fwd Search	OF	8	72,27	75,52	0,71	0,66	0,67	0,87	0,72	0,85	0,72	0,75	0,72	0,79
	RF Bkwd Search	OF	21	74,25	81,1	0,8	0,81	0,79	0,9	0,74	0,91	0,74	0,81	0,74	0,84
	Genetic	OF	17	74,75	76,2	0,76	0,77	0,75	0,89	0,74	0,89	0,74	0,76	0,74	0,8

Experiment-2IP-FS [Genetic] – VAR2

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	Genetic	DT	8	71,87	88,8	0,74	0,84	0,7	0,93	0,72	0,91	0,71	0,88	0,71	0,89
	Genetic	RF	15	72,65	80,4	0,76	0,7	0,74	0,88	0,72	0,89	0,72	0,8	0,72	0,83

Experiment-2IP-Algorithm [OF] – VAR2

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	All	OF	34	70,3	76,2	0,74	0,81	0,73	0,91	0,7	0,89	0,7	0,76	0,7	0,8
	RF Fwd Search	OF	8	78,1	83,2	0,81	0,72	0,79	0,89	0,78	0,89	0,78	0,83	0,78	0,85
	RF Bkwd Search	OF	14	77,3	78,3	0,81	0,76	0,8	0,9	0,77	0,88	0,77	0,78	0,77	0,82
	Genetic	OF	15	71,8	81,8	0,76	0,71	0,74	0,88	0,72	0,89	0,71	0,81	0,71	0,84

Experiment-2IP-FS [Genetic] – VAR3

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	Genetic	DT	7	90,9	59,4	0,85	0,6	0,83	0,85	0,9	0,86	0,9	0,59	0,9	0,68
	Genetic	RF	15	81,8	44,7	0,81	0,63	0,8	0,85	0,82	0,88	0,81	0,44	0,81	0,54

Experiment-2IP-Algorithm [OF] – VAR3

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	All	OF	39	72,7	50,3	0,75	0,64	0,74	0,87	0,73	0,86	0,72	0,5	0,72	0,6
	RF Fwd Search	OF	4	90,9	58,7	0,89	0,6	0,87	0,85	0,9	0,86	0,9	0,58	0,9	0,67
	RF Bkwd Search	OF	31	68,1	46,1	0,75	0,65	0,77	0,87	0,68	0,88	0,68	0,46	0,68	0,55
	Genetic	OF	15	81,8	48,2	0,82	0,62	0,81	0,85	0,82	0,89	0,81	0,48	0,81	0,57

Experiment-3AS-FS [Genetic] – VAR1

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	Genetic	DT	7	84,3	72,6	0,73	0,73	0,72	0,95	0,84	0,96	0,84	0,72	0,84	0,81
	Genetic	RF	32	87,5	65,1	0,9	0,7	0,89	0,95	0,88	0,96	0,87	0,76	0,87	0,76

Experiment-3AS-Algorithm [OF] – VAR1

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
None	All	OF	29	98,1	96,4	0,71	0,49	0,97	0,94	0,97	0,93	0,98	0,96	0,97	0,95
	RF Fwd Search	OF	4	98,5	94,7	0,68	0,48	0,97	0,93	0,98	0,93	0,98	0,94	0,98	0,94
	RF Bkwd Search	OF	14	98,4	96	0,77	0,48	0,97	0,93	0,98	0,93	0,98	0,96	0,98	0,95
	Genetic	OF	5	98,4	94,7	0,67	0,47	0,97	0,93	0,98	0,93	0,98	0,94	0,98	0,94
OS	All	OF	29	99,6	94,7	1	0,71	1	0,95	0,99	0,93	0,99	0,94	0,99	0,94
	RF Fwd Search	OF	8	99,7	94,2	0,99	0,63	0,99	0,95	0,99	0,93	0,99	0,94	0,99	0,94
	RF Bkwd Search	OF	25	99,8	96	1	0,62	1	0,95	0,99	0,93	0,99	0,96	0,99	0,95
	Genetic	OF	21	99,8	95,5	1	0,67	1	0,95	0,99	0,93	0,99	0,95	0,99	0,94
SMOTE	All	OF	29	99	95,1	0,99	0,69	0,99	0,95	0,99	0,93	0,99	0,95	0,99	0,94
	RF Fwd Search	OF	3	99,3	94,7	0,99	0,47	0,99	0,94	0,99	0,93	0,99	0,94	0,99	0,94
	RF Bkwd Search	OF	24	99,2	93,8	0,99	0,71	0,99	0,95	0,99	0,93	0,99	0,93	0,99	0,93
	Genetic	OF	20	99,2	95,5	0,99	0,7	0,99	0,95	0,99	0,93	0,99	0,95	0,99	0,94
US	All	OF	29	75	65,1	0,78	0,71	0,77	0,95	0,75	0,95	0,75	0,65	0,75	0,76
	RF Fwd Search	OF	4	84,3	78,4	0,77	0,71	0,73	0,95	0,84	0,94	0,84	0,78	0,84	0,85
	RF Bkwd Search	OF	23	84,3	69,6	0,88	0,71	0,88	0,96	0,84	0,95	0,84	0,69	0,84	0,79
	Genetic	OF	14	87,5	67,4	0,88	0,71	0,85	0,95	0,88	0,96	0,87	0,67	0,87	0,77

Experiment-3AS-FS [Genetic] – VAR2

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
US	Genetic	DT	3	76,9	78,8	0,66	0,82	0,67	0,96	0,84	0,96	0,76	0,78	0,75	0,85
	Genetic	RF	9	84,6	50,2	0,86	0,87	0,86	0,97	0,85	0,97	0,84	0,5	0,84	0,63

Experiment-3AS-Algorithm [OF] – VAR2

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
OS	All	OF	35	99,8	96,4	1	0,87	1	0,97	0,99	0,95	0,99	0,96	0,99	0,95
	RF Fwd Search	OF	4	100	97,3	1	0,57	1	0,94	1	0,97	1	0,97	1	0,96
	RF Bkwd Search	OF	8	100	96,4	1	0,8	1	0,96	1	0,93	1	0,96	1	0,95
	Genetic	OF	12	100	96,9	1	0,8	1	0,96	1	0,95	1	0,96	1	0,96
SMOTE	All	OF	35	98,7	95,5	0,99	0,75	0,99	0,96	0,98	0,93	0,98	0,95	0,98	0,94
	RF Fwd Search	OF	3	98,7	94,2	0,99	0,51	0,98	0,94	0,98	0,93	0,98	0,94	0,98	0,94
	RF Bkwd Search	OF	24	99,1	95,5	0,99	0,68	0,99	0,95	0,99	0,93	0,99	0,95	0,99	0,94
	Genetic	OF	16	98,8	96,4	0,99	0,84	0,99	0,97	0,98	0,96	0,98	0,96	0,98	0,96
US	All	OF	35	69,2	55,5	0,7	0,93	0,71	0,97	0,69	0,97	0,69	0,55	0,69	0,68
	RF Fwd Search	OF	4	84,6	67,4	0,84	0,66	0,82	0,94	0,88	0,95	0,84	0,67	0,84	0,77
	RF Bkwd Search	OF	32	69,2	55,9	0,79	0,89	0,82	0,97	0,69	0,97	0,69	0,55	0,69	0,68
	Genetic	OF	9	80,7	50,6	0,83	0,88	0,84	0,97	0,82	0,97	0,8	0,5	0,8	0,64

Experiment-3AS-FS [Genetic] – VAR3 – Not enough samples for US

Experiment-3AS-Algorithm [OF] – VAR3 – Not enough samples for US

Sampling	Feature Selection Search Type	Algorithm	Parameter Count	Accuracy %		ROC		PRC Area		Precision		Recall		F-Measure	
				10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation	10-Fold	Validation
OS	All	OF	38	100	94,7	1	0,79	1	0,96	1	0,93	1	0,94	1	0,94
	RF Fwd Search	OF	4	100	92	1	0,58	1	0,94	1	0,94	1	0,92	1	0,93
	RF Bkwd Search	OF	6	100	93,8	1	0,66	1	0,95	1	0,93	1	0,93	1	0,93
	Genetic	OF	6	100	81,4	1	0,77	1	0,96	1	0,95	1	0,81	1	0,87
SMOTE	All	OF	38	98,8	96	1	0,79	1	0,96	0,98	0,93	0,98	0,96	0,98	0,95
	RF Fwd Search	OF	19	100	95,1	0,99	0,49	0,99	0,94	0,99	0,93	0,99	0,95	0,99	0,94
	RF Bkwd Search	OF	24	100	96,4	1	0,68	1	0,95	1	0,93	1	0,96	1	0,95
	Genetic	OF	19	100	95,1	1	0,8	1	0,96	1	0,93	1	0,95	1	0,94

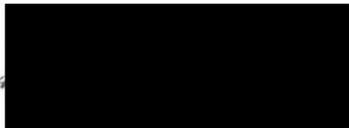
Appendix C – Letter from professional editor



5 September 2023

Editing Certificate

This letter confirms that the following doctoral thesis by Rushil Raghavjee was language edited: **Genetic Algorithm based prediction of students' course performance using learning analytics.**



Dr Karen Buckenham, PhD (KwaZulu-Natal), MA (Natal), BSc (Toronto), TESL (Toronto).

kuckenham@mweb.co.za

Appendix D – Ethical clearance letter



09 February 2021

Mr Rushil Raghavjee (201295456)
School Of Man Info Tech & Gov
Pietermaritzburg Campus

Dear Mr Raghavjee,

Protocol reference number: HSSREC/00002342/2021

Project title: Using learning analytics to monitor academic performance of students at UKZN

Degree: PhD

Approval Notification – Expedited Application

This letter serves to notify you that your application received on 25 January 2021 in connection with the above, was reviewed by the Humanities and Social Sciences Research Ethics Committee (HSSREC) and the protocol has been granted **FULL APPROVAL**.

Any alteration/s to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form, Title of the Project, Location of the Study, Research Approach and Methods must be reviewed and approved through the amendment/modification prior to its implementation. In case you have further queries, please quote the above reference number. **PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years.**

This approval is valid until 09 February 2022.

To ensure uninterrupted approval of this study beyond the approval expiry date, a progress report must be submitted to the Research Office on the appropriate form 2 - 3 months before the expiry date. A close-out report to be submitted when study is finished.

All research conducted during the COVID-19 period must adhere to the national and UKZN guidelines.

HSSREC is registered with the South African National Research Ethics Council (REC-040414-040).

Yours sincerely,

A black rectangular box redacting the signature of Professor Dipane Hlalele.

Professor Dipane Hlalele (Chair)

14 April 2023

Rushil Raghavjee (201295456)
School Of Man Info Tech & Gov
Pietermaritzburg Campus

Dear R Raghavjee,

Protocol reference number: HSSREC/00002342/2021

Project title: Using learning analytics to monitor academic performance of students at UKZN

Amended title: Genetic algorithm based prediction of students' course performance using learning analytics

Degree: PhD

Approval Notification – Amendment Application

This letter serves to notify you that your application and request for an amendment received on 12 April 2023 has now been approved as follows:

- Change in title

Any alterations to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form; Title of the Project, Location of the Study must be reviewed and approved through an amendment /modification prior to its implementation. In case you have further queries, please quote the above reference number.

PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years.

HSSREC is registered with the South African National Health Research Ethics Council (REC-040414-040).

Best wishes for the successful completion of your research protocol.

Yours faithfully



.....
Professor Dipane Hlalele (Chair)

/dd

Humanities & Social Sciences Research Ethics Committee
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Tel: +27 31 260 8350 / 4557 / 3587

Website: <http://research.ukzn.ac.za/Research-Ethics/>

Founding Campuses:  Edgewood  Howard College  Medical School  Pietermaritzburg  Westville

INSPIRING GREATNESS