Information and Knowledge Discovery from Undergraduate STEM datasets: Educational Data Mining at a South African University

Ndikhwine Netshitungulu (464332)
Supervisor: Mr. Mike Mchunu

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science.

October 2024

Declaration

I, Ndikhwine Netshitungulu,declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before to any other University, either for degree or examination purposes.

Signed:

Date: 23-10-2024

"Wouldn't it be great if there was an app that could help you weigh up the options and make a decision on the best possible outcome for you?"

Musa Kalenga

Abstract

South Africa is experiencing an acute shortage of highly trained and skilled professionals with much-needed qualifications in the disciplines of Science, Technology, Engineering and Mathematics (STEM). There is a need to accelerate the training and development of professionals in these critical, scarce-skill areas. Our universities and other institutions play an important role in supporting and facilitating the development of these much-needed professionals. The primary focus of this data mining based study was to examine a number of factors related to first time first year BSc. students enrolled in critical-skill STEM degree programmes at the University of the Witwatersrand (also known as Wits). These factors included ethnicity, gender, funding status, English language proficiency, prior computer programming experience, mathematical ability and student profiling. Wits and other Higher Education Institutions (HEIs) collect and accumulate large amounts of data about their students. Using different data mining techniques this data can be processed, analysed and leveraged to obtain meaningful and useful information about these students. This study was motivated by the need to obtain information from the longitudinal data (2015-2019) collected about first time first year students enrolled in STEM degree disciplines at Wits University. In order to conduct the study in a more structured manner, we followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which is increasingly being adopted as a project implementation standard by data mining practitioners working in different disciplines, including education. We applied different statistical and data mining techniques against our various datasets. These included Analysis of Variance (ANOVA), classification, feature selection, prediction, clustering and association rule mining techniques. Several important results emerged from our study. The result from the two-way ANOVA experiment showed the interaction between gender and ethnicity not to be significant. However, the ethnicity factor had a statistically significant effect on the APS. We also found the interaction between ethnicity and funding status to be significant. Individually, these factors were statistically significantly related to the APS. Regarding English language proficiency, the difference between English Home Language (ENA) students and English First Additional Language (ENB) students was significant. Amongst the NSC Grade 12 subjects, mathematics had the most significant relationship with the Admission Point Score (APS). In two first semester Computer Science I courses, students with prior computer programming experience significantly outperformed their peers, who lacked computer programming experience. We also found that collectively, the minimum admission requirement subjects for admission to the first year STEM disciplines were suitable predictors of the overall, first year outcome. However, the result we obtained using these subjects as predictors of degree attainment was not encouraging.

This wide ranging study has shown that data mining techniques can be used effectively in educational settings to leverage the data universities collect on their students, about whom insights and information can be obtained, information that can be used for well-informed decision making. Beyond our specific educational context, there is also much

to learn from this study and its findings by other, higher educational institutions with STEM students in a situation similar to ours.

Keywords:
Educational Data Mining, Data Mining Techniques, STEM, CRISP-DM, Feature Selection, Imbalanced Class Distribution, Performance Evaluation Metrics, Confusion Matrix, Receiver Operating Characteristics, Student Characteristics, English Language Proficiency.

List of Abbreviations

Abbreviation	**Meaning**
AISU	Academic Information Systems Unit
ANOVA	Analysis of Variance
APS	Admission Point Score
ARM	Association Rule Mining
ARRF	Attribute Relation File Format
AUC	Area Under the ROC Curve
BCO	Basic Computer Organization
C4.5	C4.5 Decision Tree Learner
CRISP-DM	Cross Industry Standard Process for Data Mining
CSI	Computer Science I
DM	Data Mining
EDM	Educational Data Mining
ENA	NSC English Home Language Subject
ENB	NSC English First Additional Language Subject
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GPA	Grade Point Average
HEI	Higher Education Institution
IAP	Introduction to Algorithms and Programming
IBk	Instance Based Learner
J48	WEKA-based Decision Tree Algorithm
JRIP	Implements a propositional rule learner, RIPPER
KDD	Knowledge Discovery in Databases
kNN	K-nearest Neighbours Algorithm
LO	Life Orientation
MAT	NSC Mathematics Subject
MLP	Multi-layer Perceptron Classifier
MLR	Multiple Linear Regression
NB	Naive Bayes Classifier
NSC	National Senior Certificate
NSFAS	National Student Financial Aid Scheme
OS	Other Subjects
PSC	NSC Physical Sciences Subject
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROC	Receiver Operating Characteristics
SAS	Statistical Analysis System
STEM	Science, Technology, Mathematics and Engineering
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
T-Test	Statistical Significance Test
WEKA	Waikato Environment for Knowledge Analysis

Acknowledgements

I would like to thank my supervisor, Mike Mchunu, for proposing this interesting topic for my research and for ensuring through his time, effort and guidance, my success in completing my studies. I would also like to thank my employer, Standard Bank, for granting me the bursary which afforded me the opportunity to further my studies. I also received tremendous support from Mr Andrew Mthethwa, my then senior and supervisor at Standard Bank, for supporting my quest for further education and for making it possible for me to work on a flexible schedule, in order to for me to be able to take time off and consult with research supervisor, whenever necessary. I would also like to thank Ms. Ashleigh Davids at Wits, who helped me to obtain from the Wits Business Intelligence Services all the data I needed to conduct my research. I would also like to extend my thanks and gratitude to Ms. Casey Sparkes, Manager of the DST-NRF Centre of Excellence in Mathematics & Statistical Sciences (CoE-MaSS), for allowing both myself and my supervisor to use one of their offices during the final preparation of my dissertation document. Lastly, my thanks go to my family, for their upbringing and emphasis on education and its value, and their encouragement during my current studies. I am also indebted to and would like to thank everyone, including my friends, who provided me with intellectual and emotional support during the course of my studies.

Contents

Declaration i

Abstract iii

List of Abbreviations v

Acknowledgements vi

1 Introduction **1**

1.1 Background Of The Problem 1

1.2 Problem Statement 4

1.3 Significance of the Study 7

1.4 Research Question 10

1.5 Research Approach 11

1.6 Research Highlights and Contribution 12

1.6.1 Prior Programming Experience and First-Semester Performance . 13

1.6.2 Research in Educational Data Mining (EDM) 14

1.6.3 The Research Methodology 14

1.6.4 The Role of Mathematics 15

1.6.5 Model Evaluation 15

1.7 Structure of the Dissertation 16

2 Background and Related Work **17**

2.1 Introduction 17

2.2 Background 18

2.2.1 Data Mining . . . 18
2.2.2 Applications of Data Mining . . . 19
2.3 Educational Data Mining . . . 21
2.4 Data Mining for the Research Problem . . . 22
2.5 Selected Data Mining and Statistical Techniques . . . 24
2.5.1 Classification . . . 24
2.5.2 Prediction . . . 32
2.5.3 Cluster Analysis . . . 33
2.5.4 Association Rule Mining (ARM) . . . 36
2.5.5 Statistical Analysis . . . 37
2.6 Related Work . . . 41
2.6.1 Ethnic Background, Gender and Academic Performance . . . 42
2.6.2 English Language Proficiency and Academic Performance . . . 44
2.6.3 Course Selection . . . 46
2.6.4 Prior Programming Experience . . . 49
2.6.5 Pre-University Academic Performance . . . 51
2.6.6 Student Characteristics and Academic Performance . . . 53
2.6.7 The Role and Influence of Mathematics in STEM Curricula . . . 55
2.7 Chapter Summary . . . 57

3 Research Methodology 59

3.1 Introduction . . . 59
3.2 Context of the Study . . . 60
3.3 Research Design . . . 61
3.4 Research Population . . . 61
3.5 Data Collection . . . 62
3.5.1 The NSC Grade 12 Dataset (2015-2019) . . . 62
3.5.2 The First Year BSc. Dataset (2015-2019) . . . 63
3.6 Data Analysis . . . 63

3.6.1 Data Mining: Using WEKA . . . 64
3.6.2 Statistical Analysis: Using SAS . . . 66
3.7 Proposed Methodology: CRISP-DM Process Model . . . 66
3.7.1 Business Understanding . . . 67
3.7.2 Data Understanding . . . 69
3.7.3 Data Preparation . . . 69
3.7.4 Modeling . . . 73
3.7.5 Evaluation . . . 73
3.8 Ethical Considerations . . . 73
3.9 Chapter Summary . . . 74

4 Results **76**

4.1 Introduction . . . 76
4.2 Ethnicity, Gender, Funding and Performance . . . 77
4.2.1 Ethnic Background, Gender and Academic Performance . . . 79
4.2.2 Ethnic Background, Funding Status and Academic Performance . . . 82
4.3 English Language Proficiency . . . 84
4.4 Influence of NSC STEM Subjects on the APS . . . 86
4.4.1 Feature Selection . . . 87
4.4.2 Linear Regression . . . 90
4.5 Prior Programming Experience (PPE) . . . 93
4.5.1 Computer Science I: First Semester Academic Performance . . . 94
4.5.2 Predicting Performance in Computer Science I . . . 97
4.5.3 Predicting the First Year Overall Result . . . 98
4.6 Minimum Admission Requirements . . . 100
4.6.1 Predicting First Year Outcome . . . 100
4.6.2 Predicting Degree Attainment (using Minimum Requirements) . . . 105
4.7 Student Characteristics and Academic Performance . . . 107
4.7.1 Cluster Analysis . . . 108

4.7.2 Association Rule Mining . . . 116
4.8 Mathematics I and Academic Performance . . . 120
4.8.1 Relationship Between Mathematics I and NSC Mathematics . . . 121
4.8.2 Predicting First Year Academic Performance . . . 122
4.8.3 Predicting Degree Attainment . . . 125
4.9 Evaluation . . . 126
4.9.1 Influence of NSC STEM Predictor Attributes . . . 128
4.9.2 Predicting Computer Science I (BCO, IAP) . . . 131
4.9.3 Predicting First Year Outcome (BCO,IAP) . . . 134
4.9.4 Predicting First Year Outcome (Minimum Requirements) . . . 137
4.9.5 Predicting Degree Attainment (Using Minimum Requirements) . . . 142
4.9.6 Predicting First Year Outcome (using Mathematics I) . . . 143
4.9.7 Predicting Degree Attainment (Using Mathematics I) . . . 146
4.10 Chapter Summary . . . 147

5 Discussion 150

5.1 Gender, Race and Funding Status . . . 151
5.2 English Language Proficiency . . . 153
5.3 Pre-University STEM Subjects . . . 155
5.4 Prior Computer Programming Experience . . . 156
5.5 Wits Minimum Admission Requirements . . . 158
5.6 Student Characteristics . . . 160
5.7 Role and Influence of Mathematics I . . . 163
5.8 Study Limitations . . . 164

6 Conclusion and Future Work 166

6.1 Future Work . . . 170

Appendix **173**

A Feature Selection and Regression Analysis Results 173

B Student Characteristics 174

C Clustering and Association Rule Mining 176

C.1 Clusters 176

C.2 Association Rules 177

D First Year and Degree Attainment Prediction 178

D.1 Mathematics - Predicting First Year Performance 178

D.2 Mathematics - Predicting Degree Outcome 178

E Model Evaluation 179

E.1 Predicting First Year Outcome (Minimum Requirements) 179

E.2 Predicting Degree Outcome (Minimum Requirements) 180

E.3 Predicting First Year Outcome (Mathematics I) 181

List of Figures

2.1 Car Loan Classification Data ((Weiss and Davison, 2010)) 24

2.2 Car Loan Decision Tree ((Weiss and Davison, 2010)) 26

2.3 Decision Tree Algorithms ((Song and Ying, 2015)) 27

2.4 Outcome classification Rule Set 28

2.5 Object Clusters ((Weiss and Davison, 2010)) 34

2.6 Types of Clustering Methods ((Chaudhari and Parikh, 2012)) 35

2.7 Sample T-test and ANOVA (Park, 2009) 40

3.1 Phases of the CRISP-DM Reference Model (Wirth and Hipp, 2000) 67

3.2 SAS Dataset 72

4.1 Multiple Linear Regression Correlation Output 91

4.2 SAS Multiple Linear Regression Output 92

4.3 Basic Computer Organization T-test Output 95

4.4 Introduction to Algorithms and Programming T-test Output 96

4.5 First Year Performance (English Home Language Training Set) 102

4.6 English Home Language Confusion Matrix 102

4.7 First Year Performance (English First Additional Language Training Set) 103

4.8 English First Additional Language Confusion Matrix 104

4.9 Predicting Degree Attainment (J48) 106

4.10 Degree Attainment Confusion Matrix (J48) 106

4.11 Estimating the best value of k 109

4.12 Using Eight Clusters 110

4.13 Using Nine Clusters (Optimal k=9) . . . 113

4.14 Using Ten Clusters (k=10) . . . 115

4.15 WEKA Object Editor Window . . . 117

4.16 Multiple Correlation (Mathematics) . . . 121

4.17 Confusion Matrix - Predicting First Year Outcome . . . 123

4.18 Correctly Classified and Incorrectly Instances . . . 124

4.19 Degree Attainment Confusion Matrix (MLP) . . . 126

4.20 Binary Classification Confusion Matrix . . . 127

4.21 Predicting the Admission Point Score (Test Set)) . . . 129

4.22 Actual Admission Point Score versus Predicted Admission Point Score (Test Set) . . . 130

4.23 Predicting Computer Science I (Basic Computer Organisation, Introduction to Algorithms and Programming) . . . 132

4.24 Predicting Computer Science I (Basic Computer Organisation, Introduction to Algorithms and Programming) . . . 133

4.25 First Year Test Set Records (Basic Computer Organisation, Introduction to Algorithms and Programming) . . . 134

4.26 Classifying Tesd Dataset Records . . . 135

4.27 First Year Outcome Binary Classification (Basic Computer Organisation and Introduction to Algorithms and Programming) . . . 135

4.28 First Year Outcome Receiver Operating Characteristics Graph (Basic Computer Organisation and Introduction to Algorithms and Programming) . . . 136

4.29 First Year Outcome Prediction Output (English Home Language) . . . 138

4.30 First Year Outcome Prediction Binary Confusion Matrix (English Home Language Test Set) . . . 138

4.31 English Home Language Receiver Operating Characteristics Graph . . . 139

4.32 First Year Outcome Prediction Output (English First Additional Language) . . . 140

4.33 First Year Outcome Prediction Binary Confusion Matrix (English First Additional Language Test Set) . . . 140

4.34 English First Additional Language Receiver Operating Characteristics Graph . . . 141

4.35 Degree Prediction Output . . . 142

4.36 Degree Attainment Confusion Matrix (using Minimum Requirements) . . . 142

4.37 First Year Outcome Prediction Output (Mathematics I) . . . 144

4.38 First Year Outcome Prediction (Using Mathematics I) . . . 144

4.39 First Year Outcome Prediction Binary Confusion Matrix (Using Mathematics I) . . . 145

4.40 Mathematics I ROC Graph . . . 145

4.41 Degree Attainment Prediction Test Set (Using Mathematics I) . . . 146

4.42 Degree Attainment Prediction Confusion Matrix (Using Mathematics I) . . . 146

1 Using Eleven Clusters (k=11) . . . 176

2 Association Rules (Support=0.01, Confidence=0.9) . . . 177

3 J48 Decision Tree: Predicting First Year Outcome . . . 178

4 English Home Language Test Set: Predicting First Year Outcome (483 records) . . . 179

5 English First Additional Language Test Set: Predicting First Year Outcome (579 records) . . . 179

6 English Home Language and English First Additional Language Test Set: Predicting Degree Outcome (432 records) . . . 180

7 Mathematics I Test Set: Predicting First Year Outcome (474 records) . . . 181

List of Tables

2.1 Test records for Outcome Classification 29

2.2 Meaning of Correlation Coefficient Values (Dancey and Reidy, 2008) . . . 38

3.1 Faculty of Science Admission Requirements 60

3.2 National Senior Certificate Race and Gender Statistics (2015-2019) 61

3.3 First Year Race and Gender Statistics (2015-2019) 62

3.4 National Senior Certificate Grade 12 Attributes 63

3.5 Selected STEM Subjects (NSC Grade 12) 64

3.6 First Year BSc. Attributes 65

3.7 BSc. Fields of Study (General & Structured Curricula) 65

3.8 Selected Attributes and Records 70

3.9 Selected Statistical and Data Mining Techniques 75

4.1 Demographic and Funding Descriptive Statistics 78

4.2 Admission Point Score Descriptive Statistics Segmented By Race, Gender and Funding Type 79

4.3 Overall Two-way ANOVA Results (Gender, Race and APS) 80

4.4 Adjustment for Multiple Comparisons: Tukey Kramer 80

4.5 Least Squares Means for effect Race level Pr $> |t|$ for H0: LSMEAN(i)=LSMEAN(j) Dependent Variable: Admission Point Score 81

4.6 Race and Funding Type Mean Admission Point Score Statistics 83

4.7 Overall Two-way ANOVA Results (Race, Funding Status and APS) . . . 83

4.8 English Home Language and English First Additional Language Statistics 84

4.9 Group Statistics (Admission Point Score) . . . 85

4.10 Normality Tests (English Home Language and English First Additional Language) . . . 85

4.11 Equality of Variances . . . 86

4.12 Independent Samples T-test . . . 86

4.13 Wits Admission Point Score . . . 87

4.14 Selected STEM Subjects . . . 88

4.15 Predicting the Admission Point Score (Without Feature Selection) . . . 89

4.16 Selected Feature Selection Techniques . . . 89

4.17 Predicting the Admission Point Score (With Feature Selection) . . . 90

4.18 Computer Science I Courses . . . 94

4.19 Multiple Linear Regression Results . . . 98

4.20 First Year Outcome Symbolic Codes . . . 99

4.21 Predicting First Year Overall Result Decision . . . 99

4.22 J48 Evaluation Summary (English Home Language and English First Additional Language datasets) . . . 104

4.23 Selected Rules (Five Attributes) . . . 119

4.24 Mathematics I - Predicting First Year Overall Result . . . 123

4.25 Summary Statistics - Predicting First Year Overall Result . . . 123

4.26 Mathematics I - Predicting Degree Achievement . . . 125

4.27 Predicted versus Actual Admission Point Score Values . . . 128

4.28 Predicted versus Actual Computer Science I Scores (Basic Computer Organisation and Introduction to Algorithms and Programming) . . . 132

4.29 Receiver Operating Characteristics Critical Points . . . 139

1 Feature Selection Results . . . 173

2 Linear Regression Results . . . 173

3 Student Characteristics . . . 174

4 k-means Cluster Analysis . . . 176

Chapter 1

Introduction

1.1 Background Of The Problem

One of the most important goals of the South African government, through its Department of Higher Education, is to provide good quality tertiary education to as many deserving and academically capable students as possible. For South Africa to improve its economic status and to achieve economic prosperity, there is an urgent need to develop a highly trained and educated workforce. In this regard the tertiary education institutions can play a major role and make an important contribution. The country and the economy rely on these institutions to produce a continuous stream of skilled graduates, particularly in scarce skills such as, vocational and technical, science, technology, engineering and mathematics disciplines. To achieve this goal the government, through the National Treasury Department, allocates a considerable amount of money every year to subsidize higher education institutions. Despite this laudable effort, it will be difficult to satisfy the demand for access to higher education by an ever increasing number of needy, academically capable students, given that insufficient money is allocated to universities to fulfil this purpose.

In order to try and address the problem of access to higher education, the National Student Financial Aid Scheme (NSFAS) (Jackson, 2002) was introduced in 1999 to provide financial assistance in the form of loans and bursaries, to needy students who qualify to study at South Africa's tertiary institutions. Between 2013 and 2017 the NSFAS has supported over two million tertiary-level students, affording them the opportunity to enrol in the country's universities and colleges. Due to increasing numbers of eligible students who apply for financial aid, the money allocated to the NSFAS from the higher education budget has also been increased, to try and meet this demand. For example, during the 2014/15 financial year R10 billion was allocated to the NSFAS, during the 2016/17 financial year the scheme was allocated R14.3 billion. Apart from the NSFAS,

students receive funding from other agencies such as Scholarships(SUG), Government Bursaries (GB), External Bursaries (EB), Corporate Bursaries (CB), Other Bursaries (B) and Sports Bursaries (SB). Students who excel in their studies receive merit (M) and Fellowship (F) awards, regardless of their financial background or status.

The nationwide student protests that took place between 2015 and 2016, under the auspices of the #FeesMustFall movement (Baloyi and Isaacs, 2015), (Naidoo and McKay, 2018), generated a higher education crisis and brought into sharp focus the glaring inadequacy of student funding. During this period almost all of South Africa's public universities and colleges were brought to a standstill by this nationwide student protest, primarily over the issue of fee increases that were being planned (Baloyi and Isaacs, 2015). The students' demands included, amongst others, a 0% fee increase, free education, an end to the outsourcing of workers employed by labour brokers, as well as the transformation and decolonization of South African universities. An important victory of this movement was to get the government to agree to its demand of a 0% fee increase. Another major victory, following the demand for outsourced workers to be hired permanently, was to obtain the commitment of some universities to provide permanent employment to their outsourced workers. This demand has been met by some universities, including ours. It is, therefore, evident that some of the issues pertaining to the crisis currently facing higher education institutions are directly related to inadequate student funding and the funding model used to support the increasing number of eligible students who view access to higher education as their hope for a better life in the future.

Given the millions of taxpayers' money that have, thus far, been spent towards funding needy students, it therefore becomes a matter of interest to evaluate the effectiveness of this funding in terms of, among other things, its impact on these students' academic performance. Insightful and meaningful information about them can be extracted from the vast amounts of data which universities collect and store in their repositories and archives. In the educational research environment, Educational Data Mining (EDM) (Baker et al., 2007) has emerged as a discipline in which different data mining techniques are used to explore and extract information from educational datasets, in order to obtain a better understanding of students and the environments in which they learn. In this data mining based longitudinal study (2015-2019), conducted at the University of the Witwatersrand (also known as "Wits University"), different data mining and statistical techniques were used to process and extract information from a historical dataset of records belonging to first year Bachelor of Science (BSc.) students enrolled in critical-skill STEM degree programmes. The data for the study was provided by the university's Academic Information and Systems Unit (AISU).

For many students, their first year of study marks the beginning of a challenging and critical journey in their academic life that, hopefully, should result in them achieving

academic success and completing their studies in record time, after which they can proceed to pursue professional careers of their choice. Given South Africa's need for highly skilled professionals who possess STEM-based qualifications, the first year of study for students enrolled in STEM programmes marks a critical entry point. The transition between high school and university is navigated during the first year of study. Students who bridge the high school to university gap and manage a successful transition, stand a much better chance of retention and progression within the system. By applying different data mining techniques on student datasets, meaningful and useful insights can be obtained about students, information which higher education institutions and other stakeholders can use in a number of ways, which include identifying students at risk for failure, predicting the possibility of retention, progression or attrition, and making well-informed decisions.

The purpose of this study was to obtain a clear picture of first-time first year students enrolled in STEM-related BSc. degree programmes at Wits University. Using data mining and statistical techniques, it was possible to extract meaningful, insightful and useful information about these students from a historical dataset of their records. This study was driven by the need to better understand this category of first year students and, given their profile in the educational context, it was deemed fit to analyze both relevant and interesting information to get to know more about them, by conducting an in-depth and detailed study about them. In this regard, this data mining based study looked at a number of issues. For example, given the importance of English as the language of teaching in South African universities and its use as one of the criteria for admitting new students in the STEM degree programmes, it was key to discover its influence and contribution towards the academic performance of funded first year STEM students, the majority of whom do not use English as their home language. In addition, apart from English, this study looked at other Grade 12 NSC subjects and the strength of their individual contribution towards the Wits Admission Points Score (WAPS) and academic performance in first year.

Given the vast sums of money the government spends in funding students at the beginning of their studies, an important issue to consider in this regard is their retention within the educational environment. We then directed our attention to the critical topic of progression from first to second year, in order to predict whether the funded first year students who enrol in the STEM-based degree programmes will succeed and proceed to second year. Apart from being able to know whether a funded first year STEM student will proceed to second year (retention), an equally important issue to consider is whether such a student will eventually qualify and obtain their STEM degree of choice, preferably in record time. By investigating these and other pertinent issues about funded first year students in STEM disciplines, we hope that this study will shed some light and

insights about this category of students, and in that way contribute to existing information and knowledge about them, which stakeholders in our educational environment and elsewhere can draw upon to make better, well-informed decisions.

1.2 Problem Statement

In 2018, South Africa was amongst the countries ranked by the World Bank as an upper middle-income economy (Abraham and Nosa, 2018). The Gross National Income (GNI) per capita in upper-middle-income economies ranges from $4,036 to $12,475. By developing a highly skilled and educated workforce, South Africa has the potential to lift itself from its current status as a developing economy to a high-income economy. For South Africa as a developing country, numerous economic and social benefits can accrue through the provision of high quality education, technical and technological skills, as well as training opportunities to her citizens. The promotion and acquisition of a tertiary education has been shown to contribute towards a country's economic growth (Nowak and Dahal, 2016). Likewise in South Africa, employment opportunities increase as individuals achieve higher levels of education and skills. Furthermore, their earning power and productivity increase as they acquire more education and skills (Radcliffe, 2017).

According to (Radcliffe, 2017), employers and companies prefer a workforce which is highly educated, skilled and motivated, a workforce comprising productive and highly independent individuals. The basis of a good economy is a sound education system, which can produce individuals and professionals capable of working and contributing their acquired knowledge in scarce-skill, science and technology-based industries (Radcliffe, 2017). A recent City & Guilds Group case study (Group, 2015) has highlighted the economic gains South Africa is likely to achieve if it were to provide adequate funding and support towards the vocational education and training sector, a development that could result in a massive reduction in the youth unemployment currently prevailing in the country. For example, in 2016 the unemployment rate amongst South Africa's youth was more than 50% (Kruger, 2016). In 2020, the youth unemployment rate stood at a massive 59.62% (Oosthuizen, 2019). This is not helped by the fact that many of South Africa's young people of working age are "neither employed nor in education or training (NEETs)" (OCDE, 2016). The fact is also highlighted by (Kruger, 2016) that social inequality is one of the serious problems South Africa needs to address. For example, (Kruger, 2016) states that the "richest 20% of the population account for over 61% of consumption while the bottom 20% account for less than 4.5%".

A strong and sound education system is required, as it can address many of the economic and social problems currently facing South Africa. The primary focus should be

to produce highly skilled graduates and professionals who, armed with suitable qualifications, can hopefully contribute towards filling the critical scarce-skills gap which South Africa is currently experiencing, and which the country is finding extremely difficult to close. Focusing specifically on the need for professionals with skills and qualifications in STEM-related disciplines, the education system, for a number of reasons, is insufficiently prepared or equipped to produce the required type and number of graduates with the scarce skills necessary to address some of the problems mentioned earlier. Starting with the high schools, particularly rural and township schools, there is a need to appoint and deploy highly qualified and competent teachers to offer STEM-related subjects across all the grades, from the lowest to the highest grade. Unfortunately, there is a dire shortage of teachers who possess STEM-related qualifications, who can be deployed to teach in our high schools, particularly rural and township high schools (for Educators, 2010), (Tshabalala and Ncube, 2014), (Muremela et al., 2023). Since our high schools do not produce sufficient numbers of Grade 12 learners who possess vocational, technical and other STEM-related qualifications, there is always a great shortage of learners with these qualifications who enrol for the first time in our tertiary institutions. As a result, tertiary institutions receive less than the required number of students they need to enrol in their STEM-related diploma or degree programmes. This means that not many graduates will be produced to satisfy the demand for the much-needed scarce skills required to develop and inject growth into the economy, and to generate social and economic prosperity for the South African population.

The high failure rate amongst first-time first year students is a problem that affects many higher education institutions, globally and locally (Singh et al., 2016). Failure to perform successfully in first year has serious ramifications. If they do not drop out, under-performing students risk failure, a situation that may affect their progress and retention. Many of our first year students require financial support in order to access higher education. Should some of them fail or drop out, another round of funding is required to enable them to repeat their first year. This is money that could have been used to fund more, equally capable and deserving students. The high first year failure rate in the higher education institutions also affects students who enrol in STEM degree programmes and whose studies are funded by agencies such as the NSFAS (Letseka and Maile, 2008), (Moodley and Singh, 2015). For a country and an economy that need the scarce skills which graduates with STEM qualifications can provide, failure in first year can be regarded as catastrophic. The fact that barriers to access to higher education have been reduced, thanks to the #FeesMustFall student movement, cannot be used as an argument to continue sustaining and supporting under-performing students for longer than is necessary (Wildschut et al., 2020), (Naidoo and McKay, 2018). With many constituencies competing for funding from government, there is only so much that can be provided to fund free education.

Acquiring a STEM-based undergraduate degree creates career opportunities which students can exploit to apply their newly acquired scarce skills, for their own benefit and that of the country and the economy. Currently, not much research is being conducted, if any, within the colleges and universities to obtain and provide insightful information about funded first year students who enrol in STEM disciplines. Not much information is available on these students, about whom a number of important and interesting questions can be asked. How good is their academic performance, compared to non-funded students? What are the characteristics of students who drop out or persist in their studies? Are these students likely to qualify (complete their degree, preferably in record time)? What is the gender balance and the enrolment choices these students make in specific STEM fields? What are the attrition and retention rates of these students, compared to non-funded students? Which courses do students who drop out or fail, enrol in, compared to those who succeed? Generally, how can the low enrolment rate in STEM fields be addressed? More specifically in our case of our institution, how can the low enrolment rate of funded first year STEM students be increased? What is the role of Grade 12 STEM courses in the success of funded first year students enrolled in STEM programmes? These and other unanswered questions about this category of students warrant an investigation that should reveal meaningful and useful information about them, information which should enable us to get to know these students better. Furthermore, different stakeholders can use this information in making well-informed decisions about them.

Meaningful and useful information about students can be obtained by processing and analysing the vast amount of data higher education institutions collect about them. In this data mining based study we used historical data (2015-2019) containing records belonging to first year students enrolled in STEM degree programmes. It is hoped that this study will provide useful information and insights regarding the role, influence and contribution of financial support on the academic trajectory of first-time first year students enrolled in STEM disciplines.

1.3 Significance of the Study

This study is being conducted at a time when South Africa is facing serious economic, social and educational challenges. Economic growth is at an all-time low, and there is a very high unemployment rate, especially amongst the country's youth of working age. The government is facing a huge problem, that of having to find enough money to satisfy the competing needs of different stakeholders in society. Following the nationwide student protests of 2015-2016, around the issue of student funding and free entry to tertiary education, the government responded to this demand by providing billions of rands to provide free education to thousands of needy, academically capable students. In order for this initiative to bear fruit, there is a need to channel these resources towards supporting many students to obtain qualifications in critical, scarce-skill STEM disciplines. The growth and success of the economy depends on these critical skills. A high-employment economy, characterised by a high growth rate, and driven by a skilled and professional workforce, offers numerous advantages and benefits to society. A good economic environment enables the state to collect more taxes, in which the money is used to fund different activities such as health, education, old-age pensions, child support grants, and infrastructure development. The benefits to society include, amongst others, social cohesion, reduced crime rate, political stability, greater social mobility and a low unemployment rate (OECD, 2013), (Great Britain. Department for Business and , BIS), (Ma et al., 2016)

Currently, not many studies in our educational context are conducted to extract insightful, interesting and useful information from student datasets. In this study we applied different data mining techniques on a historical dataset of records belonging to first year students enrolled in scarce-skill STEM disciplines. There are several ways in which a study of this nature can provide useful information about these students to the stakeholder community. In the Faculty of Science, the following minimum NSC requirements must be met for first-time first year students to be admitted to STEM disciplines of their choice: Wits Admission Point Score (APS), English Home Language (EHL/ENA) or English First Additional Language (EFAL/ENB), Physical Science and Mathematics. Student retention from the first to the second year is a useful measure of academic achievement. Since our faculty uses the three NSC minimum requirement subjects to admit first year students into STEM degree programmes, can these subjects be used to predict retention from the first to the second year of study?. Given the critical need for STEM-qualified graduates, the findings on student retention can be used to enable

academics and administrators to identify "at-risk" students, to prevent them from dropping out or failing their year of study. These students can be assisted in various ways to ensure their retention from first to second year.

In our high schools and tertiary institutions such as Wits, English is used as the language of teaching. The language plays a pivotal role in teaching different high school subjects such as Mathematics, Physical Science, Biology, and Information Technology. We expect students who are proficient in English to understand the subject content better, and therefore to perform better academically compared to their peers, whose English proficiency is relatively poor. This study explored the important and interesting idea concerning English language proficiency and its influence on the overall result of first time first year students enrolled in different STEM fields in our university's Faculty of Science. We also compared the mean Admission Point Scores obtained by ENA and ENB students, to determine whether the difference between the scores obtained by these two English language groups was significant. Our findings regarding the role and influence of English as a language of teaching in our curriculum can be used by different stakeholders in the university, such as the Wits Writing Centre to assist students identified to be weak in English to achieve a better and higher level of English language proficiency.This will enable these students, in their different courses, to better comprehend and understand the different aspects of the courses they are taking.

Apart from English and Physical Science, Mathematics is another compulsory NSC subject required for admission to STEM-related fields in our faculty. Given the important role and influence of Mathematics in STEM curricula, several studies have looked at how Mathematics overall academic achievement in first year. In the South African context research is being conducted in different fields to determine the influence of high school Mathematics in the overall academic achievement of first year students (Rankin et al., 2012), (Tewari, 2014), (Mthimunye and Daniels, 2017). High school Mathematics is amongst the key subjects students need in order to be accepted by any of the schools in our faculty. The role and influence of Mathematics in our curriculum requires a more in-depth investigation and understanding of its relationship with the academic achievement of students in their first year. There are several ways in which our findings on the role and influence of Mathematics can provide useful information to different stakeholders in our educational environment. The results may be used by lecturers, as well as school and faculty administrators, to target and provide assistance and counselling to students identified to be at risk for failure, due to poor performance in Mathematics.

The study by (Venkatas et al., 2014) used faculty-wide academic records belonging to first year optometry students, to investigate the relationship between the APS and the academic performance of funded first year students enrolled in STEM disciplines. In our context, where a similar relationship has been studied, the findings can be used in

various ways. For example, stakeholders can leverage the results obtained to make well-informed decisions on matters such as student selection, identifying appropriate STEM degree programmes, identifying at-risk students, and so on.

Another issue of specific interest to our school, where computer programming is a key component in our curriculum, is the impact of prior computer programming experience, as acquired in high school, on academic performance in first year Computer Science. More specifically, we were interested to see how students with computer programming skills would perform in two first-semester Computer Science I courses, compared to students who previously have not studied a computer programming subject, such as Information Technology. If certain students are identified during or at the end of the first semester to be "at risk" for failure, the school can take proactive steps to assist them in various ways to prevent them from failing Computer Science I. For example, faculty and school administrators can assist by providing services that include counselling, tips on study techniques and time management. Lecturers and student assistants (tutors) can provide targeted tutorials, assignments and extra lessons to help at-risk students to be successful in areas where they are weak or vulnerable.

The challenges and constraints facing funding schemes such as the NSFAS is the need for students who receive funding to finish their studies in record time or very soon thereafter. Students who finish their studies on time enable funding schemes to save more money, which can be used to provide the opportunity to equally deserving and capable students to receive funding for their own studies. It is important to be able to know whether a funded student will qualify on time. Such information can save a significant amount of money. A student identified early to be at risk of not qualifying on time can receive proactive assistance to prevent failure, and continue to eventually qualify, with the money saved being used to fund other capable and needy students. This is very important, since a significant number of students who apply for funding do not necessarily receive any financial support. In addition, knowing whether or not individual students will qualify in a particular academic track can help university admission personnel to guide and advise them to enrol in programmes that best suit their academic capabilities, thereby increasing their chances of qualifying on time.

1.4 Research Question

A number of specific questions emerged from the statement of the problem (Section 1.2), which this study sought to address. In this study we followed a quantitative research design, in which different types of quantitative research questions were used, such as declarative, comparative and relationship-based questions. To begin with, it was necessary to start by posing the main research question, the statement of which is as follows:

> *How can data mining and statistical techniques be leveraged to obtain useful and meaningful information from the historical data belonging to cohorts of first-time first year students enrolled in STEM degree programmes at a South African university?*

The following sub-questions enabled the response to the main research question:

1. What are the relationships amongst the factors of ethnic background, gender, funding status and the Wits Admission Point Score (APS) obtained by first time first year BSc. students enrolled in STEM disciplines? In order to address this question in a more focused manner, we divided it into two sub-questions, as follows:

   (a) What is the relationship between gender and ethnicity to the Admission Point Score (APS) obtained by first time first year students enrolled in STEM degree disciplines within the Faculty of Science?

   (b) What is the relationship between ethnic background and funding status to the Admission Point Score (APS) obtained by first time first year students enrolled in STEM disciplines?

2. Is the APS obtained by National Senior Certificate (NSC) students who select the English Home Language (ENA) subject significantly different from that obtained by students who select the English First Additional Language (ENB) subject?

3. How significant is the relationship between the Admission Point Score (APS) obtained by first time first year STEM students and their NSC Grade 12 STEM subject scores?

4. Is there a significant difference in performance in the two first-semester Computer Science I topics, namely Basic Computer Organization I (BCO) and Introduction to Algorithms and Programming I (IAP), between first time first year BSc. students with prior computer programming experience and their peers, who do not have prior computer programming experience?

5. How suitable are the Faculty of Science's minimum requirement subjects as predictors of first year and degree attainment outcomes?

6. What is the role of student characteristics on academic achievement amongst first time first year BSc. students enrolled in STEM disciplines at Wits University?

7. To what extent does Mathematics I (Algebra I and Calculus I) predict the academic outcome and degree attainment for first year students enrolled in STEM disciplines?

1.5 Research Approach

Given the primary focus of this study as a data mining based project, we chose the data mining *Cross-Industry Process for Data Mining* (CRISP-DM) methodology (Chapman et al., 2000) to conduct our study. CRISP-DM is a six-phase non-proprietary and freely available standard process model, which is widely adopted and used in different data mining projects. To conduct our study as a data mining project, we needed to access the data collected by Wits University about first time first year students enrolled in STEM disciplines. To obtain the required data, we approached the university's Academic Information and Systems Unit (AISU), which provided us with historical Grade 12 (high school) and first year student academic records, from 2015 to 2019. The records also contained information about the funding status (funded or non funded) of the first year students in the Faculty of Science. We followed this data collection step by describing, exploring and checking the data for quality.

The data needed to be prepared to create the dataset(s) against which data mining algorithms would be applied. Several tasks such as selecting the data, cleaning, integrating and formatting it had to be performed. To do all this, the Java-based open source toolkit, WEKA (Waikato Environment for Knowledge Analysis) (Hall et al., 2009) was used. WEKA consists of four main applications, namely the Explorer, Experimenter, Knowledge Flow and Simple CLI applications. In this study, we used the Explorer application. Using WEKA, we performed a variety of data mining tasks and experiments. These included data pre-processing, attribute selection, data visualization, classification, prediction, clustering and association rule mining. Having formatted the data during the pre-processing stage, we were thereafter able to use it during the Modelling Phase to create different classification models. We used the Statistical Analysis System (SAS) software package to perform different statistical analysis experiments. SAS, as explained by (Der and Everitt, 2008), is used for data manipulation and statistical analysis purposes. Different statistical techniques were used in this study. These included a two-sample T-test, two-way analysis of variance (ANOVA), linear regression and correlation analysis.

Having completed the data preparation process , the next phase (modeling) followed, in which the following tasks were performed: selecting the experimental techniques to be used, dividing the dataset for training and testing purposes, using algorithms to create different classification models, and evaluating the models. Briefly, the following tasks were performed:

1. From the WEKA toolkit and the SAS software package, the following classes of data mining and statistical analysis techniques were selected and used:
    - *J48, NaïveBayes, MulitilayerPerceptron, JRip, LinearRegression* (classification and prediction algorithms),
    - *SimpleKMeans* (clustering algorithms),
    - *Apriori* (association rule mining algorithm),
    - *T-test, two-way ANOVA, Pearson's correlation* (statistical analysis techniques).
2. After selecting the techniques, and before the modeling phase, the 2015-2019 dataset consisted of two subsets, which we used for training (2015-2018) and testing (2019) purposed.
3. Next, we applied the algorithms we selected earlier against the training subsets.
4. Using different performance criteria, the different models created during the training stage were assessed and ranked. These criteria included different metrics (Du, 2010), such as accuracy, precision, recall, F-measure, Area Under the Curve-Receiver Operating Characteristic (AUC-ROC), Root Mean Squared Error (RMSE), Mean Squared Error, Mean Absolute Error and the Kappa statistic. To assess the performance and interestingness of generated association rules, the support, confidence, lift and conviction measures were used.

Finally, during the Evaluation Phase of the methodology we used in this study, we assessed the models we created during the Modelling Phase, to determine their quality, suitability and effectiveness. After the Evaluation Phase, a decision can be made on whether the evaluated models can be used by different stakeholders in our environment to make well-informed decisions about the first year STEM students in our faculty.

1.6 Research Highlights and Contribution

This study contributes in a number of ways to the understanding of how educational research can benefit from leveraging data mining techniques to extract meaningful and

useful information from the vast amounts of data collected by higher education institutions about their students. In the broader South African educational research context, Educational Data Mining (EDM) has yet to entrench itself as a field in which sustained research is conducted. Currently, as far as it is known, not many tertiary institutions in South Africa conduct research in which data mining techniques are used to obtain meaningful and useful information from the data they collect about their students. This study, therefore, contributes towards addressing the knowledge gap resulting from the lack of sustained data mining based educational research being conducted in local higher education institutions. In the following subsections, we highlight some of the ways in which this study contributes to research in the field of Educational Data Mining, and to a better understanding of the first year STEM students in our faculty.

1.6.1 Prior Programming Experience and First-Semester Performance

In the School of Computer Science and Applied Mathematics (CSAM) school, first year BSc. students enrolled either in Applied Computing or Computing and Mathematics disciplines, are required at the end of the first semester to have completed two Computer Science I modules, namely Basic Computer Organization (BCO) and Introduction to Algorithms and Programming (IAP). In one of the findings from a study conducted in our school by (Modise, 2016), the performance of students in these two modules was significantly different, between students with prior computer programming experience (did Information Technology in high school) and those who had not programmed before. Students who did Information Technology at high school obtained significantly higher marks in these two modules compared to students who did not do the subject. Our study obtained results similar to (Modise, 2016)'s. We found that in these two modules, the difference in performance between students with prior computer programming experience and those who lacked such experience, was significant. This study contributes towards highlighting, in computer-related disciplines in our faculty, the importance of a prior computer programming qualification as an indicator of academic success in the first-semester of Computer Science I modules, namely BCO and IAP. Within our context, these findings can be used to identify academically vulnerable students early during their first year, after which proactive measures can be taken to prevent them from dropping out or failing at the end of the first semester, thereby ensuring their retention in Computer Science I.

1.6.2 Research in Educational Data Mining (EDM)

As far as we are aware, Wits University is one of very few higher education institutions in South Africa which place data mining and research in Educational Data Mining (Baker et al., 2007) at the centre of the educational research effort. A study conducted earlier in our school by (Mashiloane, 2016),including studies by (Modise, 2016) constitute an important contribution towards the ongoing effort aimed at promoting and encouraging the adoption of Educational Data Mining as a research discipline in the South African educational research environment. Given the vast amount of data collected and stored by tertiary institutions about their students, Educational Data Mining provides the opportunity for the data to be leveraged in order to obtain insightful and actionable information about these students. Such information can be used by institutional stakeholders and researchers for, among other things, such purposes as decision-making and generating new knowledge about the students.

1.6.3 The Research Methodology

In order to conduct this data mining-based study we had to identify and implement an appropriate research methodology to guide us through the different stages of the research. The Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). CRISP-DM is a relatively new standard methodology, which is increasingly being used within the data mining community to guide practitioners in a variety of data mining projects. The CRISP-DM project life cycle comprises the following six phases: *understanding of the business, understanding of the data , data preparation, modeling, evaluation* and *deployment*. Except for the last, deployment phase, this study successfully completed the first five phases of the methodology. Given the nature of this study and its primary focus on data mining and data mining techniques, CRISP-DM became the natural choice of a methodology we could follow in conducting our research. Whilst CRISP-DM has been adopted and is being used in different disciplines and application areas as a data mining methodology of choice, more work remains to be done to adopt and use it as a standard methodology in educational research contexts. Currently, there are few instances where CRISP-DM has been used in educational data mining research contexts. Examples include studies by (Kovačić and Green, 2010), (Vialardi et al., 2011) and (Kabakchieva, 2013).

In the local, South African educational context, research in Educational Data Mining (EDM) is still in its infancy. There is not much evidence available pointing towards a concerted effort being taken in the educational research environment to adopt CRISP-DM as a methodology of choice for executing EDM-based projects. This study, therefore,

represents a modest beginning in highlighting and promoting CRISP-DM as a viable research methodology that can be used effectively in EDM-based research projects.

1.6.4 The Role of Mathematics

Mathematics features prominently in all the STEM-related fields of study offered by our faculty. In most of the BSc. degree options on offer in the faculty, first year students must take the compulsory Mathematics I major subject. Our results have demonstrated the importance of mathematics in STEM-based curricula, in our high schools and tertiary institutions. Amongst the NSC STEM subjects commonly selected by NSC Grade 12 students, Mathematics was the strongest and significant subject in predicting the Admission Point Score (APS). Its correlation with Algebra I and Calculus I was positive, weak and significant. Algebra I and Calculus I were strongly and significantly correlated to each other and together, as Mathematics I, emerged as the strongest predictors of the overall first year result. As for predicting degree achievement, Mathematics I turned out to be a very poor predictor of degree achievement. Our findings regarding Mathematics I contribute towards a better understanding of the critical role and influence of this subject in our STEM curricula. In educational contexts similar to ours, where Mathematics features prominently in the curriculum, this subject could be investigated for its role as a predictor of success or failure in STEM degree offerings.

1.6.5 Model Evaluation

In any data mining project, such as ours, it is not enough to create models that are not evaluated in detail to determine their quality and effectiveness, before they are used in real-life application environments. During the Modeling Phase we applied different data mining algorithms against our training datasets. From this task different data mining models were created. During the Evaluation Phase, we evaluated the models in great detail, by applying them against previously unseen test datasets records. Using different evaluation metrics, we were able to determine the models' quality and effectiveness in performing their tasks. To evaluate these different models, we used several metrics. Our evaluation approach, the detailed nature of the work we did in evaluating the models, and the insights we obtained during the model Evaluation Phase, can perhaps be used to inform and guide educational researchers involved in studies similar to ours on how to conduct a systematic and detailed evaluation of their own data mining models, *before* using them in real data mining environments and applications.

1.7 Structure of the Dissertation

This document comprises six chapters, a brief summary of which is presented as follows. Chapter 2 begins with a background discussion of topics relevant to this study. These topics include data mining and its applications, and Educational Data Mining (EDM). A detailed review of the literature is also conducted in this chapter. In Chapter 3 we present and discuss in detail the six-phase CRISP-DM methodology we adopted in conducting the study. In Chapter 4 we present and interpret the results we obtained following the data mining and statistical analysis experiments we performed during the CRISP-DM Modeling Phase. In Chapter 5 we provide a detailed discussion of the results we obtained in this study. We compare and contrast these results with findings from similar studies. There is also a discussion of some of the limitations associated with our study. Finally, in Chapter 6 we recommend, based on our results some of the actions and decisions stakeholders in our context could take regarding the first year STEM students in our faculty. The chapter also summarizes the main findings of this study, and concludes by presenting a few ideas for further research.

Chapter 2

Background and Related Work

2.1 Introduction

Universities collect and store a lot of data about their students. By processing and analyzing such data, institutions are able to extract useful and meaningful information from it. In many application areas data mining techniques are increasingly being used to extract useful information patterns that lie hidden in data sets. For example, in the banking and insurance industries these techniques are being used to predict whether clients of a certain profile are likely to default on their loan or insurance payments. Data mining techniques are increasingly being applied in educational environments, where vast amounts of educational data is collected and stored by higher education institutions. Educational Data Mining (EDM) is a relatively new research field, in which data mining methods are used to obtain information from educational datasets. Such information can be analyzed to obtain more knowledge about students and their learning environments. In this study different data mining and statistical techniques were applied on historical data sets belonging to first time first year BSc. students enrolled between 2015 and 2020 at the Witwatersrand University's Faculty of Science.

This chapter begins by providing the information necessary for understanding how this study was conducted. This chapter also includes a discussion of the literature related to this study. For a better understanding of its contents, the chapter consists of the following sections. Section 2.2 introduces and discusses the concept of data mining. The discussion also includes example applications of data mining methods. The discussion in Section 2.3 is on Educational Data Mining. In Section 2.4 we provide some of the reasons for using data mining methods in this study. Section 2.5 contains a detailed discussion of the data mining and statistical techniques we used in different experiments. In Section 2.6 we provide a detailed discussion of some of the work related to this study. Finally, a summary of the chapter is provided in Section 2.7.

2.2 Background

Data mining and data mining techniques feature prominently in this study. For this reason, a detailed discussion of the subject is required, which should hopefully lead to some understanding of the issues that were tackled in this study. This section comprises two subsections. In the first, Subsection 2.2.1, the concept of data mining is introduced and discussed. In Subsection 2.2.2, some examples of areas in which data mining techniques are applied are provided.

2.2.1 Data Mining

Data mining, also known as Knowledge Discovery in Databases (KDD) (Kaufinann, 2006), focuses on obtaining knowledge from the information extracted from large databases. Such knowledge can be used by practitioners and stakeholders, depending on the particular context, to make well-informed decisions. In data mining statistical techniques and mathematical models are also used to check whether the data follows a certain pattern, and whether certain trends can be identified within the data. According to (Hegland, 2001), the uncovered patterns are characterized by a variety of measures and characteristics such as accuracy, precision, expressiveness, interpretability, parsimony and surprisingness.

Prediction and *description* are the primary goals of data mining (Tamilselvi and Kalaiselvi, 2013). In predictive data mining, models are created which perform tasks such as classification and prediction. In the classification task, a model is created whose task is to classify a previously unseen data instance, by assigning to it a particular class or category (Larose, 2014). With prediction, a model predicts a variable's numerical value. Regression analysis is an example of a predictive data mining task. Descriptive data mining identifies patterns and trends present in the data, including the relationship between the data attributes. Descriptive data mining tasks include clustering, association rule mining and outlier detection (Larose, 2014). In cluster analysis the data is partitioned into different, homogeneous groups referred to as "clusters". Similar objects are placed in their own cluster, separate from other objects, which reside in their own clusters. In association rule mining, rules are generated which describe relationships amongst dataset attributes.

There are *supervised* and *unsupervised* data mining techniques. In supervised learning an algorithm is trained to predict the class of individual data records by applying it against a dataset whose records have known class labels. The result from this learning process is a model that can be used for classification or prediction purposes. The model's effectiveness is determined by using it against a test dataset whose target variable has

unknown class labels. In unsupervised learning, unlike in supervised learning, there is no target or class variable. The purpose of the algorithm used in unsupervised learning is to determine relationships amongst the dataset variables, and to determine the degree to which such relationships are significant. Examples of unsupervised learning methods include clustering and association rule mining. In Section 2.5 we provide a detailed discussion of the data mining and statistical analysis methods we used in this study.

2.2.2 Applications of Data Mining

Data mining methods have, for a number of years been used successfully in different application areas. For example, in the business sector these techniques have been used to obtain insights from the data, including being deployed in business systems to help generate and increase business profits (Algarni, 2016). According to (Bramer, 2007), successful applications can be found in diverse areas such as satellite imaging, organic compound analysis, automatic abstraction, and so on. Even though the same data mining methods may be used in different applications, the objectives behind their use may not always be the same (Romero and Ventura, 2013). According to(Huebner, 2013), there is not much evidence to show that educational environments are using data mining techniques to obtain information and insights from the huge volumes of data which educational institutions collect and store in their repositories. According to (Luan, 2002), some of the data mining techniques currently being used in the business sector can also be used in higher education research environments. In the following subsections, we provide some examples where data mining techniques are used successfully.

Data Mining Applications in Social Media

Social media mining is an emerging field which involves the extraction, analysis and representation of useful patterns present in social media data or data derived from social interactions (Zatari, 2015). The information can be used by researchers in disciplines such as sociology, business, psychology, entertainment, news and politics.

Data Mining Applications in Health Care

Data mining techniques are used in the health care industry, for example, to enable health care insurers to detect fraud and abuse (Koh et al., 2011). They are also used to assist health care providers in decision making, such as enabling them to identify effective treatments and implement best practices. Other uses include designing appropriate

interventions and the reduction in claims and hospital admissions, thereby contributing towards better management in the provision of health care services.

Data Mining Applications in Banking

In the banking sector data mining techniques are used to achieve better performance in some of the industry's core business processes (Bhambri, 2011). For example, these techniques are used to profile customers by grouping similar customers together. The customers' profiles can be used for purposes such as marketing, risk analysis and, based on customer needs, for revising customer policies. Customer data can also be analyzed to identify information that could reveal fraudulent activities. Furthermore, data mining methods can be used in the banking sector to differentiate between customers who are most likely to repay their loans and those who are most likely not to pay their loans.

Data Mining Applications in Cloud Computing

(Petre et al., 2012) discuss the importance of data mining methods in cloud computing. The techniques are used, for example, to create structured information from data sources that are either unstructured or semi-structured. Through cloud computing, organizations are able to centralize their resources, with the assurance that using data mining techniques, they can provide efficient, reliable and secure services to their clients.

2.3 Educational Data Mining

In the recently emerged field of Educational Data Mining the primary aim is to obtain, using data mining methods, useful information from the vast amount of data which universities collect about their students and other stakeholders. Within an institution the data is obtained from different sources, and integrated by placing it in a central, main repository. In (Bienkowski et al., 2012), Educational Data Mining is defined as "a research area with a suite of computational and psychological methods and research approaches for understanding how students learn." (Siemens and Baker, 2012) mention the fact that data mining methods can be used to develop and test scientific theories, including the formulation of new scientific hypotheses. (Pavlik Jr et al., 2009) use data mining techniques to recommend activities, tasks, links to visit, problems to work on and courses students can choose to study.

(Algarni, 2016) states that EDM methods are being used to address a wide variety of challenging educational problems, which include the following: identifying students at-risk for failure, prioritizing the various needs of groups of students from a learning perspective, improving graduation rates, assessing the overall performance of institutions, maximizing resources on campus and optimizing the renewal of the curriculum. In (Bovo et al., 2013) a Hierarchical Clustering algorithm is used to predict failure in an online subject presented via a Learning Management System. In their study, (Durairaj and Vijitha, 2014) highlight the importance of higher education institutions being able to predict academic performance. (Romero et al., 2013) use association rules to improve the quality of courses and quizzes, concluding that updates to quizzes and courses impact student learning in a positive and significant way.

Educational institutions are data-rich environments. They collect and store a lot of data about their students and other stakeholders. From a data mining perspective, there is a need to leverage this data to obtain meaningful and useful information, and to generate new knowledge from it. For example, data mining techniques, in conjunction with different student attributes, can be used for academic achievement prediction, retention or at-risk status. Studies in which data mining has played a major role in higher education environments to obtain information about students include works by (Romero and Ventura, 2013), (Hegland, 2001), (Tan et al., 2006) and (Baker et al., 2010). Educational Data Mining is a relatively new discipline. There is therefore a need for more work to be done in this emerging field, in order to highlight some of the ways in which data mining techniques can be used to investigate the different topics and problems that present

themselves for meaningful study and analysis. This study, hopefully, should contribute meaningfully towards this effort.

Regarding the relevance and importance of Educational Data Mining, (Bienkowski et al., 2012) have the following to say:

> *"Educational data mining and learning analytics have the potential to make visible data that have heretofore gone unseen, unnoticed, and therefore unactionable."*

2.4 Data Mining for the Research Problem

Higher education institutions, both local and international, collect vast amounts of different types of data about their students. In other parts of the world great progress has been made to leverage this data to obtain useful information about students. In South Africa few higher education institutions, including ours, have only begun to use data mining techniques to extract information and obtain knowledge from the data they collect about their students. There is, therefore, a need in our context to intensify educational research efforts in which data mining techniques play a critical role in our understanding of students and their learning environments.

The University of the Witwatersrand collects vast amounts of data about students enrolled in the Faculty of Science, where this study is located. Within the faculty, the students belong to different schools, where they study for BSc. degrees in STEM disciplines. This study is about first year students enrolled in STEM disciplines in our faculty. Given the academic challenges faced by first time first year students, locally and internationally, and the lack of understanding and knowledge about these students, it is clear to us that through data mining, much information can be obtained about them, using the data the university and faculty collect about them. Given the dire shortage and huge demand for highly skilled professionals with STEM qualifications in the country, a data mining based study focusing on first time first year students STEM students in the faculty is more than justified.

Higher education institutions, including the University of Witwatersrand, are not producing a sufficient number of graduates with qualifications in critical-skill STEM disciplines. These graduates are desperately needed for their skills and expertise, which can be deployed to benefit different sectors of the economy and society. There are several reasons for this shortcoming. Not many first time first year students, for a number of reasons, choose a STEM-based discipline as their study programme of choice when

they enter university. Student retention is another problem. The problem of students dropping out or failing also affects first year students enrolled in STEM disciplines. Hence, fewer students proceed to successive years of study. As a consequence, even fewer students complete their studies and qualify to study further or become active in the economy and society by deploying their recently acquired critical skills. Whilst this study does not seek to provide answers to these and other problematic issues, its main focus is to obtain meaningful and useful information from the data belonging to first year students in our faculty. In this study we conducted different experiments to obtain information from the data we collected about first time first year STEM students in our faculty. Our study focused on the following topics:

- Gender, race, funding status and academic performance.
- English proficiency and academic performance.
- High school STEM subjects and the Admission Point Score (APS).
- Influence of prior computer programming experience.
- Wits minimum admission requirements and academic performance.
- First year STEM student characteristics and academic performance.
- The role of Mathematics I (Algebra I and Calculus I) in first year STEM curricula, and academic performance.

In order to address these issues, different data mining and statistical experiments were conducted. Hopefully, the results will provide useful insights and lead to a better understanding of first year students in our faculty.

2.5 Selected Data Mining and Statistical Techniques

In order to answer the research questions we posed earlier, we conducted different data mining and statistical analysis experiments. In this section we describe and discuss the different techniques we used to perform our experiments. This section comprises several subsections, each focusing on a data mining or statistical technique used in the experiments. Classification is discussed in Subsection 2.5.1. Predictive data mining is discussed in Subsection 2.5.2. A discussion on cluster analysis is provided in Subsection 2.5.3. This is followed by a discussion on association rule mining in Subsection 2.5.4. Finally, in Subsection 2.5.5 we discuss the statistical techniques we applied to analyze the different datasets we used in this study.

2.5.1 Classification

Classification is a frequently used data mining technique. It is a supervised learning approach in which objects (instances or records) are assigned to different classes or categories (Agarwal et al., 2012). A target or class variable of categorical type is used to classify or categorize a particular record. For example, in a dataset containing student records, a variable named *Result* can consist of two classes, namely Pass and Fail. Figure 2.1 shows a table of records that can be used for classification purposes. The records contain information about customers who have received car loans. The class variable, named *Default*, indicates whether a customer will default (Yes) or not default (No) on their loan.

Age	Income	Student	Credit Rating	Default
Youth	Medium	Yes	Fair	No
Youth	Low	Yes	Fair	No
Senior	Low	No	Excellent	No
Senior	Medium	No	Excellent	No
Senior	High	No	Poor	Yes
Senior	Medium	No	Poor	Yes
Senior	Low	Yes	Fair	No
Middle Age	Low	No	Fair	Yes
Middle Age	Medium	Yes	Fair	No
Middle Age	Low	No	Fair	Yes

FIGURE 2.1: Car Loan Classification Data ((Weiss and Davison, 2010))

Figure 2.1 shows that each record consists of five variables (attributes), including the class variable. The four input or predictor variables are named as follows: *Age, Income, Student* and *Credit-Rating*. Using data such as the records in Figure 2.1, a classification

model can be built by applying a data mining algorithm against the data. A set of records, such as those in Figure 2.1, which are used to create a model, are referred to as a *training set*. The classification task consists in using the model to classify records whose class labels are not known. Such records are referred to as a *test set*. To summarize the discussion on classification, (Han et al., 2011) state that the following tasks are performed during the classification process:

1. In the *model construction* step a model is created from a dataset in which the target variable in each record is assigned a class label.
2. In the *model usage* step the model created during the model construction step is used to classify unseen data records, namely records whose class labels are unknown.

In the following subsections, we discuss some of the methods we used to classify our dataset records. These include decision tree, classification rule, nearest neighbour and Bayesian classifiers.

Decision Tree Classification

We discuss in this subsection the classification method using a decision tree. The tree in Figure 2.2 is constructed using the data shown in Figure 2.1.

A decision tree contains three types of nodes: a *root node, internal nodes* and *leaf nodes*.

- The root node is located at the top of the decision tree. It has no input edges and may contain zero or more output edges. The root node in Figure 2.2 contains two outgoing edges.
- An internal node has a single input edge and two or more output edges. Like the root node, an internal node contains the name of an attribute. In Figure 2.2 the nodes containing the *Age* and *Income* attribute names are internal nodes.
- A leaf node represents a class label and has a single incoming edge and no output edges. In Figure 2.2 all the nodes containing *Default* attribute names are leaf nodes.

Once created, a decision tree such as the one in Figure 2.2 can be used to classify records with unknown class labels. The process begins at the root node, where the value

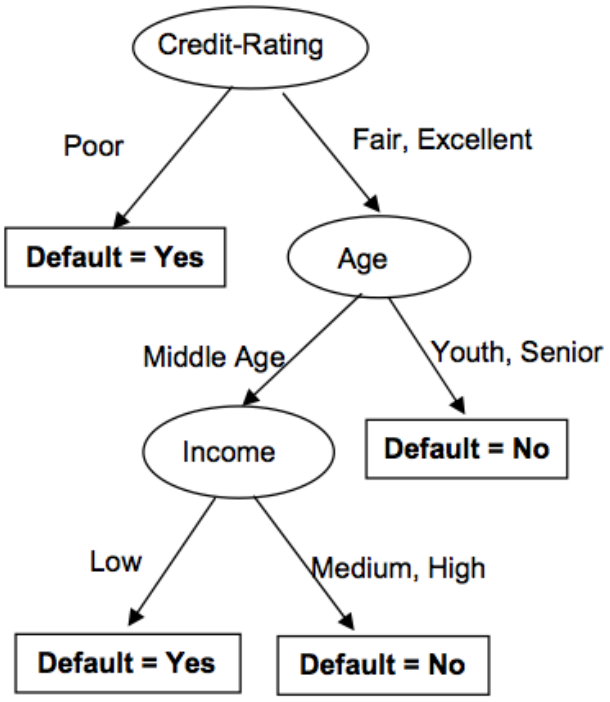


FIGURE 2.2: Car Loan Decision Tree ((Weiss and Davison, 2010))

of *Credit-Rating* Credit-Rating is tested. Based on its value, the following steps are performed:

1. If *Credit-Rating*=Poor, the decision is Yes. The customer will default if granted the car loan.
2. If *Credit-Rating*=Fair or *Credit-Rating*=Excellent, the value of *Age* is tested next.
    - If *Age*=Middle Age, then *Income* is tested next. If *Income*=Low, then the decision is Yes. Otherwise it is No.
    - *Age*=Youth or *Age*=Senior, the decision is No.

(Du, 2010) mentions the following about constructing a decision tree:

- From the same training set, it is possible to create different decision trees.
- The structure of a decision tree has a direct effect on the performance of a classification model. For a decision tree containing more levels, more tests must be conducted, which takes longer to perform the classification task.
- A classification model is considered to be more interesting if it achieves a high degree of accuracy, which increases confidence in the classification decision taken by the model.

There are several benefits associated with the decision tree technique. The output of the tree is easier to understand and interpret (Bramer, 2007). In addition, the technique is considered to be amongst the best when it comes to generating interesting classification rules (Agarwal et al., 2012).

Figure 2.3 shows some of the most widely used decision tree techniques, (Song and Ying, 2015).

Methods	CART	C4. 5	CHAID	QUEST
Measure used to select input variable	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square for categorical variables; J-way ANOVA for continuous/ordinal variables
Pruning	Pre-pruning using a single-pass algorithm	Pre-pruning using a single-pass algorithm	Pre-pruning using Chi-square test for independence	Post-pruning
Dependent variable	Categorical/ Continuous	Categorical/ Continuous	Categorica	Categorical
Input variables	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous
Split at each node	Binary; Split on linear combinations	Multiple	Multiple	Binary; Split on linear combinations

FIGURE 2.3: Decision Tree Algorithms ((Song and Ying, 2015))

In this study we used the J48 decision tree algorithm to conduct several data mining experiments. The J48 algorithm is developed from (Quinlan, 2014)'s C4.5 decision tree algorithm.

Rule-Based Classification

A rule-based classifier uses a set of "if ... Then ..." rules to classify test records. A training set is used to create a sequence of rules. A rule takes the form

$$Antecedent ==> Consequent \tag{2.1}$$

The *antecedent* (precondition) contains a conjunction of comparisons involving attributes and their corresponding values. The *consequent* is a single statement representing the predicted class. Formally, a rule is structured as follows::

$$(A_1\ op\ v_1) \wedge (A_2\ op\ v_2) \wedge ... (A_n\ op_n\ v_n)\ ==>\ class = y_j \tag{2.2}$$

where

- A_i is an attribute name,
- v_i is the value of an attribute, and
- op_n is a comparison (logical) operator belonging to the set $\{=, \neq, <, >, \leq, \geq\}$,
- y_j is a class label.

Figure 2.4 shows the classification rules generated by the WEKA-based JRIP classifier. JRIP is based on the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm proposed by (Cohen, 1995). The rules in this example can be used by the JRIP classifier to classify a record with an unknown class label.

```
JRIP rules:
===========

(APS <= 40) and (FUNDING_STATUS = F) => OUTCOME=Q1 (68.0/27.0)
(APS <= 48) and (MAT >= 80) and (APS <= 41) => OUTCOME=Q1 (47.0/19.0)
(PSC <= 66) and (FUNDING_STATUS = NF) and (ENA <= -1) => OUTCOME=Q1 (76.0/33.0)
 => OUTCOME=Q (1204.0/493.0)

Number of Rules : 4
```

FIGURE 2.4: Outcome classification Rule Set

A rule *covers* a given record if the record's attribute values match the rule's precondition. Whenever a rule covers a given record, it is said that the rule is *fired* or *triggered*. To illustrate the idea, suppose we want to classify the test records shown in Table 2.1, using the first rule (Rule 1) in Figure 2.4.

TABLE 2.1: Test records for Outcome Classification

Funding Status	**APS**	**MAT**	**PSC**	**ENA**	**ENB**
F	40.0	81.0	74.0	-1.0	69.0
NF	44.0	83.0	76.0	68.0	-1.0

- Rule 1 covers the first record since its precondition is satisfied by the record's attributes.
- Rule 1 does not cover the second record, which represents a non-funded (NF) student whose APS is greater than 40. This violates the precondition of Rule 1.

Metrics such as *coverage* and *accuracy* are often used to evaluate the quality of classification rules. Suppose D represents a training set and A the number of records covered by a rule, and B_c is the number of records with class label c, which are covered by the rule. For a rule r, we define its coverage and accuracy as follows:

$$Coverage(r) = \frac{A}{D} \tag{2.3}$$

$$Accuracy(r) = \frac{B_c}{A} \tag{2.4}$$

Looking at Figure 2.4 we notice that a classification model consists of several rules. To classify a test record, the rule triggered by the record is used to classify it. If the rules are *are ordered*, based on decreasing order of priority, a test record is assigned the class of the rule located at the top of the list. On the other hand, if the rules are *unordered* all the rules that match the test record are fired. The class of the record is then determined through a voting mechanism involving the predicted classes. A default rule is used to cater for training records which the existing rules do not cover. The default rule has no antecedent and is situated at the end of a rule list. For example the rule

$$==> \text{OUTCOME} = \text{Q}$$

in Figure 2.4 is a default rule, which is executed after all other rules have been executed and found to be false. In this study, we used the WEKA toolkit's JRIP classifier to classify the records of first year STEM students.

Nearest-Neighbour Classification

We refer to the decision tree and rule-based classifiers discussed earlier as *eager learners*. Given a training set of records, these classifiers first create a classification model, which is used thereafter to classify unseen records. In contrast to this approach, some classifiers do not immediately create a model from the training set. Instead, the model creation process is delayed until the last moment, when the model is presented with a test record to classify. Classifiers that adopt this strategy are known as *lazy learners*. To classify a test record, a distance measure is used to determine its proximity to the training set records stored in memory. The class of the training set record closest to the test record is used as the class of the test record. This approach to classification is known as *nearest-neighbour* classification.

In some cases, more than one nearest neighbour is used to classify a test record, by assigning it the majority class of the nearest k neighbours. This is the *k-nearest-neighbour* (kNN) classification method. The following pseudo-code algorithm summarizes the *kNN* algorithm for classifying a test record (Du, 2010).

```
algorithm k-Nearest-Neighbour(Tr: training set; k: integer; var: data record)
begin
   for each training example t in Tr do
     calculate distance (t,r) upon descriptive attributes
   endfor;
   Select into D k nearest neighbours of r according to their distances to r;
   r.Class := majority class in D
end;
```

The algorithm starts, inside the **for** loop, by computing the distance between the test record (r) and each of the training set records (t). Thereafter, the algorithm selects the nearest k neighbours of the test record. Using the majority voting principle, the algorithm ends by assigning the majority class of the nearest k neighbours as the class of the test record. In this study, the IBk nearest-neighbour algorithm was used to predict academic performance, with the Wits minimum admission requirement subjects being used as predictor attributes. The algorithm uses the Euclidean distance metric.

Bayesian Classification

In a number of situations, it is not possible to determine with certainty the class to which a record belongs. To deal with this uncertainty a classification approach is required which models the relationships between a set of attributes and a class variable.

The *Bayesian classifier* is suitable for this purpose. Bayes' theorem is the basis of Bayesian classification. Suppose that A and B are random variables, and that each has a value associated with a random event. We begin by defining the joint and conditional probabilities of A and B.

- The *joint probability* of A and B, namely $P(A{=}a,\ B{=}b)$ is the probability that variable A will take a as its value and that variable B will take b as its value.
- The *conditional probability* $P(B{=}b \mid A{=}a)$ is the probability that B will have value b given that A has taken the value a.

For the two variables, A and B, Equation 2.5 shows the relationship between their joint and conditional variables.

$$P(A, B) = P(B|A) * P(A) = P(A|B) * P(B) \tag{2.5}$$

From the last two expressions in Equation 2.5 we obtain Equation 2.6, a formula known as Bayes' theorem.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.6}$$

Bayes' theorem states that the posterior probability of B given A, namely $P(B|A)$, can be obtained using the prior probability of B, $P(B)$. The theorem can also be used for classification purposes. Let B represent a class attribute and $\mathbf{A}$ denote a set of descriptive attributes. $P(\mathbf{A}|B)$ is the **class-conditional** probability, $P(B)$ the prior probability of the class and $P(\mathbf{A})$ is the probability of the evidence record being classified. Using these probabilities and the evidence record, we express the posterior probability as follows:

$$P(B|\mathbf{A}) = \frac{P(\mathbf{A}|B)P(B)}{P(\mathbf{A})} \tag{2.7}$$

We want to predict the likelihood that a record $\mathbf{A}$ belongs to a certain class. For the different classes posterior probabilities ($P(b_1|\mathbf{A})$, $P(b_2|\mathbf{A})$, ..., $P(b_n|\mathbf{A})$) are computed and compared. Using the computed likelihood values, the class of a test record is determined. This is the class for which the likelihood value is the highest. In this study we used the Naïve Bayes classifier to predict the overall result of first year STEM students in our faculty.

2.5.2 Prediction

In data mining, prediction is used to predict a future state (Jindal and Borah, 2013). The type of data determines the prediction technique that should be used. Classification techniques are often used to predict categorical or discrete outcomes. Numeric prediction involves the prediction of continuous (or ordered) values. *Regression* is the most widely used approach for predicting numeric values. Regression analysis relates at least one independent variable to a single, continuous dependent variable. Regression analysis is used to predict the value of the dependent variable from one or more independent variables. When all the independent variables are continuous, regression analysis is preferred as the most suitable prediction technique. In most cases, the *linear regression* technique is used for prediction purposes. In this subsection, the following linear regression techniques are discussed: *straight-line* and *multiple* linear regression analysis.

Straight-Line Regression Analysis

In this type of analysis the response variable is predicted from a single predictor variable. In this, the simplest form of regression, the response variable (y) is modeled as a linear function of the predictor variable, x. Equation 2.8 shows the linear relationship between the two variables, x and y.

$$y = b_0 + b_1 x \tag{2.8}$$

where

- y is the predicted value of the response variable,
- b_0 is the *y-intercept* of the regression line,
- b_1 is the *slope* of the regression line,
- b_0 and b_1 are *regression coefficients*.

In data mining a model is created during the training stage from a set of records which make up the training set. The model is then applied against test records to predict the numerical values of the response variable.

Multiple Linear Regression Analysis

Multiple linear regression involves more than one predictor variable. In multiple linear regression a response variable, say y, is expressed in terms of two or more independent variables. Equation 2.9 shows the relationship between y and the two independent variables, x_1 and x_2.

$$y = w_0 + w_1 x_1 + w_2 x_2 \tag{2.9}$$

where

- y is the predicted value of the response variable,
- w_0 is the y-intercept of the regression line,
- w_1 and w_2 are slopes of the regression line,
- w_0, w_1 and w_2 are regression coefficients.

During the model building stage a coefficient is computed for each predictor variable in the model. As in straight-line (simple) linear regression, the coefficient indicates the influence of its related predictor on the response variable. In this study, we used the *LinearRegression* function in the WEKA toolkit to determine the overall academic achievement of Computer Science I students, using two first-semester Computer Science I courses.

2.5.3 Cluster Analysis

Clustering refers to the process in which objects are partitioned into groups, also known as *clusters*, consisting of objects that are similar. A cluster contains similar objects, which are not similar to objects located in other clusters (Agarwal et al., 2012), (Larose, 2014). Clustering methods use similarity metrics to determine the extent to which objects are similar or dissimilar to each other. Some of the measures generally used to determine object similarity include *Euclidean Distance, Manhattan Distance* and *Cosine Similarity*. Objects that are closer to each other are grouped together to form a cluster. The Euclidean Distance is the most popular similarity metric. It is computed as follows (Equation 2.10):

$$D(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2} \tag{2.10}$$

where

- $x_i = x_1, x_2, \ldots, x_k$ are the k attribute values of the first object
- $x_j = x_1, x_2, \ldots, x_k$ are the k attribute values of the second object

An example of object clustering is illustrated in Figure 2.5 (Weiss and Davison, 2010). There are three distinct clusters. Dashed lines are used to indicate the cluster boundaries.

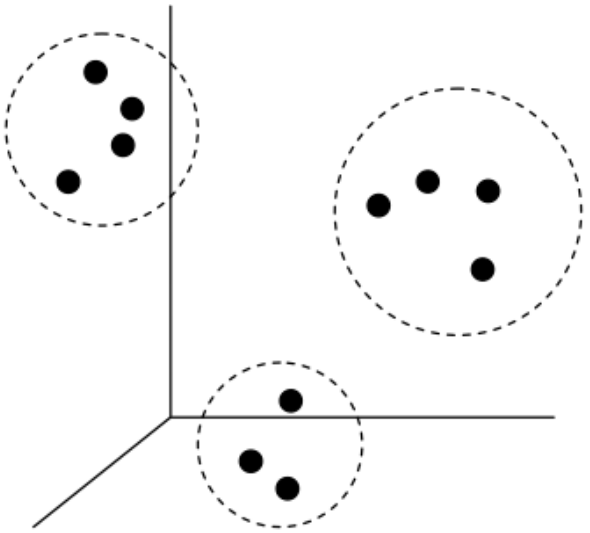


FIGURE 2.5: Object Clusters ((Weiss and Davison, 2010))

There are different types of clustering techniques. Examples include *hierarchical* and *non-hierarchical* (flat or partitioned) clustering techniques (Hegland, 2001). Hierarchical clustering creates multiple levels of clusters, organized in the form of a tree in which each cluster (node) is created from sub-clusters (child nodes). All the objects in the cluster are located at the root of the tree. In the case of non-hierarchical clustering, the entire dataset is partitioned into a single level. To create clusters hierarchical clustering algorithms use either *agglomerative* or *divisive* methods (Weiss and Davison, 2010). Initially, an agglomerative algorithm places each record or object in its own cluster. At each iteration, similar cluster pairs are merged to form a new cluster. At each step the number of clusters is decreased by one. The process ends when a stopping criterion is reached and a single large cluster has been formed. Divisive algorithms, on the other hand, start with all the records in a single cluster. Thereafter, dissimilar records are identified and placed into new clusters. The iterative splitting process stops when the terminating criterion is reached, at which point each record will have been placed in its own cluster.

Figure 2.6 shows one of several ways in which clustering techniques can be categorized.

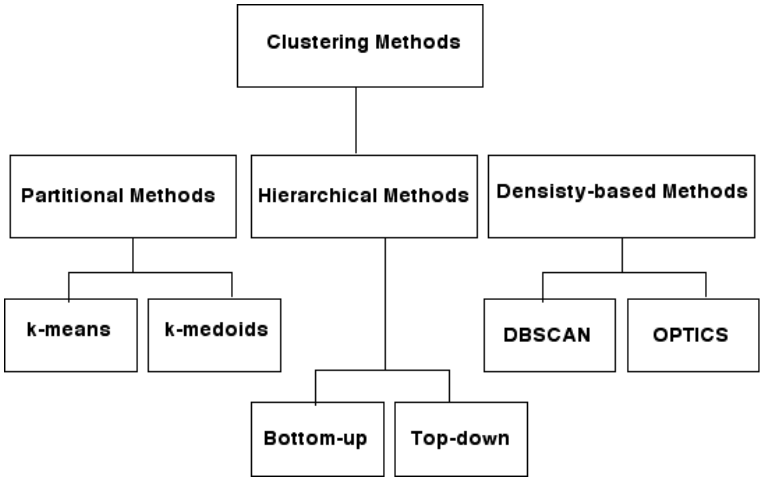


FIGURE 2.6: Types of Clustering Methods ((Chaudhari and Parikh, 2012))

One of the objectives of this study was to determine the role and influence of first year BSc. students' characteristics on their academic achievement. We found cluster analysis to be a suitable approach in achieving this objective. The *k-means* algorithm is a popular and frequently used non-hierarchical clustering algorithms. The algorithm has been used in several educational contexts. For example, to predict student results, (Bhise et al., 2013) applied the algorithm on a database of student records. (Ajith et al., 2013) grouped students according to different behaviors, using factors such as academic performance, punctuality and demographic characteristics. In this study, we used WEKA-based *k-means* algorithm to generate and analyze several clusters, based on the characteristics of first year STEM students in our faculty.

The following is a simplified description of the *k-means* algorithm we used in this study ((Tan et al., 2006), (Han et al., 2011)):

```
Algorithm: K-means
 1:Choose a value of k (the number of clusters the data will be partitioned into)
 2:Randomly select k records. Use them as cluster centroids
 3:Assign each of the records to the nearest cluster centroid
 4:Recalculate the centroids of the k clusters
 5:Repeat steps 3 and 4 until the centroids no longer move
```

In a number of fields cluster analysis is used in different applications to extract information about objects of interest. The fields include biology, where cluster analysis has been used to identify groups of genes that perform similar functions (Tan et al., 2006). In medicine, clustering can be used to identify groups of patients who show similar symptoms. In marketing, clustering can be used to identify clusters of customers with similar

purchasing behavior. Other fields where clustering techniques are frequently used include auditing, crime analysis and information retrieval. In Educational Data Mining, more work involving cluster analysis needs to be done, to leverage and extract useful information from the data higher education institutions collect about their students.

2.5.4 Association Rule Mining (ARM)

The purpose of association analysis is the discovery of significant patterns and relationships in datasets (Romero and Ventura, 2013). Association rule mining discovers frequently used data items by identifying patterns associated with related events (Bhambri, 2011). The discovered relationships are expressed using *association rules*. Generally, a rule is represented as follows:

$$X \Rightarrow Y$$

where X and Y are referred to as *itemsets*. An itemset is a non-empty set containing one or more items. The itemset X is the *antecedent* and the itemset Y is the *consequent* of the rule. In order to avoid extracting associations between the same items, a necessary constraint is that both X and Y should not contain common items (Du, 2010).

Suppose that D is a transaction database, in which T is the set of all transactions, where T= $\{t_1, t_2, ...,t_N\}$. To determine how strong an association rule is, the *support* (s) and *confidence* (c) measures are used. Given the rule $X \Rightarrow Y$, its support is the percentage of transactions containing X or Y. The support measure is defined in Equation 2.11, as follows:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{2.11}$$

where N is the total number of transactions in D . The confidence measure is the percentage of transactions containing items in X that also contain items in Y. This metric is defined in Equation 2.12, as follows:

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2.12}$$

The following steps are followed in obtaining rules from a given transaction database (Du, 2010), (Larose, 2014):

1. Find all itemsets from the transaction database whose *support* satisfies the *minimum support* threshold. These are called *frequent* itemsets.

2. From the frequent itemsets, generate association rules whose *confidence* meets the *minimum confidence* threshold requirement. These are called *strong* rules.

We used the *Apriori* algorithm to generate several rules and obtain information from the historical dataset of first time first year STEM students in our faculty. The algorithm uses the two steps described earlier to generate frequent itemsets and to extract high-confidence rules from them.

2.5.5 Statistical Analysis

This study used different statistical analysis techniques to analyse some of the first year STEM datasets obtained from the AISU. In this study we used the following statistical analysis techniques in the different experiments we performed to analyze our datasets: correlation analysis, independent samples T-test and analysis of variance.

Correlation Analysis

Correlation analysis is one of several statistical analysis techniques that is used to explore relationships among data variables (Zou et al., 2003). Correlation analysis is often used to address questions involving relationships between two variables (e.g., x and y). For example, we might want to know whether the Admission Point Score (y) varies as a function of the NSC Grade 12 Mathematics score (x). The degree of the relationship between two variables, say x and y, is determined by the *correlation coefficient* (r). Its value ranges between -1 and 1.

The *Pearson product-moment correlation coefficient* (r) is a frequently used measure. Equation 2.13 shows how this coefficient is computed between x and y.

$$r_{xy} = \frac{\sum_i ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2.13}$$

where

- $\bar{x}$ is the sample mean of variable x,
- $\bar{y}$ is the sample mean of variable y and
- $i = 1, 2, \ldots, n$.

The contents of Table 2.2 can be used to interpret the result of a correlation analysis experiment.

TABLE 2.2: Meaning of Correlation Coefficient Values (Dancey and Reidy, 2008)

Coefficient (r)	**Strength**
0	Zero
$<$ 0.1 - 0.3	Weak
0.4 - 0.6	Moderate
0.7 - 0.9	Strong

Independent Samples T-test

The T-test is a frequently used statistical analysis test for comparing samples (Rojewski et al., 2012), (Kim, 2015). The test compares the means of two independent samples. One of the reasons for choosing to use the T-test in this study is its robustness to violations of some of its assumptions. For example, the test can be used to analyse skewed data, despite the assumption that the data should be normally distributed. In one of our experiments, the test was used to analyse the NSC grade 12 records belonging to two groups of first year STEM students. The aim was to determine whether the Admission Point Score obtained by English Home Language (ENA) students differed significantly from that obtained by English First Additional Language (ENB) students.

A T-test compares the means of two populations. The mean values are used to test two hypotheses, namely the *Null Hypothesis* (H_0) and *Alternative Hypothesis* (H_1), which are stated as follows:

H_0: *The sample means do not differ from each other.*
H_1: *The sample means differ from each other.*

When conducting a two-sample T-test, two test statistics (t-statistics) are calculated. In the first calculation the variances of the groups being compared are assumed to be equal. In the second calculation the variances are assumed not to be equal. When testing whether the sample variances are equal and the p-value obtained is less than the significance level of 0.05, the hypothesis that the variances are equal is rejected. Instead, the test statistic regarding unequal variances is used. For a large p-value the t-statistic for equal variances is used.

To conduct a statistical test, the following general steps are followed:

1. The null and alternative hypotheses are stated.
2. Test the null hypothesis by performing the relevant statistical test.
3. Interpret the result of the test and, using a decision criterion, decide whether the null hypothesis should be accepted or rejected.

Analysis of Variance (ANOVA)

Similar to the T-test, the ANOVA technique is also used to compare sample means. However, unlike the T-test, which compares only two means, ANOVA can be used to determine the difference between two or more means (Rojewski et al., 2012). We differentiate between two types of ANOVA, namely *One-Way ANOVA* and *factorial ANOVA*. In one-way ANOVA only one categorical independent variable is used, which consists of two or more levels. There is only one numerical dependent variable. The one-way ANOVA technique is based on the following assumptions: the sample populations are normally distributed, the populations have the same variance and the observations are independent.

Whereas in one-way ANOVA a single independent variable is used, factorial ANOVA involves two or more independent variables. In two-way ANOVA the relationship involves a dependent variable and two independent variables.

There are two ways of labeling factorial designs. They can be labeled using the number of factors involved. If two factors are used, we refer to the design as a *two-way factorial* design, and with three factors we have a *three-way factorial* design, and so on. Another method is to label a factorial using the number of levels of each factor. For example, suppose there is a study involving two factors, Gender and Race. The Gender factor has two levels (Male, Female) and the Race factor has five levels (African, Chinese, Coloured, Indian, White). Since there are two levels of Gender and five levels of Race, the design would be referred to as 2x5 factorial design. In this study, two-way ANOVA factorial designs were used to investigate the following:

1. The relationship between the quantitative Admission Point Score (APS) and two qualitative variables, namely Gender and Race, and
2. The relationship between the quantitative Admission Point Score (APS) and two qualitative variables, namely Funding Status and Race.

Figure 2.7 summarizes the process involved in deciding which statistical test to perform based on the following information: number of samples, whether samples are dependent or independent, and the nature of the variance (equal or unequal).

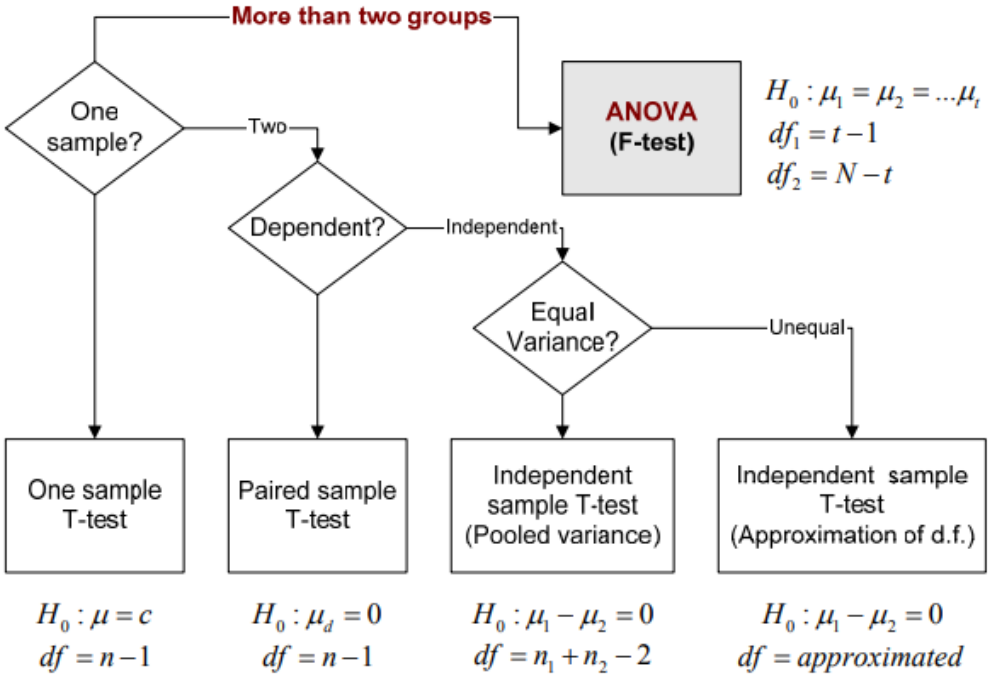


FIGURE 2.7: Sample T-test and ANOVA (Park, 2009)

In this study we used the following statistical analysis tests: independent sample T-test (unequal variance) and two-way ANOVA.

2.6 Related Work

This section provides a detailed review and discussion of the literature related to this study. It was deemed more convenient to organize the literature into different topics (themes), each focusing on the specific question this study sought to answer. These topics are discussed in different subsections, as follows. In Subsection 2.6.1, we discuss the literature on students' ethnic background and gender, and the relationship of these two factors to academic performance. The review includes studies focusing on the ethnic and gender characteristics of students enrolled in the Science, Technology, Engineering and Mathematics (STEM) disciplines. The majority of universities in South Africa use English as the language of teaching. This situation seemingly places students whose home language is English at a certain academic advantage. On the other hand, for most students, English is their second language. Some of them, due to their lack of proficiency in English, may find it difficult to cope in their studies. To look in some depth at the issue regarding English as the language of teaching, the review and discussion of the literature in Subsection 2.6.2 focuses on the connection between English language proficiency and academic achievement.

Given the influence and contribution of course selection on academic success or failure, in Subsection 2.6.3, the literature on this important topic is discussed. We also look at studies focusing primarily on students' selection of STEM courses and how these choices relate to academic performance. In Subsection 2.6.4, there is a discussion on the literature on the prior computer programming experience of pre-university students and whether such experience offers any academic benefits for students enrolled in computer-related disciplines, where computer programming plays a pivotal role. Subsection 2.6.5 reviews the literature on pre-university admission criteria and their possible impact on academic success or failure. Student funding provides access to higher education to deserving and capable students from deprived economic and social backgrounds. The review of the literature in Subsection 2.6.6 focuses on student funding and its relationship with other factors, such as academic performance. Finally, Subsection 2.6.7 discusses the literature on Mathematics and its impact on academic achievement in STEM-based disciplines.

2.6.1 Ethnic Background, Gender and Academic Performance

In a number of studies, students' ethnic background and gender have been investigated to determine, among other things, their effect on academic performance. For example (Crisp et al., 2009)'s study, conducted at an Hispanic Serving Institution (HSI) in America, investigated the impact of demographic and other factors in undergraduate students' decision to obtain a STEM qualification. The study involved 1925 Hispanic and White students who obtained degrees during the period 2006-2008. Several differences, some significant, were found between Hispanic and White students. Amongst some of the findings, Hispanic students, compared to their White counterparts, obtained lower Scholastic Assessment Test (SAT) scores in Mathematics. They also obtained lower Grade Point Average (GPA) scores in their first semester of college.

In (Crisp et al., 2009)'s study, female students were not as likely as their male counterparts to declare a STEM major, convert to a STEM major, and obtain a STEM degree. In (Thiele et al., 2016)'s study, the purpose was to investigate academic and socio-demographic factors, to determine their influence as predictors of academic performance and degree attainment. The study involved 5369 records belonging to full-time students who had successfully completed their three-year degree programmes. Factors of interest included gender, ethnicity and academic performance. Similar to (Crisp et al., 2009)'s findings, students from minority ethnic groups, in this case Black and Asian students, were outperformed by White students. White students outperformed their Black and Asian peers, who were most likely to perform at the lower second (2.2), or even a lower degree category. Apart from entering university having the lowest points Black students, after Asian students, obtained the lowest average scores.

In their review of several studies, (Miller-Cotto and Byrnes, 2016) investigated ethnic racial identity (ERI) and its association with academic achievement. There is a different view amongst researchers regarding the nature of this relationship. Some believe that the correlation between ERI and academic achievement should be positive, whereas others expect this relationship to be negative. Yet another view is that this relationship depends on the ethnic background and the context in which the participants find themselves. (Miller-Cotto and Byrnes, 2016) found the correlation between ERI and academic achievement to be small, positive and significant.

The longitudinal study by (Frederickson and Petrides, 2008) examined the academic achievement of secondary school students from different ethnic and gender backgrounds. The study analyzed questionnaires belonging to 517 male and female students, mostly from Black, White and Pakistani backgrounds. The students came from urban, suburban and rural areas and were selected from grammar and modern secondary schools. The following factors were investigated: ethnicity, socio-economic status (SES), Verbal

Reasoning Test (VRT), Key Stage 3 (KS3) and General Certificate of Secondary Education (GCSE) results. Regarding the KS3 factor, boys significantly outperformed girls in Mathematics and Science. At the GCSE level, girls significantly outperformed boys in English. Regarding the socio-economic status factor, high SES pupils significantly outperformed low SES pupils in all respects (VRT, KS3 and GCSE). The Analysis of Variance (ANOVA) results showed White pupils to be performing significantly better in all respects (VRT, KS3 and GCSE), compared to their Black and Pakistani counterparts from poor backgrounds, who enter university having lower GCSE scores. Similarly in (Thiele et al., 2016) students from low socio-economic backgrounds, mostly Black, entered university having the lowest points.

In their study (Sommerville and Singaram, 2018) investigated the magnitude and extent to which non-cognitive demographic factors influence medical students' academic performance. This mixed methods study involved a cohort of 202 students whose academic progress was tracked over a period of five years, starting from year one. The study used student marks collected over a five-year period. Interviews were also conducted involving some lecturers and third year students. The analysis of student marks involved such factors as race, language, sex, age, and financial support. The marks were analyzed using the following statistical analysis tests: T-test, ANOVA and regression analysis. There was significant difference in performance amongst Black, Coloured, Indian and White students, particularly between Black and Indian students. According to (Sommerville and Singaram, 2018), the differences could be due to the unequal distribution of funding towards primary and secondary schools attended by the different ethnic groups. Similarly, the study by (Thiele et al., 2016) found students' ethnic background to be significantly related to academic performance.

On the question of gender, (Sommerville and Singaram, 2018) found an insignificant gender difference in terms of academic performance. However, according to the academics who were interviewed in this study, their view was that female students were more diligent, emotionally mature, organized and less risky. Contrary to (Sommerville and Singaram, 2018)'s finding, (Thiele et al., 2016) found a significant difference between gender and academic achievement, both at high school and university levels. (Sommerville and Singaram, 2018) also found that language and financial aid had a significant influence on academic performance. English first language speakers obtained better marks compared to students from other language groups. Regarding financial aid status, students on merit bursaries significantly outperformed non-funded and NSFAS-funded students, the latter also being significantly outperformed by non-funded students.

2.6.2 English Language Proficiency and Academic Performance

In South African tertiary institutions, English is used as the language of instruction. However, most of the students who constitute the majority in our universities are not English first language speakers. They come from different language backgrounds. Their retention and success rates may be hampered by their lack of English language proficiency, which is necessary for understanding lecture content and reading material. Numerous studies have investigated the connection between English as a teaching language and academic achievement.

In (Manson, 2015), the purpose was to investigate the English language proficiency of first year nursing students and their academic achievement. The study investigated several issues, such as factors influencing English proficiency, students' self-opinion of their English proficiency, the relationship between high school language results and academic achievement, as well as the success rate at the end of the first year of study. Statistical and regression analysis techniques were used to perform significance tests and to predict academic performance. Positive and significant relationships were discovered between matriculation English and some of the factors under study. For example, attending a rural primary school was significantly related to performance in high school English. Compared to students from urban primary schools, attending a rural primary school was associated with significantly lower high school English scores. Having access to a computer resulted in a significantly better performance in English compared to having no access to a computer. Students from a more privileged socio-economic background obtained higher matriculation English scores. Perhaps not surprisingly, the matriculation English marks obtained by English home language speakers were significantly higher compared to those obtained by students from other language backgrounds. Nursing students who entered university with higher matriculation English scores achieved better first year results. In a similar study (Sommerville and Singaram, 2018), English first-language students performed significantly better in tests, conducted over a five-year period, than English second-language students.

(Dlomo et al., 2011)'s study examined several factors to determine their influence on first year students' achievement in an economics course. Several important findings emerged from (Dlomo et al., 2011)'s study. Students who obtained high scores (A or B) in the high school English Home Language subject also performed better in their first year economics examination. Similarly, in (Manson, 2015) found that success amongst first year nursing students was associated with high matriculation English scores.

Contrary to (Dlomo et al., 2011) and (Manson, 2015)'s findings, (Van Rooy and Coetzee-Van Rooy, 2015) did not find matriculation language scores to be adequate predictors of success amongst university students. Even though positive and significant correlations were found between the NSC's English First Additional Language (EFAL) subject and

average performance in the first, second and third years, these correlations were very weak, leading the authors to conclude that performance in EFAL could not be relied upon to predict overall success at university. Instead, the academic literacy support modules taught during the first and second semesters were better predictors of success in the first and second years of study.

In line with studies conducted by (Manson, 2015), (Dlomo et al., 2011) and (Van Rooy and Coetzee-Van Rooy, 2015), the study by (Marnewick, 2012) investigated the association between high school marks and the marks obtained in the first year modules of Information and Communication Technology (ICT) students. Regarding the association between the English score and the scores obtained in first year modules, only two out of 15 modules correlated positively and significantly with English. Given the weakness of these correlations, (Marnewick, 2012) concluded that the results were not significant enough to consider using high school English as a predictor of academic achievement in first year ICT modules. Similarly, from their results (Van Rooy and Coetzee-Van Rooy, 2015) concluded that high school English scores were not reliable predictors of academic performance at tertiary level.

The study by (Schoer et al., 2015) used the English Home Language (EHL) and English First Additional Language (EFAL) scores of 759 first year Bachelor of Education (BEd) students to determine, among other things, their academic language competence. Several findings emerged from this study. Despite the fact that more EFAL students had low Admission Point Scores (APS) on entry, compared to EHL students, the NSC English scores of the two groups were not significantly different. According to (Schoer et al., 2015), this finding does not mean that between the two groups, significant differences in language competencies do not exist. In their opinion the NBT AL component seems, for a number of reasons, to be a better instrument for identifying and determining students' language competency problems.

2.6.3 Course Selection

Academic performance in high school STEM subjects can serve as a good indicator of academic performance in STEM-based disciplines at university. For high school students to enhance their chances of success at university, it is important for them to select "potentially successful" STEM subjects in their high school curriculum. For students to experience a much easier transition towards successful studies in STEM disciplines at university, in high school they need to select subjects which are most likely to ensure their success in first year STEM disciplines. Different techniques such as feature selection and multiple linear regression can be used to identify STEM subjects which contribute the most towards academic performance. The purpose of feature selection is to select relevant attributes by removing redundant or irrelevant features from a set of input features or attributes.

There are several studies in which feature selection methods have been used. For instance in their study, which was to was to determine the effectiveness of different machine learning algorithms in predicting academic performance, (Velmurugan and Anuradha, 2016) used different feature selection techniques. Their educational dataset contained 21 attributes, which included gender, high school grades, medium of instruction, parents' educational qualifications and field of study. Experiments were conducted using six classification algorithms (J48, NaïvëBayes, BayesNet, IBK, OneR and JRip) and four feature selection techniques (CfsSubset, Chi-Squared, Information Gain and Relief). There were several findings from (Velmurugan and Anuradha, 2016)'s study. The average F-measure, computed using the F-measures of the six the data mining algorithms, was the highest for the CfsSubset feature selection technique, with seven attributes being used. When all 21 attributes were used, in the absence of feature selection, the average F-Measure was the lowest. In this case the algorithms were less effective in predicting academic performance compared to when feature selection techniques were incorporated. Overall, the CfsSubset technique outperformed the other three techniques. The IBK algorithm, combined with the CfsSubset feature selection technique, performed better than all the other algorithms, achieving the highest accuracy of 99.68%.

A study similar to (Velmurugan and Anuradha, 2016)'s was conducted by (Kavitha et al., 2012), who used six feature selection techniques and three machine learning algorithms to determine the algorithms' classification accuracy. The study used 34 features, obtained from a questionnaire, to predict student performance. As in (Velmurugan and Anuradha, 2016), the F1-Measure was used by (Kavitha et al., 2012) to determine the effectiveness of the feature selection algorithms in selecting the best feature subsets. Using 14 attributes, the Relief algorithm outperformed the other feature selection techniques by obtaining the highest F1 value. (Punlumjeak and Rachburee, 2015) used 6882 first year student records in their study, to predict academic performance. The study

used four feature selection techniques and four data mining algorithms. From the feature selection experiments, 10 relevant features were identified, and were thereafter used with each of the four data mining algorithms to predict academic achievement amongst first year engineering students. The Minimum Redundancy and Maximum Relevance feature selection method, used in combination with the k-Nearest neighbour algorithm, yielded the highest accuracy result. Based on their findings, (Punlumjeak and Rachburee, 2015) are of the view that their model can be used by students and stakeholders for academic planning and management purposes. Some of the research where feature selection methods have been used in predicting students' academic achievement include (Rahman et al., 2017) and (Zaffar et al., 2018).

Not many studies have used the multiple linear regression (MLR) technique to address the feature selection problem. In a recent study, (Örkcü, 2013) combined genetic algorithm (GA) and simulated annealing (SA) techniques to show that an alternative feature selection technique can be used in MLR analysis to obtain a minimal subset of relevant dataset features. Using real and simulation data, (Örkcü, 2013)'s hybrid feature selection technique outperformed traditional feature selection techniques. In their study (Yang et al., 2018) the MLR method was used to investigate its suitability in predicting academic achievement by first year students in a calculus course. The study's results highlighted the benefit of combining the MLR and Principal Component Analysis techniques. From the regression model's goodness-of-fit and residual analysis results, (Yang et al., 2018) concluded that MLR could be used as an indicator of students' academic achievement. The study also found the predictive performance of MLR and PCA together to be higher compared to MLR being used on its own.

In their study, (El Aissaoui et al., 2019) created a model from MLR and feature selection methods, which they used for predicting student performance. From the variables that were used in this study, the MLR model obtained using the MARS feature selection method, some of the predictor variables that were identified as having the most influence on the students' final grade include the mother's education level, age and romantic involvement.

Apart from the need to know how students will perform in their chosen courses, it may also be necessary to look at some of the factors students consider to make their subject choices. (Babad, 2001) used a questionnaire to obtain student responses regarding their choice of elective courses. The results from a random sample of 655 upper undergraduates from mixed departments showed that in making their first course selections, students considered several factors such as the standard of the course, the standard of teaching, the potential for acquiring knowledge, and so on. On the other hand, other courses were selected because they were comfortable and easy. In another study, (Palmer et al., 2017) surveyed students to learn about some of the factors they took into account

when selecting their subjects. In making their subject choices, the most important factors considered by the students include increased enjoyment, interest and awareness of their academic abilities in the sciences.

Course selection, particularly in STEM disciplines, is also impacted by other factors, such as student gender. In order for our society and the economy to meet current and future challenges, STEM-related disciplines and professions are essential. Currently, there is low participation amongst females, locally and globally, in STEM-related disciplines and professions (Piatek-Jimenez et al., 2018), (He et al., 2020). Several studies have looked into some of the factors responsible for female under-representation in STEM disciplines. For instance, (Wang and Degol, 2017) highlighted several factors to explain female under-representation in STEM disciplines. These include values based on lifestyle, occupational choices and gender-based opinions. (Bøe et al., 2011) investigated the relationship and participation of young learners in STEM-based courses. Similar to (Wang and Degol, 2017)'s finding, female students in (Bøe et al., 2011)'s study were under-represented in university STEM subjects. However, their numbers were much higher in the life sciences and health sciences. Amongst some of the issues raised by (Bøe et al., 2011), was the interest shown by students in STEM subjects. Interest in science and mathematics was somewhat low and declined as the students progressed through school. In order to improve student participation in STEM subjects, the authors suggest, among other things, the following: increasing interest and enjoyment in STEM subjects, teaching and presenting STEM subjects and studies in a way that acknowledges students' identities, having study group and organized homework sessions to increase, in particular and young female learners' expectations of positive outcomes.

(Kaleva et al., 2019)'s study raises issues similar to (Bøe et al., 2011) and (Wang and Degol, 2017) to explain low participation by female students in STEM disciplines, regarding in this instance the choice between basic and advanced mathematics. The reasons advanced by female students for not selecting advanced mathematics include lack of ability and not being mathematically competent.

In another study, (Holmes et al., 2018) investigated factors associated with predicting students' affinity for careers in STEM fields. Several significant factors were identified by the study, amongst which are the following: belonging to older student groups, being highly cultured, having a previous record of excellence in reading and being good with numbers. The study conducted by (Chen, 2013) over a period of six years (2003-2009) was to investigate attrition from STEM disciplines by a cohort of students whose bachelor degree studies began during the 2003-2004 academic year. Using a multinomial probit (MNP) model, the study looked at how various factors were linked to attrition from STEM disciplines. The author identified some of the factors responsible for increasing the probability of switching from STEM to non-STEM fields. These include taking

easier first year Mathematics courses and achieving poor results in STEM-related classes compared to classes not involving STEM courses.

2.6.4 Prior Programming Experience

In our faculty first time first year students are able to choose their field of study from 14 options of the BSc. degree offered by the university. The following degree options are available to students wishing to study Computer Science as a major subject: Bachelor of Science in the field of Applied Computing and Bachelor of Science in the field of Computing and Mathematics. Some of the subjects offered in these fields of study involve extensive computer programming. Amongst the first year students who enroll in these two fields of study, some will have learned how to write computer programs at high school, by taking the NSC Information Technology (IT) subject. On the other hand, some students will enroll in these two fields without having learned how to write computer programs at high school. An interesting question therefore arises. Do first time first year students with computer programming skills and experience achieve significantly better results in first year Computer Science courses compared to their counterparts, who lack prior computer programming experience?

Amongst educational researchers, there is a keen interest to determine whether a relationship exists between computer programming experience and academic achievement in computer-related disciplines. There are a number of studies whose primary focus is to investigate this relationship. (Byrne and Lyons, 2001) investigated the relationship between students' predictor attributes and achievement in the Programming and Logical Methods first year course. Compared to male students, female students obtained a higher average score in the programming course. However, the difference was not significant. In (Wilcox and Lionelle, 2018), female students with prior programming experience outperformed their male counterparts. The result was, however, not significant. The Science and Mathematics subjects were significantly correlated to the programming examination score. The correlation between the programming examination score and the foreign language and English subjects was not significant.

(Ayalew et al., 2018)'s study investigated several factors and their association with academic achievement in a first year programming subject. Several findings emerged from this study. The correlation between the mathematics and physics high school subjects on the one hand, and the introductory programming course, on the other, was significant. The high school chemistry and biology subjects did not correlate significantly with the programming subjects. There was a significant gender difference amongst male and female students in the programming course. Male students outperformed their female counterparts. On the contrary, in the earlier study by (Byrne and Lyons, 2001),

female students achieved better results in the programming course, compared to male students. However, the difference was not significant, a finding which (Bennedsen and Caspersen, 2007) also made in their study.

It is expected that students with prior programming experience should perform better at tertiary level in computer-related courses compared to their peers, who have no programming experience. Contrary to expectation, (Ayalew et al., 2018) did not find performance to be significant in the programming course between students with high school programming experience and those lacking such experience. In (Barlow-Jones et al., 2014) the Grade 12 computer scores were used to determine their association with the results obtained in a programming subject. The study involved students in two South African universities. The Grade 12 Computer Applications (CAT) scores were moderately correlated to the final Software Development module scores. (Kori et al., 2016) raise the important issue of the high dropout rate prevalent among higher education Information and Communication Technology (ICT) students. These authors note the fact that the low motivation amongst students could be the reason for them dropping out of their studies in large numbers. According to (Kori et al., 2016), prior experience in the ICT discipline may have an influence on higher education students' motivation and their academic results. In this study the weighted average grades of students with previous programming skills differed significantly from the grades obtained by students who lacked these skills. Students with programming skills considered their studies at university to be easier compared to students who had not programmed before. The two groups differed significantly in this regard.

Whilst prior computer programming skills may be associated with success in a first year programming subject, some researchers are also interested in students' performance beyond the first year. In (Holden and Weeden, 2003) students with prior programming experience, obtained in first year, outperformed their inexperienced counterparts, by obtaining a higher Grade Point Average (GPA) score. However, in a subsequent programming course, there was no difference between the two groups of students. They were performing at a similar level, obtaining the same average grades. In a similar study, (Wilcox and Lionelle, 2018) investigated whether prior programming experience provided any advantage towards student achievement in an introductory computer-related course (CSI). Overall, experienced students performed significantly better in the programming course compared to their inexperienced counterparts. In a subsequent Computer Science course (CS2) the overall performance between the two categories of students was similar indicating, as in (Wilcox and Lionelle, 2018), that in subsequent Computer Science courses the performance is no longer different between the two groups. Contrary to the findings in (Holden and Weeden, 2003) and (Wilcox and Lionelle, 2018), (Tafliovich et al., 2013) found that the performance gap in CS2 did not decrease, but increased between

students with prior programming skills and their peers, who lacked prior programming experience.

2.6.5 Pre-University Academic Performance

Students who wish to enroll at the Wits Faculty of Science for the first time, she or he must satisfy the minimum requirements for acceptance to their chosen fields of study. For example, a student wishing to enroll at the Wits Faculty of Science for the first time, she or he must meet the minimum criteria for acceptance to the BSc. degree option of their choice. To be accepted for a BSc. degree in the field of Actuarial Science, for example, a student in the faculty must obtain at least 80% in each of the following National Senior Certificate (NSC) subjects: Mathematics, Physical Sciences and English. An Admission Point Score (APS) of at least 40 points is also required. Each one of the 14 BSc. degree options in our faculty has specific minimum requirements for admission. Most universities, locally and globally, follow a similar practice, requiring first time first year students to satisfy certain admission criteria. An important question therefore arises. Do the minimum requirements for admission to a particular field of study in our faculty have any influence on the final result of first time first yea STEM students?

Numerous studies have focused on the contribution and relationship of first year admission criteria to academic performance. The study by (Jacobs et al., 2015) sought to predict the first semester achievement of 479 first year Life Sciences students. The study used three NSC Grade 12 STEM subjects and the NSC Grade 12 English subject. The relationship between the four NSC subjects and the first year Life Sciences subject was significant.

The National Benchmark Test (NBT) was introduced in South Africa to obtain a better understanding of senior high school students' academic abilities and preparedness for university (Wilson-Strydom, 2012). Universities use NSC results to select and place qualifying students into certificate, diploma and degree programmes. Some studies have looked at whether the NBT, similar to the NSC, could be used for the purpose of placing students in their first year degree. The study by (Van Walbeek et al., 2012) compared the NBT and the NSC to determine which, between the two examinations, could best predict academic success at university. Whilst the NSC results were found to be useful in predicting first year performance, the same cannot be said about the NBT. The study's findings showed that unlike the NSC, the NBT could not be used to place prospective first year students at university. Rather, (Van Walbeek et al., 2012) felt that the NBT could be used to complement the NSC, for the admission and correct placement of low performing applicants.

(Venkatas et al., 2014) conducted an exploratory study to determine the relationship involving Grade 12 NSC results (Mathematics, Physical Sciences and Life Sciences) and academic achievement in the first year, Bachelor of Optometry (BOptom) degree. Between the NSC scores and academic achievement in first year, there was no correlation, leading (Venkatas et al., 2014) to conclude that the NSC results were not suited to the task of predicting the academic success of first year students in the optometry degree programme. In contrast with (Venkatas et al., 2014)'s study, (Van Walbeek et al., 2012) on the other hand found that the same NSC subjects used by (Venkatas et al., 2014) could be used as first year academic performance predictors.

In another study at a South African university, (Dlomo et al., 2011) conducted a survey among first year Economics students, to determine their success rate, based on their language background and work status. The results showed that part-time working students and second language English speaking students were outperformed by first language English speaking students and students who spent more time studying outside lectures. Based on these findings, (Dlomo et al., 2011) recommend that universities adjust their admission requirements, including providing additional academic support to vulnerable students.

Several factors were investigated in (Blanchet, 2016)'s study to determine their influence in predicting success among American college students. Demographic factors included gender, race and socioeconomic status. Non-cognitive factors were also used, which included, amongst others, academic engagement. The high school aggregate score was the strongest predictor of achievement in the first semester, and retention to the second semester. The selection tests were not suitable predictors of college academic achievement. Similarly, in the South African context, (Van Walbeek et al., 2012) found the NBT to be an unsuitable selection test for entering first year university students. As in (Van Walbeek et al., 2012), the results obtained by (Van Rooy and Coetzee-Van Rooy, 2015) at a South African university showed that Grade 12 NSC scores were suitable for predicting successful academic achievement amongst first year students. However, this result applied to cases involving low risk, high performing students. Where low performing, high risk students were concerned, their school marks had little predictive value, and could not be considered as a suitable criterion for student admission and placement in first year.

2.6.6 Student Characteristics and Academic Performance

A number of studies have investigated student characteristics to determine, amongst others, their effect on academic performance. Student characteristics include funding status, age, gender, race and socio-economic status. Given the vast amounts of money governments and other agencies spend on funding students through their studies, it is also necessary in our context to take a closer look at the link between student funding and student outcomes. The purpose of (Naidoo and McKay, 2018)'s exploratory study was to discover the connection between student funding and academic achievement. The students' academic progress was followed for a duration of three years. Except for students who received academic merit awards, student funding and academic achievement (graduation) were not correlated. Based on their findings, (Naidoo and McKay, 2018) recommend the need for further research to determine the type of first year student who should be funded. The current funding model, which depends solely on the matriculation results to decide whether or not funding should be provided, is simply not adequate.

Student retention is a topic of interest in many studies, such (Carpenter, 2013)'s, which sought to determine the correlation between student funding and retention from year one to year two. Students' characteristics included the admission ACT score, gender, race and expected family contribution. Several findings emerged from (Carpenter, 2013)'s study. None of the different types of funding (equity, not equity and not awarded) were significantly different in terms of their effect on student retention. The study also showed that the time at which a financial aid application is submitted does not significantly affect progression from the freshman to the sophomore year.

Other characteristics that have been found to have been discovered to influence academic achievement include parents' socioeconomic status and education. In (Azhar et al., 2014), a questionnaire was administered to Masters students. The results showed that students from a strong financial background perform better, compared to their counterparts from a weak financial background. The study also revealed that parental education was a contributory factor towards academic performance. In their study (Adeyemo and Kuyoro, 2013) used students' attributes and previous academic achievement to achieve several objectives. One of these was to identify students' background factors which contribute towards academic achievement in first year. Four out of the 14 attributes used in the study were most influential in predicting academic performance in first year. These were the Joint Admission and Matriculation Board (JAMB) score, age on entering first year and parents' occupation. In their study (Erdogdu and Erdogdu, 2015) made several findings, which include the fact that academic performance is influenced by students' individual attributes and their home background, as well as the fact that the school attended and its location do not have any influence on academic performance. In contrast with (Erdogdu and Erdogdu, 2015), who found that the type of school attended had no

influence on academic performance, (Naidoo et al., 2014) found that students performed in a significantly different manner, depending on the type of school they attended. As in (Azhar et al., 2014), (Erdogdu and Erdogdu, 2015) also found a significant relationship between parental education and academic performance.

In (Erdem, 2013)'s study student characteristics were used, among other things, to investigate the association between student characteristics and academic achievement. The study found the attributes of gender and city of origin not to be significantly related to academic excellence. Other characteristics which had a significant relationship with academic achievement included the educational status of the father, the high school attended and attendance in class. In their study (Khan et al., 2013) identified the factors responsible for students not completing their studies. These included lack of family support, lack of commitment to study, time management and class participation. High dropout rates were also indirectly linked to factors such as wrong subject choices and changing affiliation within the system. In the South African context, (Sibanda et al., 2015) identified some of the factors responsible for academic performance amongst second year students. The following are some of the factors that were found to be associated with academic failure amongst these students: lack of commitment and dedication, regular study, class attendance and completion of assignments. In a study conducted earlier, (Khan et al., 2013) also identified lack of commitment to studies as one of the factors responsible for students dropping out or failing.

Students' expectations regarding their studies may not match the true situation they find themselves in. In other words, there may be a mismatch between expectation and actual experience. (Pather and Dorasamy, 2018) identified several factors responsible for students not completing their studies, due to unmet expectations. For a number of reasons, some of the students were dropping out of their studies. They were feeling disconnected from the university, due to factors such as lack of social and academic engagement, and lack of academic support. One of the recommendations of this study was that universities needed to be more responsive to student experiences and expectations. This could be achieved by developing strategies to prepare and support students in achieving successful outcomes.

2.6.7 The Role and Influence of Mathematics in STEM Curricula

In our faculty first time first year students can choose from among 14 options of the BSc. degree they would like to pursue. In 12 of these options the students must take the compulsory Mathematics I major subject. In the remaining two options (Bachelor of Science in Biological Sciences and Bachelor of Science in Natural Sciences) they must take the less demanding Ancillary Mathematics and Statistics I course. Mathematics therefore, in one form or another, is at the centre of all the BSc. degree offerings. In other higher education institutions, local and international, Mathematics plays a similar, pivotal role in their STEM curricula.

Educational researchers have studied and continue to study the impact of Mathematics in students' academic performance. The study by (Van der Westhuizen and Barlow-Jones, 2015) investigated the use of high school Mathematics as a criterion for student admission to programming courses. The dataset consisted of 393 records, belonging to four cohorts of students who studied between 2012 and 2015. The authors discovered a weak relationship between the Mathematics subject and programming modules. On the other hand, a strong relationship was discovered between Mathematical Literacy and the two programming modules, leading (Van der Westhuizen and Barlow-Jones, 2015) to conclude that the Mathematics scores obtained in high school cannot be used as the sole criterion for admitting first year students into programming modules.

(White and Sivitanides, 2003) used 837 first year student records collected over three years to investigate the association of mathematics with a first year visual programming course. The relationship between mathematics and computer programming was strong. In their study, (Ayalew et al., 2018) investigated the connection between gender and high school subjects, on the one hand, and an introductory programming course, on the other. The correlation between the high school mathematics subject and the introductory programming subject was significant. An earlier study by (White and Sivitanides, 2003) made a finding similar to (Ayalew et al., 2018)'s. Contrary to these findings, (Van der Westhuizen and Barlow-Jones, 2015) found a weak relationship between mathematics and the programming modules.

In their study, (Mthimunye and Daniels, 2017) used high school mathematics and science scores to determine their relationship to the scores obtained in nursing science modules taken by first year nursing science students. The second cohort comprised 266 students, who completed their first year in 2013. As in (Ayalew et al., 2018),(Tewari, 2014) and (White and Sivitanides, 2003), the importance of mathematics was also highlighted in

(Mthimunye and Daniels, 2017)'s study, which found the relationship between the Grade 12 NSC mathematics subject and first year nursing science modules to be strong.

The study by (Louw et al., 2013) was to determine the correlation between high school and first semester mathematics scores. The study found high school mathematics and physical science to be significant predictors of first year mathematics performance. In (De Winter and Dodou, 2011) three subjects (mathematics, physics and chemistry) were strong predictor variables of the first year Grade Point Average and BSc. degree achievement.

As in (Ayalew et al., 2018), the study by (Islam and Al-Ghassani, 2015) investigated whether performance in high school and gender were suitable predictors of the first year calculus subject. The hierarchical regression technique was used to analyze the data. Compared to male students, female students performed better in high school, as well as in the Calculus I course. In short, the study found gender, mathematics and overall high school performance to be significant predictors of the Calculus I subject. Given their findings on gender and high school performance, (Islam and Al-Ghassani, 2015) suggest that these factors be considered as admission criteria, to provide students with equal chances of being accepted and also to have in place a much fairer decision-making mechanism that could be used when admitting new students.

Several factors can have an impact on performance in mathematics. In numerous studies some of the factors that influence performance in mathematics have been investigated. (Kizito et al., 2016) mention the fact that in a number of higher education institutions very few students pass first-year mathematics. In their exploratory study, (Kizito et al., 2016) administered a questionnaire to 68 first year mathematics students to identify and investigate factors that influenced academic achievement in a mathematics course (MAT105). The authors discovered a significant correlational relationship between the following factors and performance in MAT105: teaching methods, teaching quality, deep or surface learning and workload. The strongest correlation was between workload and MAT105. The correlation between matriculation and MAT105 was weak, positive and insignificant. Regarding student gender, the difference performance was not significant. Workload and prior knowledge were the two significant predictor attributes of MAT105, with workload being the strongest predictor.

Similar to (Kizito et al., 2016), a study conducted earlier by (Murray, 2013) sought to determine the factors most responsible for academic achievement in a first year first semester Algebra course (MTH 111), including identifying the most influential predictors. Prior academic achievement, self-efficacy, academic resources, self-regulation and learning styles had a positive and significant correlation with performance in MTH 111. Prior academic achievement, learning styles and academic resources were the most influential predictors.

A number of studies have found mathematical background and ability to be influential factors in the academic achievement of STEM students. However, given their weak mathematical background and lack of mathematical skills, a number of first time first year students are not adequately prepared for successful study in their STEM disciplines. In a study conducted at an Australian university, (Rylands and Coady, 2009) compared first year students' performance in mathematics and related subjects to their university entrance score and performance in different levels of secondary school mathematics (advanced, intermediate and elementary). Overall, students who took secondary school mathematics at the advanced level performed better in selected first year subjects, compared to their peers who took intermediate and elementary mathematics subjects. The study also found that mathematics background was a good predictor of academic achievement in mathematically-related subjects.

Whilst (Rylands and Coady, 2009)'s study focused on the association between secondary school mathematics on academic success in first year mathematics and related subjects, the study by (Joyce et al., 2017) investigated the association of secondary school mathematics with performance in first year health sciences. As in (Rylands and Coady, 2009), students who took advanced mathematics at high school performed better in first year compared to those who took intermediate or elementary mathematics. In this case, their first year GPA was significantly higher compared to the GPA scores obtained by their peers in the other mathematics streams.

2.7 Chapter Summary

In this chapter several topics were presented and discussed in great detail. The chapter began by providing background information, to set the context for a better understanding of the study we were conducting. Given the nature of this study and its focus on data mining, it was only proper for the background discussion to introduce the concept of data mining and to provide a few examples of its applications in real life. Following the general discussion on data mining, we narrowed the discussion by focusing on Educational Data Mining (EDM), the field within which this study was conducted. We followed this with a discussion justifying the use of data mining for the research problem. We then proceeded to introduce and discuss different data mining and statistical techniques we were going to use in performing different tasks during the Modeling and Evaluation phases of the methodology we followed in conducting our research. In our experiments we used different classification, prediction, clustering and association rule mining techniques. Statistical techniques included correlation analysis, the T-test and two-way ANOVA. To complete the chapter, we discussed in detail the literature related to this study and to the research questions we needed to answer. We covered a range of

topics which included, amongst others, factors such as ethnic background, gender, funding status, English language proficiency and prior computer programming experience.

In the next chapter, Chapter 3, we discuss in detail the methodology we followed in conducting this study. This being a data mining based study, we present and discuss in detail the CRISP-DM methodology we selected to guide the research through its different phases.

Chapter 3

Research Methodology

3.1 Introduction

Chapter 2 began by providing a background discussion which focused on student funding, data mining and educational data mining. Thereafter, we presented and discussed the data mining and statistical techniques which were selected for our study. These techniques were applied on historical first year BSc. datasets, collected during the period 2015-2019. A detailed literature review was conducted which focused on a discussion of the different topics under investigation. This chapter discusses the methodology we followed in conducting this study. The sections in this chapter are organised as follows: Section 3.2 describes the context within which this research was conducted. In Section 3.3, the selected research design is described. Section 3.4 describes the research population, consisting of different cohorts of first year students enrolled in STEM disciplines between 2015 and 2019. In Section 3.5 we mention the source(s) from which we obtained the Grade 12 and first year datasets we used in this study. The data records are also described in detail in this section. In Section 3.6, we describe the different data mining and statistical analysis techniques we used to perform our experiments. Given the general focus of this study on data mining and data mining techniques, we identified and selected the CRISP-DM methodology as the methodology of choice in conducting this research. CRISP-DM is increasingly being used as the preferred, standard methodology of choice in data mining projects such as the one described in this study. In Section 3.7, a detailed discussion is provided on how, in this study, this methodology was followed through its different phases. Since confidential academic and other information about the first year students in our faculty had to be obtained, it was necessary to obtain ethical clearance before permission to access this information could be granted. The steps we took to obtain the necessary ethics clearance and the data required for our study are described in Section 3.8. Finally, a summary of this chapter is provided in Section 3.9.

3.2 Context of the Study

This study was conducted at the University of the Witwatersrand (also known as Wits), situated in Johannesburg, South Africa. In the South African context, Wits is a large research-intensive university in which students can enrol to obtain certificate, diploma and degree qualifications such as bachelor, honours, masters, and doctoral degrees. In 2019, Wits had a population of 40 881 students, including international students. In that year there were more female students (54.63%) compared to male students (45.14%). Approximately 60.95% of the total student population comprised undergraduate students. There are five faculties: Commerce, Law and Management (CLM), Engineering and the Built Environment (EBE), Health Sciences, Humanities and Science. This research was conducted within the School of Computer Science and Applied Mathematics (CSAM), one of nine schools that belong in the Faculty of Science.

To be accepted into academic programmes of their choice, first time first year students must satisfy the requirements stipulated in Table 3.1.

Table 3.1: Faculty of Science Admission Requirements

Programme	**Minimum Admission Requirements**		
BSc. in the field of:	**APS**	**Subjects**	**Scale**
Astronomy & Physics	$\geq$ 43 pts	English Home Language (EHL)	5
Chemistry with Chemical Engineering		English First Additional Language (EFAL)	5
Nuclear Sciences & Engineering		Mathematics	6
		Physical Science	6
Physical Sciences	$\geq$ 40 pts	EHL	5
		EFAL	5
		Mathematics	6
		Physical Science	5
Computer Science	$\geq$ 40 pts	EHL	5
		EFAL	5
		Mathematics	6
Biological Sciences	$\geq$ 40 pts	EHL	5
Geographical & Archeological Sciences		EFAL	5
		Mathematics	5
Geological Sciences	$\geq$ 40 pts	EHL	5
		EFAL	5
		Mathematics	6
		Physical Science	5
Actuarial Sciences	$\geq$ 40 pts	EHL	7
Mathematical Sciences		EFAL	7
		Mathematics	7
		Physical Science	7
Mathematics of Finance	$\geq$ 40 pts	EHL	5
		EFAL	5
		Mathematics	6
Bachelor of Science (General)	$\geq$ 40 pts	EHL	5
BSc. (3 years)		EFAL	5
		Mathematics	5

3.3 Research Design

In the South African higher education context, as well as globally, most students who enter university for the first time face various challenges, which include high failure rates, difficulty in transitioning from high school to university and for those who require funding for their studies, the lack of adequate finances to meet some of their personal and academic needs. The goal in this research was to obtain information about a particular group of first time first year students enrolled in critical-skill STEM disciplines at Wits University's Faculty of Science. In order for us to understand these students better and to know more about them, we collected their high school and first year records, from which we were able to extract meaningful and useful information about them. Since we would be using statistical and data mining techniques to analyze, as well as extract information from the data, we considered the quantitative research to be the most appropriate approach to adopt in conducting this study. In this study several student-related factors were considered. These included ethnicity, gender, funding status, high school performance, and performance in first year.

3.4 Research Population

There was no direct participation by students in this study. First year BSc. student records provided by the university's Academic Information and Systems Unit (AISU) were used. There were two datasets. The first dataset comprised 38 231 records, which contained information about Grade 12 National Senior Certificate (NSC) students who were enrolled, between 2015 and 2019, as first year BSc. students in the Faculty of Science. The records include demographic information such as race and gender (Table 3.2), as well as home language. To ensure confidentiality and to protect their identity, we removed any information that could be used to identify these students. In each case a unique, encrypted identity number was assigned to each record.

Table 3.2: National Senior Certificate Race and Gender Statistics (2015-2019)

Race			**Gender**		
	Frequency	**Percent (%)**		**Frequency**	**Percent (%)**
Black	24657	64.49	Female	21425	56.04
Chinese	207	0.54	Male	16793	43.93
Coloured	1568	4.10	Unspecified	13	0.03
Indian	5181	13.55			
White	6568	17.18			
Unspecified	50	0.13			

The second dataset contained records of first, second and third year students enrolled in STEM disciplines between 2015 and 2019. Only first year records were used to determine

the outcome at the end of the year. To determine degree attainment, first year datasets from 2015 to 2017 were used. Apart from demographic attributes, other variables in the dataset include financial aid status, field of study, year of study, progress outcome (overall academic performance), individual subject results and enrolment status. We excluded records belonging to part-time, dropout and repeating students, after which a total of 5 801 records remained in the final dataset. Table 3.3 shows the race and gender components of the first year student population.

TABLE 3.3: First Year Race and Gender Statistics (2015-2019)

Race			**Gender**		
	Frequency	**Percent (%)**		**Frequency**	**Percent (%)**
Black	4204	72.47	Female	2731	47.08
Chinese	33	0.57	Male	3069	52.90
Coloured	163	2.81	Unspecified	1	0.02
Indian	702	12.10			
White	699	12.05			

3.5 Data Collection

At Wits University, the Academic Information and Systems Unit (AISU) gathers and maintains data about different entities and stakeholders. Some of this data is provided on a confidential basis to requesting entities within the university. After obtaining permission from the university's Ethics Committee to access the data required for this study, a request submitted to the AISU for the provision of data belonging to a cohort of first time first year STEM students, who were enrolled during the period 2015-2019. Two different datasets were obtained from the AISU. The first contained the students' NSC Grade 12 records. The second contained their first year BSc. records.

3.5.1 The NSC Grade 12 Dataset (2015-2019)

A total of 38 231 NSC Grade 12 records from 2015 to 2019 were obtained from the AISU. These were provided as Excel spreadsheets. The records contained information which included demographic, academic performance and admission rating details. The record attributes and their types are described in the following table (Table 3.4).

Table 3.5 describes some of the NSC Grade 12 subjects taken by students who enrolled in first year STEM disciplines in the Faculty of Science.

TABLE 3.4: National Senior Certificate Grade 12 Attributes

Attribute	Type	Description
StudentID	String	Encrypted Student Number
Gender	Discrete	Student Gender (Male or Female)
Race	Discrete	Ethnic Background
Language	Discrete	Home Language
Test Segment Code	String	NSC Grade 12 Subject Code
Test Segment Name	String	NSC Grade 12 Subject Name
Test Segment Percentile	Numeric	NSC Grade 12 Subject Score
Admission Rating	Numeric	Admission Point Score

3.5.2 The First Year BSc. Dataset (2015-2019)

A number of records from the first year BSc. dataset were removed. Records belonging to part-time, dropout, and repeating students were excluded. In order to process and analyze the data, and to use it in different experiments, we used 5 801 records. The record types and attributes are described in Table 3.6.

Table 3.7 shows some of the STEM disciplines which were selected by the first year students described in this study.

3.6 Data Analysis

Data analysis involves the preparation, exploration and analysis of data. During this process the data is cleaned, transformed and analysed to extract insightful and potentially useful information. In this study, different analysis techniques were used to determine relationships among different variables and to explore similarities and differences amongst first year BSc. students. Data analysis techniques include correlation, regression, T-test and analysis of variance (ANOVA). In this data-mining based study, two software tools were used to analyse the first year BSc. datasets. We used the open source WEKA toolkit (Hall et al., 2009) to perform some of the tasks related to the research questions requiring data mining experiments had to be performed in order to answer our research questions. For statistical analysis, the proprietary Statistical Analysis Software (SAS) package was used. In the following subsections (3.6.1 and 3.6.2), we discuss briefly how we used WEKA and the SAS package in this study.

TABLE 3.5: Selected STEM Subjects (NSC Grade 12)

NSC Subject Code	Description
ACTNSC	Accounting
LSCNSC	Life Sciences
BIO-SG	Biology-SG
MATNSC	Mathematics
PSCNSC	Physical Sciences
PHC-SG	Physical Science-SG
BUSNSC	Business Studies
ECONSC	Economics
GEGNSC	Geography
MP3NSC	Mathematics Paper 3
ADV	Advanced Mathematics
TMANSC	Technical Maths
APMNSC	Advanced Programme Mathematics
ADMNSC	Additional Mathematics
AMPNSC	Agricultural Management Practices
AGSNSC	Agricultural Sciences
AGTNSC	Agricultural Technology
BEC-HG	Business Economics-HG
ECO-HG	Economics -HG
TSCINSC	Technical Science
TPSNSC	Technical Sciences
APNNCV	Applied Engineering
ENPNCV	Engineering Processes
EGDNSC	Engineering Graphics & Design
MATNCV	Mathematics
PRENCV	Professional Engineering
CVTNSC	Civil Technology
ELTNSC	Electrical Technology
MCTNSC	Mechanical Technology
CATNSC	Computer Applications Technology
INTNSC	Information Technology

3.6.1 Data Mining: Using WEKA

WEKA contains a suite of machine learning algorithms such as classification, prediction, clustering, association rule mining and feature selection algorithms (Joshi and Shetty, 2015). There are also data visualization and data pre-processing tools. In this study WEKA was chosen not only because it is freely available, but also due to the fact that it is fully implemented in Java, a factor that enhances its portability and platform-independence. WEKA's data mining algorithms can be invoked in different ways, which include using the command line interface (CLI), the graphical user interface (GUI) and a Java application programming interface (API). We used different data mining

Table 3.6: First Year BSc. Attributes

Attribute	**Type**	**Description**
StudentID	String	Encrypted Student Number
Year	String	Enrolment Year
Funding Type	Discrete	Financial Aid Status
Plan Code	String	Field of Study Code
Plan Description	String	Field of Study
Term Registration	String	Year of Study
Registration Status	String	Registered or not Registered
Progress Outcome Type	String	Overall Result Decision
Progress Outcome Type Description	String	Overall Result Decision
Course Code	String	Individual Course Code
Course Title	String	Individual Course Description
Final Mark	Numeric	Individual Course Result
Final Grade	String	Individual Course Decision Code
Final Year of Study	Discrete	Yes or No
Course Status	Discrete	Enrolled or Dropped the Course
Gender	Discrete	Student Gender (Male or Female)
Race Description	Discrete	Ethnic Background
Language	Discrete	Home Language

Table 3.7: BSc. Fields of Study (General & Structured Curricula)

Attribute	**Type**	**Description**
SMAJACOM10	String	Applied Computing
SMAJACSI10	String	Actuarial Science
SMAJASTR10	String	Astronomy Astrophysics
SMAJCHEM10	String	Chemistry
SMAJCOMM10	String	Computer Science
SMAJGEN00	String	General
SMAJGEOL10	String	Geology
SMAJMCBI10	String	Medical Cell Biology
SMAJSTAT10	String	Mathematical Statistics
SMAJACHE10	String	Applied Chemistry
SMAJAPES10	String	Animal, Plant and Environmental Sciences
SMAJBIO10	String	Biochemistry and Cell Biology
SMAJBIOL10	String	Biological Sciences
SMAJCAMS10	String	Computational and Applied Mathematics
SMAJCHEM11	String	Chemistry with Chemical Engineering
SMAJECOS10	String	Economic Science
SMAJECOT10	String	Economic Theory
SMAJGENE10	String	Genetics & Development Biology
SMAJHPHS10	String	Human Physiology

algorithms to answer some of the research questions that were posed in this study. We used classification, prediction, clustering and association rule mining algorithms.

3.6.2 Statistical Analysis: Using SAS

SAS is a widely used statistical analysis software package. It was developed in the early 1970s at North Carolina State University (Spector, 2003). In this study the SAS Enterprise Guide application was used to prepare and analyse the first year BSc. student records that were obtained from the AISU. The SAS application contains a variety of statistical analysis procedures. To perform different statistical analysis tasks we used the SAS Enterprise Guide package (version 7.1).

3.7 Proposed Methodology: CRISP-DM Process Model

Wits University, like other higher education institutions, is a data-rich environment. Over the years Wits has amassed, and continues to collect, increasing amounts of data about its students and other role players. Such an environment provides an opportunity for information from which useful insights and knowledge can be obtained about different stakeholders, including students. Various stakeholders in the institution can also use the information obtained from this data to make well-informed, student-focused decisions. In order to leverage the potential of this data as a source of meaningful and useful information, higher education institutions have begun to rely on data mining techniques as a means of extracting such information from student datasets.

Given the nature of this study and its focus on the application of data mining and statistical techniques, it was important and convenient to select and follow a methodology that would be best aligned with the nature of our study. For this reason, we adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) as the methodology of choice in undertaking this study. Currently, CRISP-DM is the standard process model adopted by most data analysts and practitioners in their data mining projects (Wirth and Hipp, 2000). The reasons for selecting and using CRISP-DM in this study include the fact that it is a freely available and application-neutral standard process model, which data mining practitioners and researchers can use to guide and assist them in executing their data mining based projects (Larose, 2014). Furthermore, we found CRISP-DM to be a logical and convenient choice, given its ability to clearly specify and perform the required tasks in each one of its six phases.

In CRISP-DM a data mining project consists of six phases, adaptive and iterative phases, namely *business understanding, data understanding, data preparation, modeling, evaluation* and *deployment* (Wirth and Hipp, 2000). These phases are illustrated in Figure 3.1.

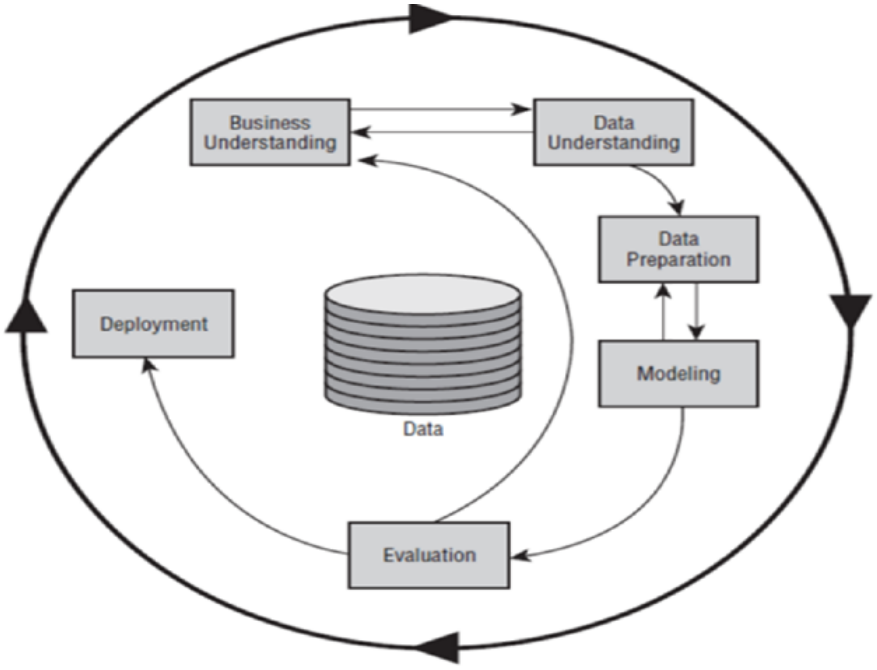


FIGURE 3.1: Phases of the CRISP-DM Reference Model (Wirth and Hipp, 2000)

The phases are not necessarily performed in a fixed sequence. There may be a need to go back to previous phases, as is shown by the arrows linking the different phases. The direction of the arrows indicates the flow of output from one phase to the next. The cyclic aspect of the activities in a data mining project is illustrated by the outer circle. A detailed description of these phases and a discussion of how they were implemented in this study is provided in the following subsections.

3.7.1 Business Understanding

This is the first phase of the CRISP-DM process model, which (Larose, 2014) also refers to as the *research understanding phase*. Several tasks are involved during this phase. These include a statement of the business or research objective(s), an assessment of the research environment, a determination of the data mining goal(s) and producing a plan for achieving the stated objective(s). In order to implement this phase, we took into account the context in which this study was conducted.

Wits University is one of South Africa's premier universities, with a strong local and international reputation. Every year thousands of first year students are accepted, who enroll in certificate, diploma and degree programmes offered by the university's five faculties. These students come from different educational, ethnic and socio-economic backgrounds. The challenge for Wits is not only to embrace this diversity, but to also ensure that the majority of these students succeed in their chosen educational streams, despite their diverse backgrounds. Through the Faculty of Science, Wits contributes

to the country's need for highly skilled and qualified graduates and professionals in the critical-skill STEM disciplines. In order to address this critical need, there are several challenges to overcome. For a sizeable number of first year students to access university and by extension the Faculty of Science, some of them require financial assistance. With adequate funding, in the form of loans and bursaries, the faculty can attract more first year students into its STEM degree programmes, a situation that could benefit the economy and the country as a whole.

The high failure and dropout rates amongst first time first year students enrolled in STEM disciplines (Jacobs, 2010) constitute another challenge. There is therefore a need for an in-depth study and understanding of some of the factors which influence academic achievement amongst students in this category. Obtaining information from the data collected about them can go a long way towards enhancing our understanding of the challenges they face, as well as the academic environment in which they learn. Such information would make it easier for relevant stakeholders to focus their attention on areas of need, where the challenges are most severe.

As its primary goal this study sought to determine whether interesting and actionable information could be uncovered from the data belonging to different cohorts of first time first year BSc. students enrolled, between 2015 and 2019, in critical-skill STEM disciplines at Wits University's Faculty of Science. Following from this goal, we identified the following data mining objectives:

1. To examine the effects of demographic attributes, namely ethnicity and gender, on academic achievement, as measured by the aggregate Admission Point Score (APS).
2. To determine the impact of the high school English Home Language (ENA) and English First Additional Language (ENB) subjects on the APS.
3. To identify the influence and contribution of selected NSC Grade 12 STEM subjects towards the APS.
4. To examine the influence of prior computer programming experience on academic achievement in the first year Computer Science subject.
5. To examine the predictive value of the Faculty of Science's minimum admission requirements subjects (Mathematics, Physical Sciences, English) on performance in first year and degree attainment.
6. To examine the influence of demographic and other student characteristics on the academic achievement of first year STEM students.
7. To examine the predictive power of Mathematics I on academic achievement in first year STEM disciplines, and degree attainment.

The plan for achieving these objectives was to use different data mining and statistical analysis techniques such as classification, prediction, correlation analysis and analysis of variance (ANOVA). To perform data mining tasks, we used the open source toolkit, WEKA. For statistical analysis we used the proprietary statistical analysis software package, SAS.

3.7.2 Data Understanding

In this phase the aim is to understand the data to be used in a project. It is important to obtain such an understanding, in order to be certain that the data being used for the project is relevant, suitable and of acceptable quality (Du, 2010). During this phase the tasks include data collection and integration, description and exploration and quality assessment.

The data for the study was obtained from a single source, namely Wits University's Academic Information and Systems Unit (AISU). It was therefore not necessary for this data to be integrated with data from other sources. We requested and were given two datasets by the AISU. The first dataset, consisting of 38 231 records, contained Grade 12 NSC results belonging to first time first year students enrolled in STEM disciplines between 2015 and 2019. A description of the record attributes and their types is provided in Table 3.4 (Subsection 3.5.1). Some of the STEM-related subjects taken by these students in Grade 12 are described in Table 3.5. The second dataset contained 5 568 records, which belonged to some of the students, who were now enrolled in first year degree programmes in the Faculty of Science, between 2015 and 2019. A description of the record attributes and their types is provided in Table 3.6 (Subsection 3.5.2). Table 3.7 shows the different STEM disciplines from which an incoming first year student can select her or his study programme.

3.7.3 Data Preparation

During this phase several tasks are performed which include selecting, cleaning, constructing, integrating and formatting the data. In order to achieve each one of the data mining objectives, we selected the relevant attributes and records from the Grade 12 and first year datasets we obtained from the AISU. Table 3.8 shows the attribute names and the number of records that were going to be used in our experiments during the Modelling Phase. Where necessary, the table also shows the number of test set and training set records that were used in different experiments.

TABLE 3.8: Selected Attributes and Records

Objective	Attributes	Records	Training	Testing
1	Gender, Funding Status, Ethnicity, Funding Type	5568		
2	ENA, ENB, APS	5643		
3	APS, NSC STEM Subjects (Table 3.5)	38347	38231	116
4	BCO, IAP, CSI Outcome, Progress Outcome Type, Prior Programming Experience			
4a		202		
4b		230	202	28
4c		198	123	75
5	Year, Funding Status, Final Year Indicator, APS, STEM Subjects, Progress Outcome Type, Qualification			
5a		5576	4517	1059
5b		1802	1370	432
6	Year, Funding Type, Gender, Ethnicity, Home Language, APS, Progress Outcome Type, ENA or ENB	6223		
7	Year, Funding Status, Final Year Indicator, Math1034 (Algebra I), Math1036 (Calculus I), Progress Outcome Type, Qualification			
7a		1786		
7b		16586	1184	474
7c		926	691	235

Having selected the attributes and records that were to be used in different data mining and statistical operations, the next task was to clean the data, in order to improve its quality. This task involved different activities, which included handling missing values, dealing with inconsistent data and where necessary, using appropriate default values. A possible approach for handling missing values would be to remove the record(s) or field(s) containing missing values. However, in this study, we did not follow this approach in dealing with missing data, given the risks involved. For example, a biased subset of the

data may result from the removal of records containing missing values (Larose, 2014). Similarly with columns, removing a column or columns because few column values are missing, may result in loss of information. Given the disadvantages associated with this approach, we decided to follow a different approach. For missing numerical values, we decided to use an out-of range value. We used -1 to replace missing numerical values. Where character values were involved, we replaced them with a question mark character ("?").

In the final task during this phase we prepared the subsets of the data we were going to use during the next, Modelling Phase. We stored the data subsets in different formats, depending on whether SAS or WEKA was going to be used to analyse a particular data subset. In WEKA, a special format is used to represent a dataset. This is the *attribute relation file format* (ARFF). A dataset consisting of several instances (records) and related attributes is stored in a ARFF file. The following is an example of a ARFF file for a vehicle classification dataset (adapted from (Bigus and Bigus, 2001)).

```
% ARFF file for vehicle data

@relation typesOfVehicles

@attribute motor {yes,no}
@attribute numWheels numeric
@attribute numDoors numeric
@attribute size {small,medium,large}
@attribute vehicle {cycle,automobile}
@attribute vehicleType {bicycle,tricycle,motorcycle,sedan,SUV}

@data
no,2,0,small,cycle,bicycle
no,3,0,small,cycle,tricycle
yes,2,0,small,cycle,motorcycle
yes,4,4,medium,automobile,sedan
yes,4,4,large,automobile,SUV
```

The first line is a comment, as indicated by the '%' sign. Next comes the internal name of the dataset, which is indicated by the *relation* keyword. In this case the relation name is *typesOfVehicles*. The *attribute* keyword is used to define attributes of different types. There are nominal, numeric and other types of attributes. For a nominal attribute

definition, the set of values it can take are enclosed in curly braces. In the example the following are nominal attributes: *motor, size, vehicle*. The *numeric* keyword is used to define numeric attributes such as *numWheels* and *numDoors*. The last attribute is the class attribute, which indicates the class or category to which a record belongs. In the example this is the nominal attribute, *vehicleType*. The *data* keyword indicates the beginning of the data section, in which dataset instances (records) are written on separate lines, with the values of the attributes in each instance being separated by commas. There are five instances in the example.

In SAS, a dataset is stored in a SAS library, as a file containing the *.sas* extension as part of its name. The file contents are organized as a table consisting of rows and columns (variable names). Figure 3.2 shows the contents of a SAS file created using the SAS Enterprise Guide software package.

	PatientID	DrugDose	Disease	BloodP
1	69	2	B	13
2	162	4	A	-47
3	181	1	B	12
4	209	4	A	-4
5	308	2	A	4
6	331	4	C	37
7	340	4	C	-19
8	350	1	B	-9
9	360	2	B	-17
10	363	4	A	-41
11	382	2	C	-5

FIGURE 3.2: SAS Dataset

As in WEKA, SAS attributes are of different types, such as *date, character or numeric*. In a SAS data file numeric attributes are shown alongside a blue icon, and character attributes appear next to a red icon. To identify date attributes, SAS uses a calendar icon. In Figure 3.2 the attributes *PatientID, DrugDose* and *BloodP* are numeric. The *Disease* attribute is of character type.

3.7.4 Modeling

The tasks during this phase include selecting and using different data mining techniques to be applied against datasets such as those described in Table 3.8. The techniques include classification, prediction, clustering and association rule mining. In cases involving classification or prediction, the dataset may be divided into training, testing and validation subsets. Thereafter, different data mining techniques are applied against the training set, a process that results in the creation of one or more models. After the training phase, the performance of the models is evaluated by applying them against one or more test datasets. Table 3.9 shows the statistical and data mining techniques we applied against different datasets.

In Chapter 4 we present, interpret and evaluate the results we obtained from the different statistical and data mining experiments we performed as part of the Modelling Phase.

3.7.5 Evaluation

This phase involves the following tasks: evaluating results, reviewing the data mining process and deciding what steps to take next. The main objective in this phase was to evaluate the performance of the models created during the Modelling Phase, the purpose being to determine how well these models performed against test datasets. The results obtained during this phase are presented in Section 4.9.

At this point we need to mention the fact that we did not implement the final phase of the CRISP-DM life cycle, namely the Deployment Phase. This was intentional. This phase involves tasks that would have taken more time to execute, thereby extending the scope of this study beyond what was intended.

3.8 Ethical Considerations

This study did not involve human participants. Therefore, we did not have to request the Ethics Committee for students to participate in this study. Ethical clearance was obtained to conduct the study and to access the records of first year BSc. students who constituted the target group. Thereafter, permission to access the records was sought and granted by the university's Academic Information and Systems Unit (AISU). In order to ensure their anonymity, students' names, numbers and any information that could be used to identify them were removed from the records. A unique, encrypted student number was used to identify each record.

3.9 Chapter Summary

This study was motivated by the need to leverage the power of data mining and statistical techniques to extract meaningful and useful information from datasets of first time first year students enrolled in STEM degree programmes at Wits University. The chapter began by providing the context for the study. Since the purpose was to process, analyse and extract information from student datasets, we considered a quantitative research design to be the most appropriate approach for conducting this study. Historical records belonging to first time first year students enrolled in STEM disciplines in the Faculty of Science were used in the different statistical and data mining experiments we conducted. From the AISU we obtained Grade 12 NSC and first year BSc. student records, which we processed and analysed using the SAS package and the data mining WEKA toolkit. Given the focus of this study on statistical and data mining techniques, CRISP-DM was chosen as the most suitable methodology to follow. A detailed description and discussion were provided on how, within our context, the different phases of this methodology were implemented. CRISP-DM has become a leading methodology in data mining based projects, a fact which influenced our decision to choose and adopt this methodology in our study. Using student records required ethical clearance from the university's Ethics Committee. Permission was granted, which enabled us to use the data in the different experiments we conducted.

In Chapter 4, we present our results from the statistical and data mining experiments we conducted in order to provide answers to the research questions we posed in Chapter 1.

Table 3.9: Selected Statistical and Data Mining Techniques

Question	**Technique**	**Purpose**
1a. Gender, Ethnicity, APS	2-way ANOVA	Difference
1b. Ethnicity, Funding Type, APS	2-way ANOVA	Difference
2. ENA, ENB, APS	Two-sample T-test	Difference
3a. NSC STEM Subjects	Feature Selection	Prediction
3b. NSC STEM Subjects	Multiple Linear Regression	Relationship
4a. Prior Programming Experience (PPE)	Two-sample T-test or Mann-Whitney	Difference
4b. CSI First Semester courses (CSI Performance)	Multiple Linear Regression	Prediction
4c. CSI First Semester courses (First Year Performance)	J48, JRIP, MLP, Naive Bayes	Prediction
5a. WITS Admission Requirements (First Year Performance)	J48, MLP, Naive Bayes, JRip	Prediction
5b. WITS Admission Requirements (Degree Attainment)	J48, MLP, Naive Bayes, JRip	Prediction
6a. Student Characteristics	Simple K-means algorithm	Clustering
6b. Student Characteristics	Apriori Algorithm	Association
7a. Mathematics I (First Year Performance)	J48, MLP, Naive Bayes, JRip	Prediction
7b. Mathematics I (Degree Attainment)	J48, MLP, Naive Bayes, JRip	Prediction

Chapter 4

Results

4.1 Introduction

In Chapter 3, we presented and discussed the research methodology we followed in conducting this study. This is the CRISP-DM methodology, which we decided was most suited to the data-mining based study we were conducting. In this chapter, we present and interpret the results we obtained from the statistical and data mining experiments we conducted. Each experiment is linked to an geared towards providing an answer to a specific research question.

Section 4.2 presents our findings on the relationship between demographic characteristics, funding status, field of study and academic performance. Section 4.3 explores the relationship between English proficiency and academic performance. Subject selection in the Grade 12 NSC curriculum and its relationship to high school performance is investigated in Section 4.4. Section 4.5 presents the findings on prior programming experience and its influence on performance in two first semester Computer I courses. In Section 4.6, we present the results obtained using the Grade 12 minimum admission requirement subjects as predictors of the first year outcome and degree attainment. Section 4.7 presents the findings on the association involving demographic characteristics, financial status, field of study and academic performance. Section 4.8 presents the findings on the predictive power of Mathematics I on first year performance and degree attainment. In Section 4.9 we present the results we obtained during the Evaluation Phase, after applying the different models we created during the Modelling Phase. Finally, Section 4.10 provides a summary of the main points covered in this chapter.

4.2 Ethnicity, Gender, Funding and Performance

South African higher education institutions face numerous difficulties, amongst which is the need to address the serious challenge of how to successfully meet the expectations of a diverse community of students. The students come from different ethnic, socio-economic and educational backgrounds. In the South African higher education context and other places where student diversity is norm, educational researchers continue to undertake studies which seek to determine, among other things, how different factors related to students' diversity affect their academic performance. In this study we used the ethnic background, gender and funding status attributes to determine their relationship to the academic achievement of first time first year students enrolled in STEM disciplines in the Faculty of Science.

In order to obtain a clearer and better understanding of the significance and role played by these factors in our educational context, we posed the following research question:

> *What are the relationships amongst the factors of ethnic background, gender, funding status and the Wits Admission Point Score (APS) obtained by first time first year BSc. students enrolled in STEM disciplines?*

In Subsection 2.6.1, we presented a detailed review of the literature focusing on these and other factors. In a study conducted by (Mashiloane, 2016) at Wits University's Faculty of Science, different clustering techniques were used to determine the influence of ethnic and gender attributes on the academic performance of Computer Science I students. The study used 802 Computer Science I student records. In this study we used the two-way Analysis of Variance (ANOVA) statistical technique to answer the research question involving these three factors, namely ethnic background, gender and funding status. The original dataset consisted of 5801 student records, which were collected during the period 2015-2019. The records were prepared for analysis using the statistical analysis software package, SAS Enterprise Guide. After preparation, a total of 5568 records remained, which we then used to conduct our analysis. Table 4.1 summarizes the demographic and funding status data we used to answer our research question.

A two-way ANOVA experiment involves a quantitative, dependent variable, and two categorical, independent variables, also known as *factors*. In our study gender, ethnicity and funding status are the independent variables and the APS is the dependent variable. Two-way ANOVA explores the relationship between the independent and dependent variables. Each factor consists of two or more levels. For example, the gender factor has two levels, the ethnicity factor has five levels and the funding type factor has nine levels.

TABLE 4.1: Demographic and Funding Descriptive Statistics

Variable	**Description**	**Frequency**	**Code**
Gender	Male (M) Female (F)	2922 2646	1 2
Funding Status	Funded (F) Male Female Non-funded (NF) Male Female	3898 2105 (54%) 1793 (46%) 1670 785 (47%) 885 (53%)	1 2
Funding Type	Bursary (B) Corporate Bursary (CB) External Bursary (EB) Government Bursary (GB) Internal Bursary Undergraduate (IBUG) NSFAS Undergraduate (NUG) non-funded (NF) Scholarship (SUG)	94 1062 200 959 87 728 1670 767	1 2 3 4 6 7 8 9
Ethnicity	Black (BL) Chinese (CH) Coloured (CL) Indian (IN) White (WH)	4047 30 155 674 662	1 2 3 4 5

Table 4.2 shows the mean and standard deviation values for the two factors, gender and race. The statistics on funding are also included.

TABLE 4.2: Admission Point Score Descriptive Statistics Segmented By Race, Gender and Funding Type

Attribute	Mean APS	Std. Dev.
Gender		
Male	44.28	4.94
Female	44.10	4.04
Race		
Black	43.73	4.37
Chinese	44.56	4.37
Coloured	43.70	4.20
Indian	45.54	4.47
White	45.81	5.01
Funding Type		
B	44.69	4.02
CB	45.55	4.66
EB	45.21	5.06
GB	43.65	3.44
IBUG	45.13	3.80
NUG	43.06	3.83
NF	41.81	3.89
SUG	49.01	2.80

4.2.1 Ethnic Background, Gender and Academic Performance

In this subsection we present our results on the relationship involving students' ethnic background and gender, on the one hand and their performance, on the other. In order to investigate this relationship, we posed the following research question:

> *What is the relationship between gender and ethnicity to the Admission Point Score (APS) obtained by first time first year students enrolled in STEM degree disciplines within the Faculty of Science?*

Since our study involved a dependent variable and two independent variables, we decided that the two-way ANOVA technique would be more appropriate for answering our research question. Using the SAS Enterprise Guide software package, we conducted a two-way ANOVA analysis experiment. The two independent variables, namely Gender and Race, consisted of two and five levels, respectively. As shown in Table 4.1, the number 1 is used to represent the male gender, and the number 2 is used to represent the female gender. The numbers representing different ethnic groups are also shown in the table. The dependent variable is the quantitative aggregate score, the APS. Table 4.3

shows the results we obtained after performing the 2x5 ANOVA analysis experiment, with Gender and Race as independent variables, and the APS as the dependent variable. The level of significance was set at α=0.05.

TABLE 4.3: Overall Two-way ANOVA Results (Gender, Race and APS)

source	**df**	**SS**	**Mean Square**	**F**	**Pr>F**
Gender	1	0.156698	0.156698	0.01	0.9293
Race	4	3871.889451	967.972363	48.59	< 0.0001
Gender*Race	4	153.348168	38.337042	1.92	0.1034

Looking at these results, the most important result is considered first, namely the Gender*Race interaction. The interaction effect between gender and ethnicity on the APS was not significant ([F=1.92, p=0.1034]). In other words, male and female students did not perform differently in terms of the APS, across all five levels of the ethnicity factor. Similarly, students from different ethnic groups did not perform differently across the two levels of the the gender factor. That is, amongst students from ethnic groups, males and females produced similar performance. Regarding Gender, the result was not statistically significant ([F=0.01, p=0.9293]). In other words, the differences in the APS obtained by male and female students were not significant. However, the Race factor had a significant effect on the APS ([F=48.59, p=$<$ 0.0001]). There was a significant difference in performance (APS) among students from different ethnic backgrounds.

Since the interaction between Gender and Race was insignificant, we proceeded to investigate each factor individually. The main effect of Gender on academic achievement (APS) is not statistically significant (p=0.9293), indicating that male and female students' performance is not significantly different. Regarding the Race factor, our results show that there is a statistically significant main effect for Race. This means that there is a significant difference in the APS achieved by students from different ethnic backgrounds. Since the race factor consisted of five levels, we had to determine which levels of this factor were significantly different from each other. We performed a multiple comparison test amongst the five means of the Race factor. The Tukey multiple comparison test we conducted produced the following results (Table 4.4):

TABLE 4.4: Adjustment for Multiple Comparisons: Tukey Kramer

Race	**APS LSMEAN (M)**	**LSMEAN NUMBER**
1	43.73	1
2	44.56	2
3	43.70	3
4	45.54	4
5	45.81	5

As shown in table 4.4, a code number (LSMEAN Number) is assigned to each race group. Table 4.5 shows the results obtained following the multiple, pairwise comparison test

we performed to determine which of the five means of the Race factor were significantly different.

TABLE 4.5: Least Squares Means for effect Race level
Pr > |t| for H0: LSMEAN(i)=LSMEAN(j)
Dependent Variable: Admission Point Score

i/j	**1**	**2**	**3**	**4**	**5**
1		0.84	1.00	< .001	< .001
2	0.84		0.86	0.76	0.56
3	1.00	0.86		< .001	< .001
4	< .001	0.76	< .001		0.81
5	< .001	0.56	< .001	0.81	

Having obtained a significant main effect for the Race, the following is a summary of the results displayed in Table 4.5:

- Black students' performance (M=43.73, SD=4.37) was not significantly different compared to Chinese (M=44.56, SD=4.37) and Coloured (M=43.70, SD=4.20) students. Black students were significantly outperformed by Indian (M=45.54, SD=4.47) and White (M=45.81, SD=5.01) students.
- The performance of Chinese students (M=44.56, SD=4.37) was not significantly different compared to that of students from other ethnic groups.
- The performance of Coloured students (M=43.70, SD=4.20) was not significantly different compared to Black and Chinese students. However, their APS results were significantly lower compared to those obtained by Indian and White students.
- Indian students performed at a significantly higher level (M=45.54, SD=4.47) compared to Black and Coloured students. However, compared to Chinese and White students, their performance was not significantly different.
- White students' performance (M=45.81, SD=5.01), compared to Black and Coloured students, was significantly different. However, compared to Chinese and Indian students, their performance was not significantly different.

Our findings, looking at the overall means, also reveal the fact that among the different ethnic groups Whites are most likely to obtain the highest APS results (M=45.81, SD=5.01), followed by Indians (M=45.54, SD=4.47), Chinese (M=44.56, SD=4.37), Blacks (M=43.73, SD=4.37) and Coloureds (M=43.70, SD=4.20).

In answering the research question we posed earlier in this subsection, we summarize below the results we obtained from the two-way ANOVA experiment we conducted.

- The main effect of gender on the APS is not significant.
- There is a significant main effect of Race on the APS, across Gender.
- The interaction effect between the Gender and Race factors on the APS is not significant.

4.2.2 Ethnic Background, Funding Status and Academic Performance

Most of the students in our tertiary institutions obtain access to higher education through the financial assistance they receive from different funding agencies, such as the National Student Financial Aid Scheme (NSFAS). Funding is provided to deserving and academically capable students from different ethnic backgrounds. The specific aim in this part of our study was to determine whether first year STEM students who receive different types of funding in first year, obtained significantly different APS results in high school.

In order to obtain information regarding relationships amongst the students' ethnic background, funding status and academic achievement, we posed the following research question:

> *What is the relationship between ethnic background and funding status to the Admission Point Score (APS) obtained by first time first year students enrolled in STEM disciplines?*

It is perhaps important to stress the fact that this question seeks to determine how the differently funded first year students performed, not in their first year, but in high school, as indicated by their APS results.

Table 4.6 shows the mean values of the APS for each combination of the two factors, race and funding type.

In order to answer the above research question, we conducted a two-way ANOVA analysis experiment involving two independent factors, namely Race and FundingStatus. From this experiment we obtained the following results (Table 4.7).

These results show a significant interaction between ethnicity and funding status ([F=2.23, p=0.0005]). The significance of the interaction points to the fact that the effect of one factor is dependent on the level of the other factor. The main effect for Race was statistically significant ([F=5.62, p=0.0002]). Students from different ethnic backgrounds

TABLE 4.6: Race and Funding Type Mean Admission Point Score Statistics

Funding Type	**Race**					**Funding Type Means**
	BL	CH	CO	IN	WH	
B	44.96	0.00	41.50	42.00	43.54	43.00
CB	45.24	0.00	45.30	48.13	47.33	46.50
EB	44.87	0.00	45.00	50.77	48.75	47.34
GB	43.53	45.50	44.00	45.18	43.89	44.42
IBUG	45.05	0.00	40.00	45.25	51.00	45.32
NUG	42.92	41.00	43.18	45.42	46.00	43.70
NF	41.49	42.35	41.68	42.43	42.23	42.03
SUG	48.21	48.50	48.10	49.04	49.69	48.71
Race Means	44.58	44.33	43.59	46.03	46.55	45.01

TABLE 4.7: Overall Two-way ANOVA Results (Race, Funding Status and APS)

source	**df**	**SS**	**Mean Square**	**F**	**Pr>F**
Race	4	334.626271	83.656568	5.62	0.0002
FundingStatus	8	5254.433139	656.804142	44.13	<0.0001
Race * FundingStatus	24	798.189101	33.257879	2.23	0.0005

obtained significantly different APS results across all levels of funding type. For example, the results we obtained earlier show that generally, Indian and White students significantly outperformed Black and Coloured students, by obtaining higher APS results, across the majority of funding types. A significant main effect was obtained between students who receive different types of funding for their studies ([F=44.13, p<0.0001]). Looking back at their performance in high school, their aggregate APS results were significantly different across all race groups. That is, students who received different types of funding in their first year, obtained significantly different APS results.

In answering the research question on the ethnic background, funding status and academic performance of our first year STEM students, we made the following findings:

- We found a statistically significant main effect for Race on the APS.
- There is a significant main effect of FundingStatus on the APS, across both levels of Race.
- We found a significant interaction between Race and FundingStatus on the APS.

4.3 English Language Proficiency

In almost all South African universities English is used as the language of teaching. Regardless of their home language background, entering students require proficiency in English. Most students are not first language speakers of English. At high school, except for home language subjects, English is used to deliver content in all the other subjects. At high school, the majority of students whose home language is not English, will select English First Additional Language (ENB) as their English subject. For first language speakers, their language of choice is English Home Language (ENA). Table 4.8 summarizes the English subject statistics about first year BSc. students who enrolled at Wits University during the period 2015-2019.

TABLE 4.8: English Home Language and English First Additional Language Statistics

Subject	**Race**	**Volume**	**Mean APS**	**Std. Dev.**
ENA	Black	1139	43.11	4.10
ENB	Black	2959	43.97	4.34
ENA	Chinese	29	44.72	4.35
ENB	Chinese	1	40.00	-
ENA	Coloured	150	43.65	4.21
ENB	Coloured	10	44.00	4.10
ENA	Indian	673	45.54	4.47
ENB	Indian	12	45.00	5.00
ENA	White	615	45.87	4.93
ENB	White	55	44.65	5.79

The students who enter Wits University, therefore, have different English language capabilities and proficiency. By focusing on this issue, there was a need to investigate the impact of English on the APS, by determining whether the mean APS obtained by ENA students was significantly different from that obtained by ENB students. To direct the investigation of this issue in a more focused manner, we posed the following research question:

Is the APS obtained by National Senior Certificate (NSC) or equivalent students who select the English Home Language (ENA) subject significantly different from the APS obtained by students who select the English First Additional Language (ENB) subject?

There were two separate groups of students: those who did ENA and those who did ENB. A total of 5643 records were used, of which 2606 belonged to ENA students, with 3037 records belonging to ENB students. The statistics for the two groups of students are shown below (Table 4.9).

TABLE 4.9: Group Statistics (Admission Point Score)

English	**N**	**Mean**	95% CL Mean		**Std. Dev.**	95% CL Std Dev		**Std. Error Mean**
ENA	2606	44.4401	44.2636	44.6167	4.5964	4.4749	7247	0.0900
ENB	3037	43.9928	43.8371	44.1484	4.3739	4.2666	4.4868	0.0794

The mean APS score for ENA students is 44.44 (Std. Dev.=4.60) and for ENB students the mean APS score is 43.99 (Std. Dev.=4.37). The results show that on average, students taking ENA are more likely to obtain a higher mean APS score compared to students who take ENB as their English subject. However, we wanted to know if the mean ENA and ENB scores were significantly different, statistically. Therefore, the relevant statistical significance test was performed, whose results will be presented in later sections. The first task was to test these separate datasets for normality. For this purpose, two normality tests were performed, namely the Kolmogorov-Smirnov and Anderson-Darling tests. Table 4.10 shows the results from the two tests.

TABLE 4.10: Normality Tests (English Home Language and English First Additional Language)

	Kolmogorov-Smirnov			**Anderson-Darling**		
Subject	**Statistic**	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
ENA	-	-	<.01	-	-	<.005
ENB	-	-	<.01	-	-	<.005

Both tests compare the values of a test statistic (p) to the significance level, alpha (α). If $p \leq \alpha$, the data does not satisfy the normal distribution condition. If $p \geq \alpha$, the condition of a normal distribution is satisfied. The Kolmogorov-Smirnov test shows that the ENA and ENB datasets are not normally distributed. The same conclusion can be reached regarding the Anderson-Darling test result. Having checked our datasets for normality, we used the SAS software package to perform an independent samples T-test, which produced the following result (Table 4.11):

TABLE 4.11: Equality of Variances

Method	Num DF	Den DF	FValue	Pr>F
Folded F	2605	3036	1.10	0.0085

The F value produced by the test is 1.10 (p=0.0085). From this result, we conclude that the two groups' (ENA and ENB) variances are significantly different from each other. We conclude from this result that it is invalid to assume homogeneity between the variances of the two datasets. The independent samples T-test conducted involving the two groups yielded the following result (Table 4.12).

TABLE 4.12: Independent Samples T-test

Method	Variance	df	t Value	Pr > t	Mean Difference	95% CL Mean	
Pooled	Equal	5641	3.74	0.0002	0.4474	0.2130	0.6818
Scatterthwaite	Unequal	5418.8	3.73	0.0002	0.4474	0.2121	0.6827

We found the difference between the ENA and ENB variances to be significant (Table 4.11). Since we are dealing with variances that are not equal, we read the p-value from the second row of Table 4.12, which shows the T-test result to be significant (p=0.0002). From this result, and in answering our research question, we conclude that the difference between the mean APS achieved by ENA students and that achieved by ENB students is significant ($t_{5418.8}$=3.73,p=0.0002).

4.4 Influence of NSC STEM Subjects on the APS

This study focuses on a specific target group, which comprises first time first year BSc. students enrolled in different degree programmes in the field of Science, Technology, Engineering and Mathematics (STEM). At high school the majority of these students study a variety of STEM-related subjects. The importance and influence of STEM subjects in preparing NSC Grade 12 students for the demanding and academically challenging STEM-based curricula of their choice at university cannot be underestimated. It is therefore necessary to determine which, amongst these courses, contribute the most towards the aggregate score, namely the APS. Obtaining a high APS places students in a good position of being accepted into their preferred first year STEM discipline. In order to identify what was considered to be the most important pre-university STEM subjects, we posed the following research question:

> *How significant is the relationship between the Admission Point Score (APS) obtained by first time first year STEM students and their NSC Grade 12 STEM subject scores?*

Table 4.13 shows how Wits uses a slightly different scale from the one used by the National Senior Certificate (NSC) to calculate the Admission Point Score (APS) from the performance scores obtained in Grade 12 by students who wish to study at the university.

TABLE 4.13: Wits Admission Point Score

NSC Scale of Achievement	NSC %	Wits APS	Wits APS %	Wits APS for Maths & English	Wits APS for Life Orientation	Wits APS for Other Subjects
		8	90-100	8+2=10	4	8
7	80-99	7	80-89	7+2=9	3	7
6	70-79	6	70-79	6+2=8	2	6
5	60-69	5	60-69	5+2=7	1	5
4	50-59	4	50-59	4	0	4
3	40-49	3	40-49	3	0	3
2	30-39	0	30-39	0	0	0
1	0-29	0	0-29	0	0	0

To calculate the total APS, Wits uses the best seven NSC subjects, which include Life Orientation. Where students obtain English and Mathematics scores between 60% and 100% inclusive, Wits uses a slightly different scale to assign points to subject scores.

In attempting to answer this question two different approaches were adopted. Firstly, in order to identify the NSC STEM subjects that influence the APS the most, different feature selection experiments were performed. The findings are presented in Subsection 4.4.1. Secondly, in Subsection 4.4.2, the findings on the linear regression experiment conducted to confirm whether the subjects identified using feature selection have the same influence when used to predict the APS are presented.

4.4.1 Feature Selection

In many instances datasets may contain records with a large number of attributes or features. Amongst these features, some may be irrelevant or redundant, and may need to be removed. This is in order to improve algorithm performance , and to reduce the high computational cost associated with processing datasets containing many features. Educational institutions accumulate and store a lot of data about different stakeholders, including students. The data records may contain hundreds of features, some of which may not be relevant or useful. Different data mining techniques are used to select relevant dataset attributes, to be used in performing different tasks such as classification and prediction.

There are three types of feature selection techniques: *filter, wrapper* and *embedded* methods. Filtering algorithms use search and evaluation methods to determine which

features are relevant. Following the selection of relevant features, an algorithm is applied against a dataset using only the selected features. Whereas filtering methods are used independently from data mining algorithms, wrapper methods, however, incorporate algorithms as part of the feature selection process. That is, search methods are wrapped with different classifiers, which are used to evaluate features. In embedded feature selection, it is data mining algorithms that determine features that should be used, as well as those that should be omitted. In this study only filtering techniques were used. These include Correlation-based Feature Selection (CFS), Gain Ratio (GR), Chi-Squared, and Information Gain.

From the list of NSC Grade 12 subjects shown in Table 3.5 the following STEM subjects shown in Table 4.14 were selected.

TABLE 4.14: Selected STEM Subjects

NSC Subject	**Description**	**Volume**	**Mean**	**Std dev**
LSCNSC	Life Sciences	25202	75.57	9.81
MATNSC	Mathematics	31343	72.39	13.52
PSCNSC	Physical Sciences	26324	71.98	13.21
GEGNSC	Geography	13440	73.74	9.88
MP3NSC	Mathematics Paper 3	2 772	62.72	18.56
INTNSC	Information Technology	1686	74.99	13.60
ENGNSC	English First Additional Language (ENB)	13563	73.17	7.10
ENGNSC	English Home Language (ENA)	22353	72.97	7.54
CATNSC	Computer Applications Technology	2164	70.88	12.06
APMNSC	Advanced Programme Mathematics	1892	65.05	18.21
AGSNSC	Agricultural Sciences	1607	70.64	9.46

First, a *.arff* file consisting of 38231 records was prepared. The STEM subjects in Table 4.14 were then used to predict the APS, without using feature selection techniques. The following attributes from the .arff file were removed, given the very small numbers of students enrolled in the related subjects: ADVNSC, TMANSC, TPSINSC and TSCINSC. The remaining 12 NSC STEM attributes, including the two English subjects (ENA and ENB) were used. From the WEKA toolkit the following data mining algorithms were selected and executed: decision tree (REPTree), rule-based (M5Rules), function-based (MultilayerPerceptron,MLP) and nearest neighbour (IBk). All the records were used as the training set. Our results from these experiments are displayed in Table 4.15.

It can be seen from these results that overall, based on all the five measures of performance, the nearest neighbour algorithm (IBk) has outperformed the other three algorithms. The algorithm has the largest correlation coefficient (0.8729), indicating a very strong relationship between the actual and predicted APS. For each error metric, IBk also has the smallest value. The REPTree algorithm performed second best on all five

TABLE 4.15: Predicting the Admission Point Score (Without Feature Selection)

Evaluation Metrics	**REPTree**	**M5Rules**	**MLP**	**IBk**
Correlation Coefficient	0.7724	0.7429	0.7149	0.8729
Mean Absolute Error (MAE)	2.2709	2.4549	3.3501	0.7488
Root Mean Squared Error (RMSE)	4.6169	4.866	5.6943	3.5464
Relative Absolute Error (RAE)	45.2093%	48.8739%	66.6956%	14.9067%
Root Relative Squared Error (RRSE)	63.5102%	66.9451%	78.3305%	48.7845%

metrics compared to M5Rules and MLP which, respectively, were third and fourth in terms of performance.

After using the 12 selected attributes in Table 4.14 to predict the APS, we thereafter used different feature selection methods to identify the strongest predictor attributes of the APS. In Table 4.16 the following feature selection techniques are described, available in the WEKA toolkit, which were used to identify relevant predictor attributes: CfsSubsetEval, ClassifierSubsetEval, ReliefAttributeEval and WrapperSubsetEval.

TABLE 4.16: Selected Feature Selection Techniques

Evaluator	**Description**
CfsSubsetEval	Determines a feature's ability as a predictor, as well as the extent to which features are redundant.
ClassifierSubsetEval	Uses training data or a seperate hold out test dataset to evaluates attribute subsets.
ReliefAttributeEval	Determines the strength of an attribute through a repeated sampling of an instance, and uses the value of the attribute to locate the nearest instance belonging to the same and different category.
WrapperSubsetEval	Uses a learning scheme to evaluate sets of attributes.

In Table 1, Section A of the appendix, the results from the application of different different feature selection methods are displayed. The ClassifierSubsetEval and WrapperSubsetEval algorithms, in conjunction with the RandomSearch method, identified the most relevant predictor attributes, namely MAT, PSC, GEG, INT and AGS. To this group of attributes the APM attribute was included, which appeared prominently in the list of attributes generated by the CfsSubsetEval algorithm. Having performed experiments in which the APS was predicted using a full set of attributes (without feature selection), we next wanted to see whether improved results could be obtained using the six attributes identified through feature selection. Again, the same set of algorithms used earlier to predict the APS were selected, namely REPTree, M5Rules, MLP and IBk. In Table 4.17 the results obtained using the selected attributes are presented.

These results, compared to the results shown in Table 4.15, indicate the following:

- The fewer attributes obtained through feature selection do not necessarily result in better algorithm performance. The results show that using a complete set of attributes yields better performance.

TABLE 4.17: Predicting the Admission Point Score (With Feature Selection)

Evaluation Metrics	**REPTree**	**M5Rules**	**MLP**	**IBk**
Correlation Coefficient	0.7042	0.6847	0.6449	0.7595
Mean Absolute Error (MAE)	2.7926	2.918	3.3752	2.0381
Root Mean Squared Error (RMSE)	5.1617	5.298	5.7814	4.7285
Relative Absolute Error (RAE)	55.5955%	58.0937%	67.1939%	40.5744%
Root Relative Squared Error (RRSE)	71.0049%	72.8795%	79.5297%	65.0451%

- In both cases (using and not using feature selection) the nearest neighbourhood algorithm (IBk) has again outperformed the other three algorithm.

Despite the fact that feature selection did not result in better algorithm performance, the main focus of the question was to determine which NSC STEM subjects were strong predictors of the APS. Using different feature selection techniques, the following NSC STEM subjects were identified as having the strongest influence as predictors of the APS: MAT, PSC, GEG, INT, APM and AGS. From these results, it is also clear that for the datasets the IBk algorithm could be the best candidate for predicting the APS values of unseen NSC STEM student records.

4.4.2 Linear Regression

Linear regression is another technique that can be used to identify attributes which have a strong influence in predicting the response variable. There is *simple linear regression* and *multiple linear regression*. In simple regression, only one variable is used to predict the values of the response or dependent variable. In multiple linear regression at least two predictor variables are used to predict the dependent variable. In this study, we used the multiple linear regression technique to identify statistically significant, NSC STEM predictor attributes of the APS. Since the response and predictor variables were numeric, multiple linear regression was the most suitable technique to consider using. Equation 4.1 represents a multiple linear regression model, in which the dependant variable is expressed as a linear combination of weighted predictor variables.

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \tag{4.1}$$

where

- y is the response or class variable,
- x_1, x_2, $\ldots$, x_n are the predictor attribute values and
- w_0, w_1, w_2, $\ldots$, w_n are the corresponding weights

The WEKA toolkit contains a linear regression algorithm (LinearRegression), which we applied against a dataset of 38231 NSC STEM records. The dataset contained the following NSC STEM subject attributes: LSC, MAT, PSC, GEG, MP3, INT, CAT, ENA, ENB, APM, AGS. This is the same list of subjects shown in Table 4.14, which were used in the experiments presented in Subsection 4.4.1. Our experiment produced the following regression model:

APS =
(0.0042 * LSC) + (0.0815 * MAT) + (0.0186 * PSC) - (0.0048 * GEG) +
(0.0097 * MP3) - (0.0041 * INT) - (0.0110 * CAT) + (0.0531 * ENA) +
(0.0610 * ENB) + (0.0499 * APM) - (0.0045 * AGS) + 31.4476

Looking at the above model equation, the MAT coefficient value is the largest (0.0815), indicating the fact that the NSC Mathematics subject is the strongest predictor of the APS compared to the other NSC STEM subjects. Another Mathematics-based subject, APM (Advanced Programming Mathematics), emerged as the next strongest predictor of the APS amongst the STEM subjects, with a coefficient value of 0.0499. The importance of English (ENA and ENB) as a medium of instruction is also highlighted by the model. Both English subjects contribute strongly towards the APS, with ENB showing a somewhat stronger contribution (0.0610) compared to ENA (0.0531).

As part of the result, the following output was also obtained (Figure 4.1):

```
=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.5547
Mean absolute error                      3.9865
Root mean squared error                  6.0487
Relative absolute error                 79.3658 %
Root relative squared error             83.2058 %
Total Number of Instances            38231
```

FIGURE 4.1: Multiple Linear Regression Correlation Output

The value of the correlation coefficient is 0.5547, indicating a a moderate correlation between the true and predicted value of the APS. We also used the SAS software package to analyze our dataset. The following output, shown in Figure 4.2, was obtained:

This figure contains information on the statistical significance of the regression model which, in this case, is reported to be statistically significant (F=1544.11.72, Pr=<0.0001). The Parameter Estimates section contains the results of the regression model produced

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: APS

Number of Observations Read	38231
Number of Observations Used	38231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	621625	56511	1544.11	<.0001
Error	38219	1398736	36.59792		
Corrected Total	38230	2020360			

Root MSE	6.04962	R-Square	0.3077
Dependent Mean	41.17151	Adj R-Sq	0.3075
Coeff Var	14.69371		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	31.44763	0.11461	274.39	<.0001
LSC	1	0.00425	0.00101	4.19	<.0001
MAT	1	0.08145	0.00173	47.10	<.0001
PSC	1	0.01864	0.00145	12.82	<.0001
GEG	1	-0.00476	0.00092815	-5.13	<.0001
MP3	1	0.00966	0.00186	5.19	<.0001
INT	1	-0.00407	0.00206	-1.98	0.0478
CAT	1	-0.01102	0.00186	-5.93	<.0001
ENA	1	0.05314	0.00203	26.12	<.0001
ENB	1	0.06098	0.00216	28.26	<.0001
APM	1	0.04987	0.00220	22.68	<.0001
AGS	1	-0.00446	0.00226	-1.97	0.0486

FIGURE 4.2: SAS Multiple Linear Regression Output

by the SAS package. The model consists of the y intercept (31.44763) and several predictor variables, whose associated coefficients (parameter estimates) appear under the Parameter Estimate column.

Two NSC STEM attributes, namely INT (Information Technology) and AGS (Agricultural Sciences) are weak, but statistically significant predictors. The following subjects emerged as strong, statistically significant predictors: Life Sciences (LSC), Mathematics (MAT), Physical Sciences (PSC), Geography (GEG), Mathematics Paper 3 (MP3), Computer Applications Technology (CAT), English Home Language (ENA), English First Additional Language (ENB) and Advanced Programming Mathematics (APM). The results show Mathematics (MAT) to be the most influential subject, with the largest coefficient value of 0.08145. An example interpretation of this result is as follows: increasing MAT by one unit corresponds to a 0.08145 increase in the dependent variable, the APS. English, as a medium of instruction, also contributes significantly towards the APS. In the case of English First Additional Language (ENB) the coefficient value is 0.06182. For English Home Language (ENA) the value is 0.05366. The linear regression results are summarized Section A, Table 2 of the appendix.

Our findings using feature selection and multiple linear regression techniques have shown

that either one of these techniques can be used to identify the most influential predictor attribute or attributes of the APS. In this case, both techniques were able to identify the following NSC STEM subjects as the strongest predictors: Mathematics (MAT), Physical Sciences (PSC), Advanced Programme Mathematics (APM), Information Technology (INT), Geography (GEG) and Agricultural Sciences (AGS). Whilst four of these NSC STEM subjects were strong, statistically significant predictors, the INT (p-value=0.0478) and AGS (p-value=0.0486) subjects, on the other hand, were barely significant in their contribution towards the APS.

4.5 Prior Programming Experience (PPE)

The South African high school curriculum includes Information Technology (IT) as one of the STEM subjects students can select as part of the NSC subjects they can study. In the IT subject Students, among other things, learn how to write computer programs using programming languages such as *Java* and *Python*. At high school, the IT syllabus is offered over a period of three years, from Grade 10 to Grade 12. When Grade 12 NSC IT students complete their studies, they will hopefully have acquired the necessary background and IT skills to see them through the challenging undergraduate Computer Science curriculum, which involves a very significant amount of computer programming.

Amongst others, this study sought to determine whether prior programming experience, as measured through performance in the Grade 12 NSC IT subject, could be used to differentiate between two groups of first time first year Computer Science I students, namely those who have done NSC IT at high school, and those who have not done the subject.The aim was to determine whether these students would perform significantly differently in two first-semester Computer Science I courses, based on whether or not they had prior computer programming experience. Furthermore, the two courses were also used to predict performance in Computer Science I, and to predict the final, first year overall outcome. In order to better address these three objectives and to obtain the necessary answers, we posed the following research question:

> *Is there a significant difference in performance in the two first-semester Computer Science I topics, namely Basic Computer Organization I (BCO) and Introduction to Algorithms and Programming I (IAP), between first time first year BSc. students with prior computer programming experience and their peers, who do not have prior computer programming experience?*

This question we addressed by performing three experiments. In Subsection 4.5.1 we look at whether a significant performance difference in BCO and IAP exists between

students with experience and those who lack experience in computer programming. In the next task, in Subsection 4.5.2, we discuss how the two first semester Computer Science I modules (BCO and IAP) were used to predict performance in the four-module Computer Science I course at the end of the second semester. In this task we used the linear regression analysis technique to predict numeric Computer Science I scores. In Subsection 4.5.3 we describe the experiments we performed, using different data mining algorithms, to predict first year students' performance at the end of the academic year.

4.5.1 Computer Science I: First Semester Academic Performance

In our School, Computer Science I students must take four courses, described in Table 4.18.

TABLE 4.18: Computer Science I Courses

Course Code	Description	Points	Semester
COMS1015	Basic Computer Organisation (BCO)	9	First
COMS1018	Introduction to Algorithms and Programming (IAP)	9	First
COMS1017	Introduction to Data Structures and Algorithms (IDSA)	9	Second
COMS1016	Discrete Computational Structures (DCS)	9	Second

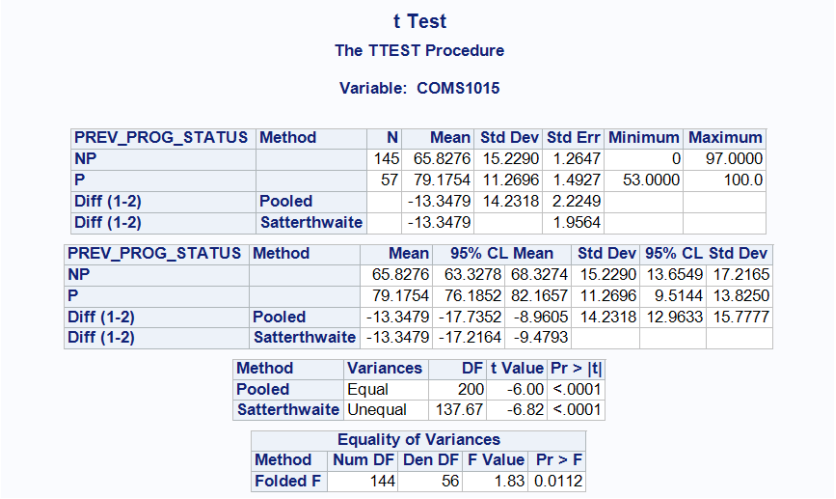
Apart from students whose major subjects include Computer Science, these courses are also taken by students who enrol in the following structured STEM curricula: *Bachelor of Science in the field of Applied Computing, Bachelor of Science in the field of Computing and Mathematics* and *Bachelor of Science in the field of Mathematics of Finance.* Given the importance and influence of Computer Science I in these structured STEM degree programmes, we focused on prior computer programming as a factor of interest, to determine its effect on first semester performance in BCO and IAP between two groups of Computer Science I students. One group consisted of students who did the Information Technology subject at high school (prior computer programming experience). The other group were students whose high school subjects did not include Information Technology.

In answering the question we used 526 records, which belonged to first time first year Computer Science I students. We used the SAS Enterprise software package to prepare the records for statistical analysis. We identified several records belonging to students who did not persist in BCO and/or IAP during the first semester. After removing all these records, the final dataset contained 202 records, which we used in the statistical analysis experiments.

BCO First Semester Performance

In our Computer Science I curriculum, BCO is taught in the first semester. This course is taken by two groups of first year students: those with prior programming experience (PPE) and those who lack such experience (without PPE). We wanted to investigate whether the two groups of students would obtain significantly different results in BCO, based on prior computer programming experience. We considered these two groups to be independent, in which case we found the independent-groups T-test to be most suitable for analyzing the BCO dataset.

The dataset comprised 202 first year Computer Science I records, of which 145 belonged to students without prior computer programming experience. The remaining 57 records belonged to students who did Information Technology at high school. Using the SAS Enterprise Guide software package and the 202 BCO records, we conducted a T-test to determine whether, based on prior programming experience, the students' performance in BCO would differ significantly. Figure 4.3 shows the T-test output produced by SAS.



t Test

The TTEST Procedure

Variable: COMS1015

PREV_PROG_STATUS	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
NP		145	65.8276	15.2290	1.2647	0	97.0000
P		57	79.1754	11.2696	1.4927	53.0000	100.0
Diff (1-2)	Pooled		-13.3479	14.2318	2.2249		
Diff (1-2)	Satterthwaite		-13.3479		1.9564		

PREV_PROG_STATUS	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
NP		65.8276	63.3278	68.3274	15.2290	13.6549	17.2165
P		79.1754	76.1852	82.1657	11.2696	9.5144	13.8250
Diff (1-2)	Pooled	-13.3479	-17.7352	-8.9605	14.2318	12.9633	15.7777
Diff (1-2)	Satterthwaite	-13.3479	-17.2164	-9.4793			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	200	-6.00	<.0001
Satterthwaite	Unequal	137.67	-6.82	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	144	56	1.83	0.0112

FIGURE 4.3: Basic Computer Organization T-test Output

In this figure the first two tables provide descriptive statistics such as the mean, confidence limit (CL), standard deviation (Std Dev) and standard error (Std Err). The mean BCO mark obtained by PPE students is 79.175, which is much higher compared to the mean BCO mark of 65.828 obtained by their non-PPE peers. The bottom table provides information about whether the variances between the BCO scores of the two groups are equal or not. The normal assumption is that the variances are equal (*homogeneity of variance*). SAS uses the **Folded F** procedure to determine if two different datasets have equal variances. In this case the variances were not equal (Pr>F=0.0112).

Having determined that the variances between the two BCO scores were not equal (p-value<0.05), we consulted the third table in Figure 4.3 to decide on which T-test result should be used. Since we found the variances to be unequal we use the second row

of the table (**Satterthwaite's** approximation). We found the result to be statistically significant (DF=137.67, t-value=-6.82, (Pr> $|t|$)<0.0001). The conclusion from this result is that there is a significant difference between the BCO results of students with prio programming experience, and those who lack such experience.

IAP First Semester Performance

To address this part of the question we used the 202 first year records belonging to the same students who took BCO as one of their first semester courses. Of this number, 145 records belonged to non-PPE students, and 57 records belonged to PPE students. As in the case of BCO, we wanted to determine whether performance in IAP was significantly different between PPE and non-PPE students. The T-test experiment we conducted to answer this question produced the output shown in Figure 4.4.

t Test

The TTEST Procedure

Variable: COMS1018

PREV_PROG_STATUS	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
NP		145	71.1793	16.3532	1.3581	0	100.0
P		57	89.5789	8.2461	1.0922	68.0000	100.0
Diff (1-2)	Pooled		-18.3996	14.5460	2.2740		
Diff (1-2)	Satterthwaite		-18.3996		1.7428		

PREV_PROG_STATUS	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
NP		71.1793	68.4950	73.8636	16.3532	14.6629	18.4874
P		89.5789	87.3910	91.7669	8.2461	6.9618	10.1159
Diff (1-2)	Pooled	-18.3996	-22.8838	-13.9155	14.5460	13.2495	16.1260
Diff (1-2)	Satterthwaite	-18.3996	-21.8375	-14.9617			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	200	-8.09	<.0001
Satterthwaite	Unequal	188.13	-10.56	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	144	56	3.93	<.0001

FIGURE 4.4: Introduction to Algorithms and Programming T-test Output

In this case, as in BCO, students with prior programming experience obtained a higher IAP mean score (89.579) compared to the mean score obtained by students without prior programming experience (71.179). Looking at the bottom table, the result shows that the variances for the two groups are not equal ((Pr>F)< 0.0001). Since the variances of the two datasets are not equal, in Figure 4.4 we consult the second row of the third table, which shows that the mean IAP scores obtained by PPE and non-PPE students are significantly different (DF=137.67, t-value=-10.56, (Pr> $|t|$) < 0.0001).

In answering the research question, regarding the influence of computer programming experience in the performance of first time first year Computer Science I students, our findings show that

- there is a significant performance difference in BCO between students with prior computer programming experience and those who lack such experience. Students with prior computer programming experience obtained significantly higher BCO scores (N=57, Mean=79.18, SD=11.27), compared to their peers who have not programmed before (N=145, Mean=65.83, SD=15.23).
- there is a significant performance difference in IAP between students with prior computer programming experience and those who lack such experience. Students with prior computer programming experience obtained significantly higher IAP scores (N=57, Mean=89.58, SD=8.25), compared to their peers who have not programmed before (N=145, Mean=71.18, SD=16.35).

4.5.2 Predicting Performance in Computer Science I

Having investigated and determined the influence of prior programming experience on performance in BCO and IAP, the next objective was to determine whether these two first semester courses could be used as predictors of performance in Computer Science I at the end of the second semester. Since we wanted to predict numeric Computer Science I scores, we found the multiple linear regression technique to be most suitable for this task. Using the WEKA software toolkit, we prepared a *.arff* file consisting of 202 records and three attributes (two independent predictor variables and one dependent variable). After selecting and applying the WEKA toolkit's *inearRegression* algorithm against our dataset, we obtained the following regression equation (model):

$$\text{CS1} = (0.3949 * \text{COMS1015}) + (0.5641 * \text{COMS1018}) + 0.7672$$

This model can be used to predict final Computer Science I scores at the end of the academic year. Looking at the regression coefficients, the COMS1018 (IAP) coefficient (0.5641) is greater compared to the COMS1015 (BCO) coefficient (0.3949). This means that between the two courses, IAP contributes more to the CSI numeric score compared to BCO. In addition to the linear regression model, other output from WEKA includes model evaluation results, obtained using different metrics. The output is displayed in Table 4.19.

TABLE 4.19: Multiple Linear Regression Results

Evaluation Metric	**Value**
Correlation Coefficient	0.93
Mean Absolute Error	4.40
Root Mean Squared Error	5.55
Relative Absolute Error	38.27 %
Root Relative Absolute Error	35.92 %
Total Number of Instances	202

The correlation between CSI and the two predictor attributes, BCO and IAP, is very high (R=0.93). the error values are also low. The value of the squared correlation coefficient R^2=0.86, which means that together BCO and IAP are responsible for 86% of the variability in CSI. This is a good result. The model can be applied against test datasets to predict final, CSI scores.

Testing the Regression Model

After obtaining the multiple regression model we needed to test it, to determine how well it could predict the known target values in a given test dataset. We applied the model on the 2019 Computer Science I test dataset, in order to determine its effectiveness in predicting Computer Science I scores. In Section 4.9, we present the results we obtained after evaluating the performance of our regression model.

4.5.3 Predicting the First Year Overall Result

It is often useful to be able to determine or predict the overall academic performance of a student when she or he completes their year of study. One advantage of such an approach is to obtain an idea regarding whether individual students will succeed in their studies and proceed to the next academic year (persistence), discontinue (drop out) or fail to proceed to the next academic year. Early prediction of success or failure can enable academics, student counsellors and administrators (school and faculty) to provide appropriate academic and other student support measures. Students who are predicted to succeed can be encouraged (and perhaps incentivised) to continue doing well in their studies. On the other hand, students for whom failure is predicted, proactive steps to prevent such failure can be taken. For these "at-risk" students preventive measures that could be taken include counselling, tutorial support and extra lessons, time management and study skills, and so on.

Given this background and following from the research question we posed earlier, regarding first semester performance in BCO and IAP, we thought that it would be interesting

to know whether these courses could be used together to predict the overall outcome at the end of first year. In order to answer our research question, we prepared a dataset containing first year BSc. student records. The 123 records we used belonged to students who registered for and completed the two courses during the period 2015-2016. From the WEKA toolkit we selected and applied the following data mining algorithms against the dataset: *JRIP* (Rule-based), *J48* (Decision tree), *Naive Bayes (NB)* (Bayesian) and *MLP* (Function). The *.arff* file we prepared to conduct our experiments contained two numeric predictor attributes (BCO and IAP) and the categorical, class attribute (Progress_outcome). A description of the codes we used to classify the overall result decision (outcome) is provided in Table 4.20.

TABLE 4.20: First Year Outcome Symbolic Codes

Description	**Outcome**
In progress	INP
Renewal of registration permission with conditions	M1C
Recommendation that registration renewal not be permitted	M1E
Renewal of admission permitted with conditions	M2C
Permission to renew admission refused	M2E
Representations made to Wits Re-admissions Committee	MBJ
Exclusion waived-must return to the same year of study	MBR
Failed minimum requirements - Needs permission to re-register in same faculty	MBZ
Minimum requirements not met	MRNM
Permitted to proceed	PCD
Must return to complete the requirements for the year of study	RET
Term withdrawn	WDR

TABLE 4.21: Predicting First Year Overall Result Decision

	JRIP	**J48**	**NB**	**MLP**
Correctly Classified (%)	81.30	82.92	78.86	77.23
Incorrectly Classified (%)	18.69	17.07	21.13	22.76
Kappa Statistic	0.44	0.41	0.38	0.09
Mean Absolute Error	0.09	0.09	0.09	0.10
Root Mean Squared Error	0.22	0.21	0.22	0.22

Following the application of each of the four data mining algorithms against the dataset, the results from our experiments are displayed in Table 4.21. The most accurate algorithm is J48 (82.92%), followed by JRIP (81.30%). The least accurate algorithm is MLP (77.23%). The algorithm with the highest Kappa statistic value is JRIP (0.44). Concerning the Root Mean Squared Error, J48 has the lowest error value (0.21). Looking at all the other measures, apart from accuracy, we judged the J48 algorithm to be the best performing algorithm compared to the other algorithms. We therefore decided to use J48 during the Evaluation Phase (Section 4.9) to predict the overall outcome, from a test dataset of first year student records.

4.6 Minimum Admission Requirements

At Wits, first time first year students wishing to study for a BSc. degree must satisfy certain minimum requirements. For each field of study there are specific requirements. For example, a first year student wishing to enrol for a BSc. degree in the field of *Mathematical Sciences*, she or he must satisfy the following minimum requirements for admission to the programme: Mathematics ($\geq$80%), Physical Sciences ($\geq$80%), English ($\geq$80%) and Wits Admission Point Score (APS, $\geq$40%). The importance of minimum admission requirements as an entry point to a chosen field of study cannot be overemphasized. Acceptance or rejection in a particular field of study depends on whether or not these requirements are satisfied.

This study also sought to determine whether the minimum admission requirements could be used to predict how first time first year BSc. students enrolled in different STEM fields would perform in their studies. First, we used these requirements to predict how first year students would perform at the end of the academic year. Second, the requirements were used to predict degree attainment after three years of study. In order to conduct the investigation in a more focused manner, we posed the following research question:

> *How suitable are the Faculty of Science's minimum requirement subjects as predictors of first year and degree attainment outcomes?*

For first time first year students to be accepted in the field of their choice within the Faculty of Science, they must satisfy the minimum admission requirements specified by the faculty. In almost all the fields of study, the following Grade 12 NSC subjects, including the APS, are required: Mathematics, Physical Sciences and English. In this study we used the APS and the marks obtained in the three NSC subjects to perform the tasks related to the question we had to answer. First, we used the data to predict the overall, first year results. Thereafter, we used the data to predict whether or not a first time first year student enrolled in a specific STEM discipline would qualify, by obtaining a degree at the end of their studies. In Subsection 4.6.1 we present our findings on predicting the first year outcome. The findings we present in Subsection 4.6.2 focus on the work we did to predict degree attainment or qualification.

4.6.1 Predicting First Year Outcome

New students find the first year experience at university challenging, given the need for them to make huge adjustments to a new environment. Due to differences in academic and social skills some students, compared to others, thrive and adjust better to the new

environment. Not only do these students come from different educational backgrounds, they are also diverse in terms of their ethnic, cultural and economic backgrounds. Students who adjust easily tend to perform better and succeed in their studies. On the other hand, students who struggle to adjust may under-perform or drop out of their studies. In order for academics and other stakeholders to provide meaningful support to these students to prevent their failure, one of the solutions is being able to predict their performance. Students who are predicted to fail can then be advised, counseled and supported to prevent them from failing.

In this study we wanted to determine whether the Faculty of Science's minimum admission requirement subjects (APS, NSC Mathematics, NSC Physical Science and NSC English together were suitable predictor attributes of overall academic performance amongst first time first year STEM students. We also considered the fact that in the NSC curriculum students select either English Home Language (ENA) or English First Additional Language (ENB) as their English subject. For this reason we prepared two WEKA-based *.arrf* files consisting of Grade 12 NSC records, collected during the period 2015-2019. The first *.arff* file contained 2123 records, belonging to ENA students. The second .arff file contained 2462 records, in which the students took ENB as their English subject. Both files contained four numeric predictor attributes (APS, MAT, PSC and ENA/ENB) and a categorical class attribute (Progress_outcome). A description of the codes used to indicate the overall result decision (Progress_outcome) is provided in Table 4.20.

To predict the first year outcome we selected the WEKA toolkit's J48 algorithm and applied it against our training datasets. We tested the resulting model during the Evaluation Phase, as described in Section 4.9.

First Year Outcome (ENA Students)

Figure 4.5 shows the output we obtained after applying the J48 algorithm against the full training set of 2123 records belonging to first year students who took the ENA subject at high school.

It can be seen from these results that about 73.4% of the training set records are correctly classified according to the class to which they belong. This means that the model is 73.4% successful in predicting the first year outcome of students who took the ENA subject in high school. The Kappa statistic measures the strength of agreement between the actual (observed) and expected classification of a dataset. In this experiment the Kappa statistic value is 0.29, which reflects fair agreement between the actual and expected result. In Section 4.9 the model produced in this experiment will be evaluated, and a

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances        1558               73.3867 %
Incorrectly Classified Instances       565               26.6133 %
Kappa statistic                          0.2948
Mean absolute error                      0.0801
Root mean squared error                  0.2001
Relative absolute error                 79.4222 %
Root relative squared error             89.2837 %
Coverage of cases (0.95 level)          97.0325 %
Mean rel. region size (0.95 level)      27.414  %
Total Number of Instances             2123
```

FIGURE 4.5: First Year Performance (English Home Language Training Set)

test dataset will be used for this purpose. The model created during the training phase is stored in a WEKA-based file named *ENAQ5Training.model*.

The numbers of successfully categorized and unsuccessfully categorized records are used to construct a table known as a *confusion matrix*, shown in Figure 4.6.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h    i    j   <-- classified as
   0   0   0   0   0   0   0   0    2    0 |   a = INP
   0   2   0   0   0   0   0   0   13    3 |   b = M1C
   0   0   0   0   0   0   0   0    2    0 |   c = M1E
   0   1   0   1   0   0   0   0    5    2 |   d = M2C
   0   0   0   0   3   0   0   0   10    3 |   e = M2E
   0   0   0   0   1  45   0   0  126   16 |   f = MBR
   0   0   0   0   0   0   3   0   33    5 |   g = MBZ
   0   0   0   0   0   3   0   3   13    2 |   h = MRNM
   0   2   0   1   0   4   1   0 1415    9 |   i = PCD
   0   1   0   0   0   6   1   2  298   86 |   j = RET
```

FIGURE 4.6: English Home Language Confusion Matrix

Since there are many classes to which the training set records can be assigned, some of these records are assigned to classes to which they do not belong. For example, a total of

- 394 records actually belong to the RET class. However, only 86 (21.8%) of these are correctly assigned to the RET class. The remaining 308 (78.2%) records are incorrectly assigned to other classes, mostly to the PCD class. In this case 298 (75.6%) students who should be repeating their first year of study (RET) are incorrectly classified as students who have passed their first year (PCD). This is a serious misclassification error.
- 41 records actually belong to the MBZ class, representing students who have failed minimum requirements and should obtain permission to re-register in the same faculty. However, only three records (7.3%) are correctly classified as MBZ records. The remaining 38 (92.7%) records are misclassified as either PCD (80.5%) or RET (12.2%). The most serious classification error is for the majority of excluded MBZ students to be classified incorrectly as having passed (PCD).

It is worth noting that when it comes to the predominant class (PCD), most of the records belonging to this class are classified correctly. Out of the 1432 records that belong to this class, 1415 (98.8%) of them are classified correctly as PCD records. The remaining 17 (1.2%) records are incorrectly classified into different classes. If we look at the records belonging to the other, minority classes, most of them are classified incorrectly as PCD records. The classifier is unable to correctly classify minority class records. This is an example of the *imbalanced learning problem*, which is characterized by a skewed class distribution, leading to inaccurate classification of the records across different classes (He and Garcia, 2009).

First Year Outcome (ENB Students)

After applying the J48 classifier against the dataset of 2462 records of first year BSc. students who took the ENB subject at high school, we obtained the following output (Figure 4.7):

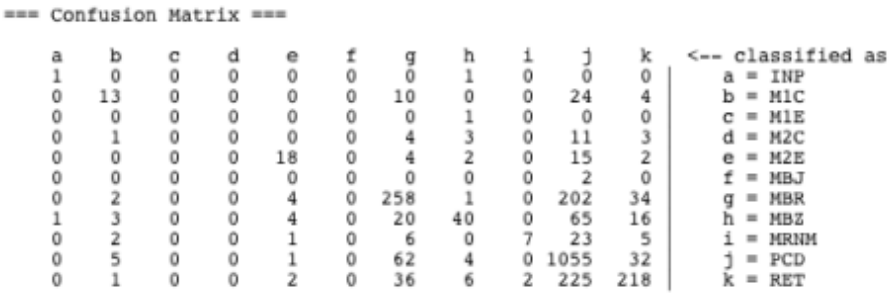
```
=== Summary ===

Correctly Classified Instances        1610               65.394  %
Incorrectly Classified Instances       852               34.606  %
Kappa statistic                          0.4503
Mean absolute error                      0.0866
Root mean squared error                  0.208
Relative absolute error                 68.6441 %
Root relative squared error             82.909  %
Coverage of cases (0.95 level)          98.9846 %
Mean rel. region size (0.95 level)      30.415  %
Total Number of Instances             2462
```

FIGURE 4.7: First Year Performance (English First Additional Language Training Set)

In this case the J48 algorithm is 65.4% accurate in predicting correctly the classes to which the records belong. The accuracy is lower compared to the value obtained in Figure 4.5, for ENA students. However, the moderate Kappa statistic value of 0.45 is much higher. In Section 4.9 the model produced in this experiment will be evaluated to determine its performance against the dataset consisting of unseen records. The model is contained in a file named *ENBQ5Training.model*. Figure 4.8 shows the confusion matrix produced following the execution of the J48 algorithm.

Looking at the confusion matrix of Figure 4.8, we make a few observations. For example, a total of

- 490 records belong to the RET class. Only 218 (44.5%) of these are correctly classified as RET records. Out of the remaining records, 225 (45.9%) records



```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j    k   <-- classified as
   1    0    0    0    0    0    0    1    0    0    0 |    a = INP
   0   13    0    0    0    0   10    0    0   24    4 |    b = M1C
   0    0    0    0    0    0    0    1    0    0    0 |    c = M1E
   0    1    0    0    0    0    4    3    0   11    3 |    d = M2C
   0    0    0    0   18    0    4    2    0   15    2 |    e = M2E
   0    0    0    0    0    0    0    0    0    2    0 |    f = MBJ
   0    2    0    0    4    0  258    1    0  202   34 |    g = MBR
   1    3    0    0    4    0   20   40    0   65   16 |    h = MBZ
   0    2    0    0    1    0    6    0    7   23    5 |    i = MRNM
   0    5    0    0    1    0   62    4    0 1055   32 |    j = PCD
   0    1    0    0    2    0   36    6    2  225  218 |    k = RET
```

FIGURE 4.8: English First Additional Language Confusion Matrix

are incorrectly labeled as PCD, which is even higher than the correctly classified RET records. More students who should be repeating first year are actually being permitted to proceed to second year, a serious misclassification error on the part of the algorithm.

- 149 records actually belong to the MBZ class, representing excluded students. Out of this number, 40 (26.8%) students are correctly classified as being excluded students. A serious misclassification error involves 65 (43.6%) records which are labeled as PCD, incorrectly indicating that excluded students have have passed and can proceed to second year.

In this case too, PCD is the majority class, containing 1159 records, of which 1055 (91.0%) are actual PCD records. Again, as in our previous experiment, the classification of minority class records is biased towards the majority PCD class.

Table 4.22 summarizes the J48 algorithm's classification results from the experiments we conducted using the ENA and ENB datasets.

TABLE 4.22: J48 Evaluation Summary (English Home Language and English First Additional Language datasets)

Metrics	**ENA**	**ENB**
Total Number of Instances	2123	2462
Correctly Classified Instances	1558 (73.39%)	1610 (65.39%)
Incorrectly Classified Instances	565 (26.61%)	852 (34.61%)
Kappa statistic	0.2948	0.4503
Mean Absolute Error	0.0801	0.0866
Root Mean Squared Error	0.2001	0.208
Relative Absolute Error	79.4222%	68.6441%
Root Relative Squared Error	89.2837%	82.909%
Coverage of cases (0.95 level)	97.0325%	98.9846%
Mean rel. region size (0.95 level)	27.414%	30.414%

It should be noted that the results displayed in Table 4.22 were obtained by applying the J48 algorithm against two different datasets (ENA and ENB). These two results

can therefore not be compared directly, even though the same algorithm was used in both cases. Having said this, there is however some similarity in how the J48 algorithm classified the records in each dataset. For example

- both datasets are dominated by a single class, namely PCD. As a result of this class imbalance the J48 algorithm is able in both instances to correctly classify the majority of the records belonging to the dominant, PCD class. For the ENA dataset, 98.8% of PCD records are correctly classified as PCD. In the case of the ENB dataset, 91.0% of PCD records are correctly identified as such. In short, J48 was able to classify most of the records belonging to the majority class. However, in both instances (ENA and ENB), J48 was not very successful in correctly categorizing minority class records.

From our findings it can be concluded that the J48 algorithm has been successful during the training phase in predicting the final outcome of first year BSc. students, using the Faculty of Science's minimum requirement subjects as predictor attributes. Given the limited time available to conduct this study, we considered it to be outside the scope of this study to try and address the class imbalance problem we identified from the results of our experiments. There are different approaches of addressing this problem. Identifying the most suitable approach after trying different options would be time consuming. For this reason, we felt that in order to do justice to this problem, it would be better addressed in a different, future study, where more time could be spent to address it extensively in a scholarly, rather than in an *ad hoc* manner. In Section 4.9 the performance of the two the J48 models (*ENAQ5Training.model* and *ENBQ5Training.model*) created during the training phase will be determined by applying them against unseen, test datasets records.

4.6.2 Predicting Degree Attainment (using Minimum Requirements)

There is a dire shortage in South Africa, as in many other countries, of graduates who possess qualifications and skills in critical, STEM-related disciplines. Not many first time first year students enroll in sufficient numbers in STEM disciplines. Amongst those who are accepted some of them face a number of challenges, which include difficulty in bridging the academic gap between high school and first year, as well as dropping out or failing to proceed to the next academic year. Compared to the number of students who enroll in STEM disciplines at the beginning of first year, not all of these students qualify and obtain degrees after three or four years of study. Even amongst those who graduate some of them, for a number of reasons, do not complete their studies in record time. Given the importance of STEM-related disciplines in different sectors of

our economy and society, of utmost importance is the ability to develop some knowledge and understanding of STEM students' chances of graduating in regulation time. For this reason, part of the research question we needed to answer was to determine whether the Wits Faculty of Science's minimum admission requirement subjects were suitable predictors of degree attainment.

In answering this part of our research question we prepared a file containing 1370 student records, which were collected during the period 2015-2016. We used them as the training set. The file contained the following predictor attributes: APS, ENA,ENB,MAT and PSC. These we used to predict the categorical class attribute, OUTCOME. As in the previous experiment, we used the J48 algorithm in predicting whether a first time first year student would qualify in record time (Q) or be unable to qualify in record time (NQ). Again in this case, as in Subsection 4.6.1, we used the J48 algorithm to predict degree achievement. We obtained the following results (Figure 4.9):

```
Correctly Classified Instances        1251               91.3139 %
Incorrectly Classified Instances       119                8.6861 %
Kappa statistic                          0.1321
Mean absolute error                      0.1553
Root mean squared error                  0.2787
Relative absolute error                 90.8013 %
Root relative squared error             95.4235 %
Total Number of Instances             1370
```

FIGURE 4.9: Predicting Degree Attainment (J48)

At 91.3% the classifier is highly accurate in predicting degree attainment. A smaller percentage (8.7%) of the number of records are classified incorrectly. Figure 4.10 is the confusion matrix produced by the J48 classifier.

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1241    0 |    a = Q
  119   10 |    b = NQ
```

FIGURE 4.10: Degree Attainment Confusion Matrix (J48)

The 1241 records of students who qualified in record time are all classified correctly as belonging to class Q. There are 129 records of students who will not qualify and obtain their degrees in record time (NQ). Out of this number, a very high percentage (92.2%) of them are classified incorrectly as students who have qualified, a serious misclassification error. Only 7.7% of the NQ records are classified correctly.

We saved the model which produced the output in Figure 4.10 in a file named *DegreeJ48Training.model*. We tested the model during the Evaluation Phase, in Section 4.9. Again, the dataset is biased towards the dominant majority class (Q). The J48 classifier faces a more complex task of having to accurately predict the class to which minority records should belong.

4.7 Student Characteristics and Academic Performance

There is a keen interest amongst educational researchers to investigate and understand the role and influence of student characteristics in academic environments. Amongst others, these characteristics include ethnicity, age, gender, beliefs, study programme, environment, economic and social background, admission year, and admission score. Some of the ways in which student characteristics are leveraged to obtain meaningful and insightful information from student datasets include

- predicting academic performance,
- categorizing students based on academic achievement,
- determining how students' characteristics contribute towards their persistence or dropping out of their studies,
- developing profiles of academically successful or unsuccessful students,
- identifying factors responsible for student success or failure, and
- identifying slow learners, as well as students at risk for failure.

Statistical and more recently, data mining techniques, are increasingly being used to process, analyse and generate information from the vast quantities of data collected by higher educational institutions about their students. The new discipline of Educational Data Mining, within which this study is located, involves a variety of techniques which are used to obtain insightful and meaningful information, and to generate new knowledge from educational datasets. These techniques include classification, prediction, clustering and association rule mining. We used cluster analysis and association rule mining techniques to answer the following question:

> *What is the role of student characteristics on the academic performance of first time first year BSc. students enrolled in STEM disciplines at Wits University?*

In Subsection 4.7.1 we describe the approach we followed in answering this question, by using a cluster analysis technique. We also used a association rule mining technique, which we discuss in Subsection 4.7.2.

4.7.1 Cluster Analysis

Cluster analysis can be used to describe the general characteristics of data objects. Clustering is the process through which data objects are partitioned into homogeneous groups consisting of similar objects. A *cluster* consists of similar data objects, which are different from data objects contained in other clusters. Once the objects have been clustered, their general characteristics as cluster members become easier to study and analyze. There are widespread applications of cluster analysis. Application areas include image recognition, outlier detection, business, biology, medicine, psychology and pattern recognition. In the field of education, Educational Data Mining (EDM) has emerged as a relatively new discipline, whose primary focus is to extract information from student datasets through different data mining methods. For example, clustering techniques can be used to separate students into different, homogeneous groups, using their characteristics as the basis for clustering. We used cluster analysis to investigate the relationship between first year BSc. students' characteristics and their academic results.

To address the above research question we used a dataset of 6703 records, which were collected during the period 2015-2019. From this dataset we selected 13 attributes, which are the characteristics we were going to use in our cluster analysis experiments. These attributes are described in Table 3, Section B of the appendix. From the WEKA toolkit, we selected and applied the *SimplekMeans* clustering algorithm against our first year BSc. dataset. Several clusters were generated, from which we were able to obtain meaningful information about the first time first year STEM students in our faculty. Before we could conduct our clustering experiments, we first had to determine how many clusters we needed to use.

Estimating the Number of Clusters

We used the *"elbow method"* to determine how many clusters we needed to use in our experiment. The method is one of several techniques used by data mining practitioners to estimate how many clusters to use when performing cluster analysis tasks. The elbow method constructs a graph, also referred to as an *"elbow graph"*, in which the *Within cluster sum of squared errors* (WSS) measure is plotted against a range of numbers. Each number (**k**) represents the number of clusters to be used. In order to determine how many clusters we needed to use we performed the following steps:

1. Using the WEKA toolkit and the dataset we prepared for cluster analysis, we executed the *SimpleKMeans* clustering algorithm using different values of k, where k=1...20.

2. For each value of k, a corresponding value of WSS was calculated. In Section C, Table 4 of the appendix, different WSS values are displayed, along with their corresponding values of k.

3. We then plotted a graph of WSS values against corresponding values of k. We obtained the following graph (Figure 4.11):

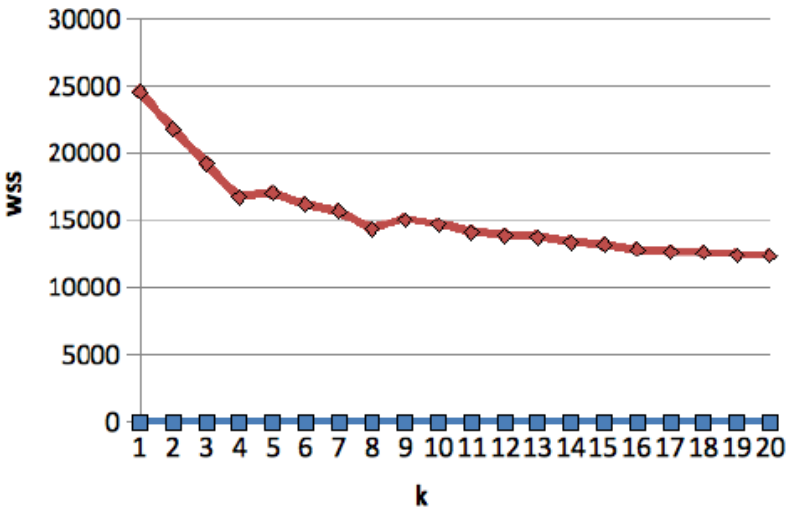


FIGURE 4.11: Estimating the best value of k

4. In Figure 4.11 the point where the graph bends and WSS decreases steadily as k increases, represents the optimal value of k, where k=9. This is the value we decided would represent the number of clusters we were going to specify for the WEKA-based *SimpleKMeans* algorithm to generate, and for us to analyse.

Having determined the optimal number of clusters we were going to use, we proceeded to apply the *SimpleKMeans* algorithm against our prepared dataset. We decided that in order to obtain more information from the cluster analysis, we would also use values of k in the neighborhood of the optimal value (k=9) obtained earlier. We used the following values: k=8 and k=10.

In the following subsections we provide an interpretation of our clustering results.

Using Eight Clusters (k=8)

Figure 4.12 shows the resulting output after executing the *SimpleKMeans* algorithm for k=8.

Cluster centroids:

		Cluster#							
Attribute	Full Data	0	1	2	3	4	5	6	7
	(6700)	(493)	(241)	(1024)	(669)	(1603)	(1045)	(909)	(716)
Year	2016.7355	2017.3631	2016.751	2016.0645	2016.0852	2016.7224	2016.5837	2017.5347	2017.102
FUNDING_STATUS	Funded	Funded	NonFunded	NonFunded	Funded	Funded	Funded	Funded	Funded
Gender	M	F	F	F	F	M	M	F	F
Race	BL	BL	BL	BL	BL	BL	BL	BL	IN
Language	EN	SO	XH	EN	SE	SO	EN	ZU	EN
APS	44.3921	44.002	41.5394	41.5811	45.2511	44.4498	46.4105	43.4059	47.0154
MAT	71.7413	70.6207	56.9585	58.4795	71.0075	73.378	82.2641	72.1826	77.5587
PSC	69.3072	68.7728	53.5353	52.2207	70.5994	73.1354	77.2842	73.1485	73.1229
Progress_outcome	PCD	PCD	MBR	PCD	PCD	PCD	PCD	RET	PCD
ENGLISH_TYPE	ENB	ENA	ENB	ENA	ENB	ENB	ENA	ENB	ENA
ENGLISH	67.7676	67.3002	60.5021	59.251	68.5097	65.078	72.9885	72.2684	74.7095
PLANCODE	GEN0	GEN0	BIOL	GEN0	GEOL	GEN0	COMM	GEN0	GEN0

FIGURE 4.12: Using Eight Clusters

From the clusters in Figure 4.12 we see that our first year BSc. students are enrolled in different STEM disciplines, which include the following: General BSc. (GEN0), and BSc. in the fields of Biological Sciences (BIOL), Geological Sciences (GEOL), Computing and Mathematics (COMM).

By interpreting and profiling these clusters we obtained the following, interesting information:

1. **GEN0 Clusters**

   Out of a total population of 6700 records, 4745 (70.8%) of these records are divided amongst five GEN0 clusters, namely clusters 0,2,4,6 and 7.

   - Cluster 0 and Cluster 2

     These two clusters contain records of funded (Cluster 0) and non-funded

(Cluster 2) students, who are academically successful (PCD outcome). One would expect that students whose home language is English and who also take the English Home Language (ENA) subject in high school, should perform better in their high school subjects compared to students whose home language is not English. Interestingly, even though students in Cluster 0 do not speak English at home, their overall performance (including ENA) is much better compared to students in Cluster 2, whose home language is English. For future first year students in need of funding and whose profile is similar to students in Cluster 0, funding should be provided, since these students are most likely to pass the first year of their GEN0 degree studies.

- Cluster 4 and Cluster 6

  These clusters represent funded students. Their home language is not English. At high school they took English First Additional Language (ENB) as their English subject. These students performed at almost the same level (APS, MAT, PSC and ENB). However, looking at the final outcome, we see that students in Cluster 4 are successful (PCD), whereas those in Cluster 6 have to repeat (RET), despite the high level of performance exhibited by students in this cluster. This is somewhat surprising. Cluster 4 is a male-dominated cluster (M) and Cluster 6 is female-dominated (F). A future study could investigate whether there is a significant difference amongst male and female students, in the overall outcome of first year students who register for the general BSc. degree (GEN0).

- Cluster 7

  This cluster represents the best performing students amongst all the GEN0 students. They receive funding and at home they speak English. At high school they did the English Home Language (ENA) subject. In all their subjects, they perform above the average, population results. Apart from Cluster 5, the students in Cluster 7 perform much better in all subjects, including the APS, compared to students in other clusters. Given their funded status and their excellent results, more funding should be provided to support entering students of a similar profile, who are most likely to be equally successful in their studies.

2. **BIOL Cluster**

- This is Cluster 1. It contains the smallest number of records. This is a low performance cluster representing mainly female, non-funded students whose home language is not English, and who took the English First Additional Language (ENB) subject in high school. Compared to the ENB scores in other clusters, students in Cluster 1 have obtained the lowest ENB mean score (60.5%), an indication perhaps that the language of instruction may be a barrier in these students' mastery or understanding of the content delivered to them in English, in their first year BIOL classes. This English language "problem" may be responsible for these students' poor overall outcome (MBR). They have to return to repeat first year, after their exclusion status has been waived. Also, these students perform in all respects below the overall average scores (APS, MAT,PSC and ENB). Students in Cluster 2 share similar characteristics with students in Cluster 1, in terms of demography and academic scores (APS, MAT and PSC). Interestingly, even though students in Cluster 2 obtain the lowest average ENA score (59.3), they still pass their first year (PCD). Could Cluster 2 students have been advantaged by the fact that they took the ENA subject at high school instead of ENB, or is it the different content offered in BIOL and GEN0 that differentiates between the overall outcome of Cluster 1 (BIOL) and Cluster 2 (GEN0) students? At this point we do not have answers to these questions. However, it may be a good idea for our faculty to look at the possibility of setting a threshold, entry score for the ENB subject, just in case English as a medium of instruction in BIOL classes poses a challenge for students whose home language is not English and who also took ENB at high school.

3. **GEOL Cluster**

- This is Cluster 3. The students are funded for their first year studies. They do not speak English at home, and at high school they took ENB as their English subject. These are very successful students, whose performance is only eclipsed by students in Cluster 5 and Cluster 7. Similar to the students in these two clusters, GEOL students are above-average students, obtaining higher than the average score of the overall student population (in APS, MAT, PSC and ENB). Prospective GEOL students of a profile similar to Cluster 3 students, and who receive funding, stand a better chance of doing well in their first year.

4. **COMM Cluster**

- This is Cluster 5. The students are funded, home language speakers of English. In terms of performance, they are as good as the GEN0 students in Cluster 7. These are high performers and like Cluster 7 students, their performance is higher than the overall average of the student population. Students with COMM degree qualifications are in great demand. Given the very high performance of these students, entering students whose characteristics are similar to Cluster 5 students should be funded, since it is highly likely that they will succeed in their first year.

Using Nine Clusters (k=9)

From Figure 4.11 we were able to determine how many clusters we were going to usein our experiment. The nine clusters generated by the SimpleKMeans algorithm are displayed in Figure 4.13.

Cluster centroids:

		Cluster#								
Attribute	Full Data	0	1	2	3	4	5	6	7	8
	(6700)	(482)	(228)	(1046)	(636)	(1336)	(1006)	(784)	(710)	(472)
Year	2016.7355	2017.3651	2016.8377	2016.0679	2016.1085	2016.9259	2016.5358	2017.5612	2017.1239	2016.2987
FUNDING_STATUS	Funded	Funded	NonFunded	NonFunded	Funded	Funded	Funded	Funded	Funded	Funded
Gender	M	F	F	F	F	M	M	F	F	M
Race	BL	BL	BL	BL	BL	BL	BL	BL	IN	BL
Language	EN	SO	XH	EN	SE	SO	EN	ZU	EN	SE
APS	44.3921	43.8942	41.7456	41.5402	45.3208	44.7777	47.2117	43.4515	46.6704	42.2818
MAT	71.7413	69.7905	59.8772	58.4962	71.1258	76.3451	84.5567	71.7041	76.6254	62.0169
PSC	69.3072	67.8423	56.7061	51.8728	70.783	76.5569	79.839	72.8648	72.1521	60.3814
Progress_Outcome	PCD	PCD	MBR	PCD	PCD	PCD	PCD	RET	PCD	RET
ENGLISH_TYPE	ENB	ENA	ENB	ENA	ENB	ENB	ENA	ENB	ENA	ENB
ENGLISH	67.7676	66.2012	63.7149	59.1663	69.0456	67.3211	74.9135	72.8316	73.9437	56.9958
PLANCODE	GEN0	GEN0	BIOL	GEN0	GEOL	GEN0	ACSI	GEN0	GEN0	GEN0

FIGURE 4.13: Using Nine Clusters (Optimal k=9)

On the whole, the nine clusters generated in Figure 4.13 are very much similar to the clusters obtained in Figure 4.12. However, Cluster 5 is now more "sharply" defined, as a ACSI cluster, representing students enrolled in the highly competitive and much sought-after BSc. degree in Actuarial Science (ACSI). For a student to be permitted to study for this degree, the following requirements must be met: Mathematics ($\geq$80%), Physical Sciences ($\geq$80%), English ($\geq$80%), and APS ($\geq$40). Given these requirements, only the very best students are accepted into the ACSI programme. The output in Cluster 5 matches and confirms these requirements in terms of the APS (47.2), MAT (84.6%), PSC (79.8%) and ENA (74.9%). The students represented in Cluster 5 receive

funding for their studies, speak English at home and ENA is their English subject at school.

Using Ten Clusters (k=10)

Figure 4.14 shows the output after executing the k-means algorithm using 10 clusters.

Cluster centroids:

Attribute	Full Data (6700)	Cluster# 0 (491)	1 (246)	2 (539)	3 (630)	4 (1364)	5 (862)	6 (781)	7 (629)	8 (477)	9 (681)
Year	2016.7355	2017.3238	2016.6707	2016.0056	2016.1079	2016.9296	2016.2483	2017.5595	2017.2576	2016.3061	2016.5947
FUNDING_STATUS	Funded	Funded	Nonfunded	Nonfunded	Funded	Funded	Funded	Funded	Funded	Funded	Nonfunded
Gender	M	F	F	F	F	M	M	F	F	M	F
Race	BL	BL	BL	BL	BL	BL	BL	BL	IN	BL	WH
Language	EN	SO	XH	EN	SE	SO	EN	ZU	EN	SE	EN
APS	44.3921	43.778	41.7561	41.5158	45.3492	45.0982	45.7854	43.4558	48.4849	42.2306	42.8076
MAT	71.7413	72.6191	60.5528	39.7588	71.6206	77.9919	79.21	71.9872	83.1399	61.8428	74.7254
PSC	69.3072	71.2159	57.0163	34.7365	71.2476	78.2001	74.4722	73.1716	79.5723	59.9161	66.2526
Progress_outcome	PCD	PCD	MBR	PCD	PCD	PCD	PCD	RET	PCD	RET	PCD
ENGLISH_TYPE	ENB	ENA	ENB	ENA	ENB	ENB	ENA	ENB	ENA	ENB	ENA
ENGLISH	67.7676	69.9654	64.2886	39.7347	69.5317	68.3622	71.3283	73.1037	77.5835	56.8721	74.743
PLANCODE	GEN0	GEN0	BIOL	GEN0	GEOL	GEN0	BIOL	GEN0	ACSI	GEN0	GEN0

FIGURE 4.14: Using Ten Clusters (k=10)

Most of the clusters in Figure 4.12 and Figure 4.13 are the same as the clusters in Figure 4.14. However, three clusters in Figure 4.14 stand out, namely Cluster 1, Cluster 5 and Cluster 7.

1. **Cluster 1 and Cluster 5**

   - These are both BIOL clusters. Cluster 1 students are non-funded, their home language is not English and in high school they took ENB as their English subject. On the other hand, Cluster 5 students receive funding, their home language is English, and in high school they took ENA as their English subject. In all respects, Cluster 1 students perform below the overall population average (APS, MAT, PSC and ENB), and fail to proceed to second year (MBR outcome). On the contrary, Cluster 5 represents above average students, who go on to pass their first year (PCD outcome). Given the different results between these two groups of students, the challenge facing our faculty is how to deal with entering students whose characteristics are similar to students in Cluster 1. It would seem that from their high school results, these may be students at risk for failure, in which case proactive steps may need to be taken early on in first year to provide them with all types of academic support and counselling. On the other hand, entering students with characteristics similar to Cluster 5 students should be funded in greater numbers. They are very good students, whose performance is as strong as those in Cluster 7.

2. **Cluster 7**

   - The students in this cluster are enrolled in the highly competitive ACSI degree programme. They are funded, speak English at home and in high school they took ENA as their English subject. Students in this cluster outperform students in all the other clusters, in all subjects, including the APS. They are by far, above-average students, who successfully complete their first year and proceed to second year (PCD outcome). Entering students, whose level of performance in high school is similar to that of students in this cluster, should be encouraged to enroll in this critical-skill ACSI degree programme, in which they are most likely to perform very well.

4.7.2 Association Rule Mining

Association rule mining (ARM) focuses on discovering interesting relationships among dataset attributes or characteristics. The purpose is to discover rules that depict relationships between attributes. Association rules are used in different domains, which include text mining, bio-informatics, telecommunications, retail industry and medicine. Tertiary institutions store vast amount of data about their students. From this data, interesting relationships among data attributes can be discovered by using association rule mining techniques, yielding information that could contribute towards our understanding and knowledge of some of the factors that influence students' academic prospects.

In this study we used the ARM technique to discover interesting relationships among the attributes in our dataset. For our association analysis experiment, we used a dataset of 5574 student records and the following student characteristics: funding status, gender, race, final outcome and English subject (ENA or ENB). To analyse the data we used the *Apriori* algorithm, which is implemented in the WEKA toolkit.

Figure 4.15 shows the WEKA GenericObjectEditor window which we used to set different parameter values.

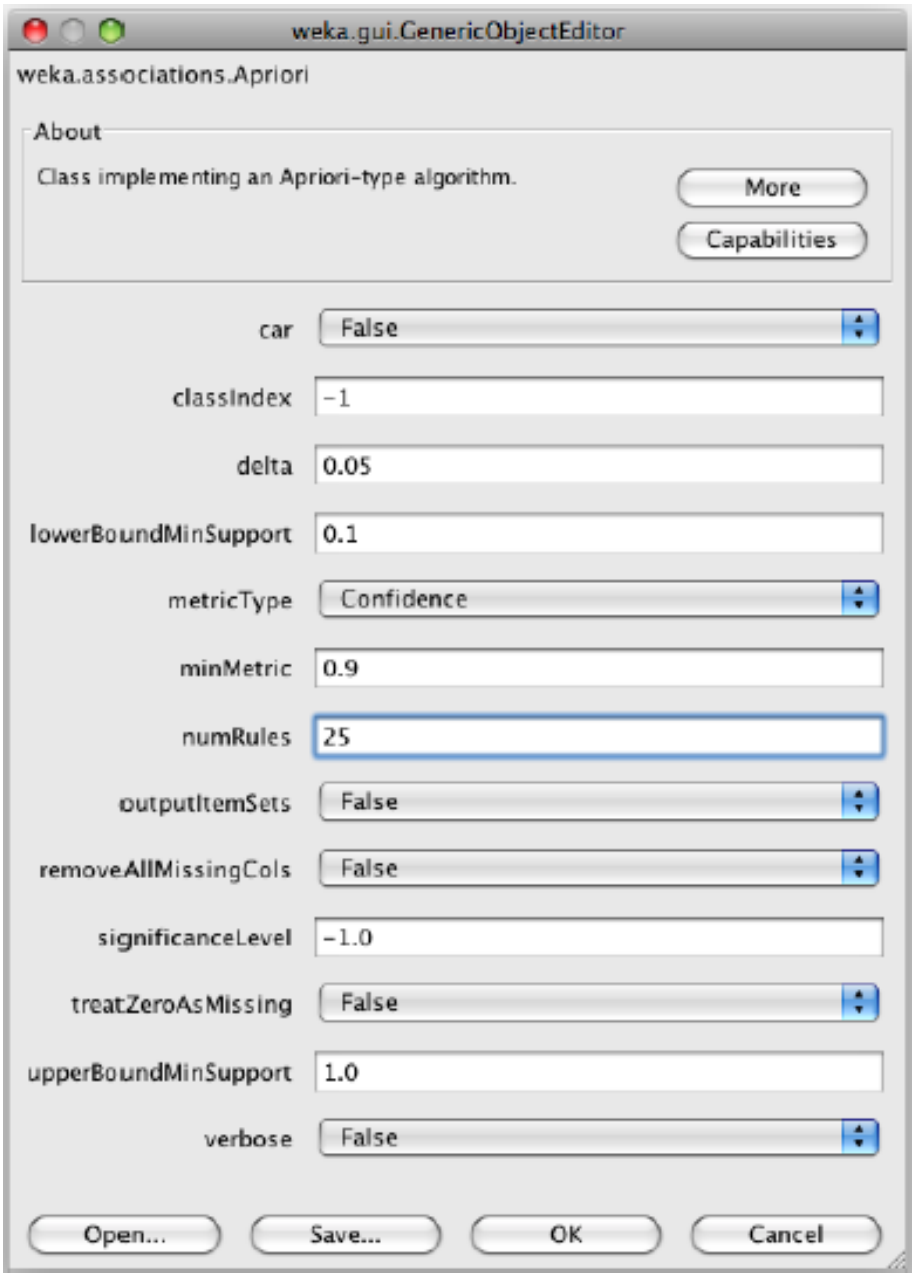


FIGURE 4.15: WEKA Object Editor Window

For the experiment we set the following parameters:

1. **lowerBoundMinSupport=0.1**
    - The minimum *support* threshold

2. **delta=0.05**

    - This is the fraction of support that is reduced iteratively from the maximum support level (**upperBoundMinSupport**) towards the minimum support level (**lowerBoundMinSupport**).

3. **metricType**

    - This parameter can be set to any one of the following metrics: *Confidence, Lift, Leverage* and *Conviction*. In the experiment we selected the *Confidence* metric.

4. **minMetric=0.9**

    - This is the minimum value of the confidence metric which we selected amongst the **metricType** parameter options.

5. **numRules=25**

    - This is the specified maximum number of rules to be generated.

6. **upperBoundMinSupport=1**

    - This is the maximum support threshold.

Before setting the minimum support threshold to 0.1, we tried several values, starting from 0.5 and gradually decreasing the values. We did not obtain a sufficient number of rules in order for us to perform a meaningful analysis. Only after setting the minimum support threshold to 0.1 did we obtain more than just a few rules.

After setting the relevant parameters we applied the *Apriori* algorithm against the dataset we wanted to analyse. In Section C.2 of the appendix, Figure 2 shows the 58 rules we obtained after executing the *Apriori* algorithm in WEKA (support=0.01, confidence=0.9). From this list we selected a few rules which we considered, within our context, to be important and interesting. The rules we selected are listed in Table 4.23.

These rules are *rare*. In order to obtain them we used a much lower support threshold value of 0.01 (1%), and five attributes (funding status, gender, home language, outcome and study plan). Since these rules satisfy the specified support and confidence threshold values, they are referred to as *strong* rules. The lift measure is derived from the support measure and the confidence measure. Given a rule X==>Y, we refer to X as the antecedent of the rule and Y as the consequent of the rule. The lift measure determines the *correlation* between the left hand side and the right hand side of a rule. When the lift value is greater than 1 the left hand side and the right hand side are *positively* correlated, which means that the left hand side and the right hand side occur more frequently together. If the value is less than 1, the rule's left hand side and right hand

TABLE 4.23: Selected Rules (Five Attributes)

Rules	Conf	Lift
8. Funding=Funded, Gender=M, Lang=EN, Plan=STAT 66 ==> Outcome=PCD 66	1	1.57
9. Funding=Funded, Plan=BIOL 86 ==> Outcome=PCD 84	0.98	1.54
12. Funding=Funded, Plan=HPHS 69 ==> Outcome=PCD 67	0.97	1.53
13. Plan=STAT 167 ==> Funding=Funded 162	0.97	1.36
14. Plan=STAT 167 ==> Funding=Funded, Outcome=PCD 162	0.97	2.09
19. Gender=F, Lang=EN, Outcome=PCD, Plan=ACSI 80 ==> Funding=Funded 77	0.96	1.35
20. Plan=MICR 77 ==> Outcome=PCD 74	0.96	1.51
34. Funding=Funded, Gender=F, Lang=EN, Plan=ACSI 81 ==> Outcome=PCD 77	0.95	1.49
38. Funding=Funded,Gender=M,Lang=EN,Plan=ACSI 148 ==> Outcome=PCD 140	0.95	1.49
42. Gender=M, Plan=PHYS 71 ==> Outcome=PCD 67	0.94	1.48
43. Gender=M, Lang=EN, Plan=STAT 70 ==> Funding=Funded 66	0.94	1.32
44. Gender=M, Lang=EN, Plan=STAT 70 ==> Funding=Funded, Outcome=PCD 66	0.94	2.03
46. Funding=non-funded, Plan=HPHS 84 ==> Outcome=PCD 79	0.94	1.48
49. Plan=GENE 89 ==> Outcome=PCD 83	0.93	1.47
51. Gender=M, Lang=ZU, Outcome=MBR 100 ==> Funding=Funded 93	0.93	1.3
56. Gender=F, Plan=GENE 70 ==> Outcome=PCD 64	0.91	1.44
57. Lang=ZU, Outcome=PCD, Plan=GEN0 111 ==> Funding=Funded 101	0.91	1.27
58. Gender=M, Lang=ZU, Outcome=PCD 322 ==> Funding=Funded 290	0.90	1.26

side are negatively correlated, in which case the left hand side occurs more often than the right hand side. If the lift value is 1 the antecedent and consequent are *independent*, meaning that they are not correlated in any way. In other words, the antecedent and consequent occur together by chance. Together the support, confidence and lift metrics are used to measure the strength of an association rule.

Looking at some of the rules in Table 4.23, we note the following:

- **Rule 8** is a high confidence rule which indicates that of the funded male students who study actuarial science and whose home language is English, all of them (100%) proceed to the next year of study. The lift value for this rule indicates that such students are 1.57 times more likely to proceed to the next academic year.
- The lift for **Rule 42** indicates that male students who study Physics are 1.48 times more likely (94% confidence) to succeed and proceed to the next academic year.
- We obtain useful information when we compare **Rule 13** and **Rule 14**. For example the lift of

  Plan=STAT==>Funding=Funded (Rule 13)

  is 1.36 (Confidence=97%). Including the **Outcome** attribute results in the rule

  Plan=STAT==>Funding=Funded, Outcome=PCD (Rule 14)

whose lift value is higher, at 2.09 (Confidence=97%). Whilst from **Rule 13** we get the idea that Actuarial Science students are likely to receive funding for their studies and more importantly, **Rule 14** suggests that these students will most likely pass and proceed to the next academic year, a fact that is supported by the high lift value that points to a strong correlation between enrolling in the STAT programme and academic success (PCD). Based on this relationship, incoming funded students with academic potential can be advised by relevant stakeholders to enroll in this STEM field, given the high possibility (no guarantees) of success in this field.

4.8 Mathematics I and Academic Performance

Mathematics features very prominently in almost all of the 13 structured curricula options on offer to first time first year STEM degree students. For these students it is compulsory for them to include Algebra 1 (MATH 1034, 15 points) and Calculus I (MATH 1036, 21 points) amongst the courses specified for their field of study. Given the prominent role played by Mathematics in the first year curriculum we were keen to investigate its relationship with pre-university (NSC) Mathematics subjects, including its predictive power on the overall outcome in first year.

The research question we posed for a detailed investigation of this issue, is as follows:

> *To what extent does Mathematics I (Algebra and Calculus) predict the academic outcome and degree attainment for first year students enrolled in STEM disciplines?*

We begin, in Subsection 4.8.1, by looking at the relationship between pre-university Mathematics subjects (NSC Mathematics and NSC Mathematics Paper 3) and first year Mathematics subjects (Algebra I and Calculus I). The aim was to investigate the relationship (correlation) between these two groups of Mathematics subjects. As part of this correlation analysis we also included the Admission Point Score (APS). In Subsection 4.8.2 we present our findings on the use of Mathematics I as a predictor of the overall, first year outcome. Equally important is the ability to determine whether a student will qualify and obtain a degree at the end of their academic programme. Again, we used Mathematics I as a predictor attribute. Our findings are presented in Subsection 4.8.3.

4.8.1 Relationship Between Mathematics I and NSC Mathematics

We conducted a multiple correlation analysis experiment, to investigate pre-university Mathematics subjects and their relationship to first year Mathematics courses. In this study we used the following variables: MATH1036 (Calculus I), MATH 1034 (Algebra I), MAT (NSC Mathematics), MP3 (NSC Mathematics Paper 3) and APS (Admission Point Score). Figure 4.16 shows the multiple correlation results produced by the SAS Enterprise software package when 1786 first year student records were analysed.

Pearson Correlation Coefficients, N = 1786 Prob > |r| under H0: Rho=0					
	MATH1036	MATH1034	MAT	MP3	APS
MATH1036	1.00000	0.90571	0.21246	0.00555	0.47658
		<.0001	<.0001	0.8148	<.0001
MATH1034	0.90571	1.00000	0.21956	-0.00878	0.43667
	<.0001		<.0001	0.7107	<.0001
MAT	0.21246	0.21956	1.00000	0.03344	0.30443
	<.0001	<.0001		0.1578	<.0001
MP3	0.00555	-0.00878	0.03344	1.00000	0.01348
	0.8148	0.7107	0.1578		0.5690
APS	0.47658	0.43667	0.30443	0.01348	1.00000
	<.0001	<.0001	<.0001	0.5690	

FIGURE 4.16: Multiple Correlation (Mathematics)

Looking at this output we observe the following

- The correlation between the two Mathematics I courses, namely MATH1036 and MATH1034, is positive, strong (r=0.91) and significant (α <.0001).
- The correlation between MATH1036 and NSC MAT is positive, weak (r=0.21) and significant (α <.0001).
- Whilst it is positive, the correlation between MATH1036 and NSC MP3 is very weak (r=0.00555) and insignificant (α=0.8148).
- The correlation between MATH1036 and the APS is positive, moderate (r=0.48) and significant (α <.0001).
- The correlation between MATH1034 and NSC MAT is positive, weak (r=0.22) and significant (α <.0001).
- Whilst it is negative, the correlation between MATH1034 and NSC MP3 is very weak (r=-0.00878) and insignificant (α=0.7107).

- The correlation between MATH1034 and the APS is positive, moderate (r=0.44) and significant (α <.0001).
- The correlation between NSC MAT and NSC MP3 is positive, weak (r=0.03) and insignificant (α=0.1578).
- The correlation between NSC MAT and the APS is positive, weak to moderate (r=0.30) and significant (α <.0001).
- The correlation between NSC MP3 and the APS is positive, weak (r=0.01) and insignificant (α=0.5690)

These results highlight a few issues relevant to our context. For instance, for certain applicants whose NSC MAT scores are just below the required threshold for acceptance into their preferred fields of study, the faculty has often used performance in NSC Mathematics Paper 3 as an extra criterion in deciding whether to accept or reject these students. However, given the poor and insignificant correlation between NSC Mathematics Paper 3 on the one hand and MATH1034, MATH1036 and NSC MAT, on the other, the usefulness of this approach in our context stands to be questioned. The weak and insignificant relationship between the NSC MAT and NSC MP3 subjects may point to the need for the content in the NSC MP3 subject to be revised and amended, to align it closely with the content in the MAT subject. Such a change may result in a stronger and significant relationship between these two NSC Mathematics subjects. The weak and insignificant correlation between NSC MP3 and Mathematics I may be due to the fact that the NSC MP3 subject content is not aligned with or relevant to Mathematics I. It is worth noting that few NSC students choose the NSC MP3 subject compared to the majority, for whom the NSC MAT subject is compulsory.

4.8.2 Predicting First Year Academic Performance

Mathematics plays a very important role in the different options of the BSc. degree offered by our faculty. Given its importance in our STEM curricula, Mathematics I could be used in our context to predict the overall, first year outcome. Having decided that we wanted to conduct an in-depth investigation into this issue, we first had to perform the following tasks:

- We prepared a *.arff* data file containing the following attributes: Year, FundingStatus, MATH1036, MATH1034 and Progress_outcome. We used the 1184 records for training purposes.
- We selected the following WEKA-based data mining algorithms, which we applied against the training set: JRIP, J48, NB and MLP.

Table 4.24 shows the output produced by the different algorithms we applied against the training set.

TABLE 4.24: Mathematics I - Predicting First Year Overall Result

	JRIP	**J48**	**NB**	**MLP**
Correctly Classified (%)	73.65	**79.56**	72.72	76.27
Incorrectly Classified (%)	26.35	**20.44**	27.28	23.73
Kappa Statistic	0.52	**0.66**	0.56	0.60
Mean Absolute Error	0.08	**0.05**	0.07	0.06
Root Mean Squared Error	0.20	**0.17**	0.18	0.17

The J48 algorithm has in all respects outperformed the other three algorithms. It has the highest accuracy (79.56%), the highest and substantial Kappa Statistic value (0.66) and the lowest error values. Figure 4.17 shows the J48 confusion matrix.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   <-- classified as
 649   0  16   7   0   0   0   0   0   0   0 |   a = PCD
  13  67  47   5   0   0   0   2   0   0   0 |   b = MBR
  39  23 166   5   1   0   0   0   0   0   0 |   c = RET
   2  14   5  48   0   0   0   1   0   0   0 |   d = MBZ
   1   5  16   3   5   0   0   1   0   0   0 |   e = M1C
   0   0   0   4   0   0   0   0   0   0   0 |   f = MRNM
   1   7   2   6   0   0   2   0   0   0   0 |   g = M2C
   0   3   6   4   0   0   0   5   0   0   0 |   h = M2E
   0   0   0   2   0   0   0   0   0   0   0 |   i = M1E
   0   0   0   0   0   0   0   0   0   0   0 |   j = MBJ
   0   0   0   0   1   0   0   0   0   0   0 |   k = INP
```

FIGURE 4.17: Confusion Matrix - Predicting First Year Outcome

These results are summarized in Table 4.25

TABLE 4.25: Summary Statistics - Predicting First Year Overall Result

1 **Class**	2 **Instances Per Class**	3 **Instances Per Class** (%)	4 **Correctly Classified Instances**	5 **Correctly Classified Instances** (%)	6 **Incorrectly Classified Instances**	7 **Incorrectly Classified Instances** (%)
PCD	672	56.8	**649**	**96.6**	23	3.4
MBR	134	11.3	**67**	**50.0**	67	50.0
RET	234	19.8	**166**	**70.9**	68	29.1
MBZ	70	5.9	**48**	**68.6**	22	31.4
M1C	31	2.6	5	16.1	**26**	**83.9**
MRNM	4	0.3	0	0.0	**4**	**100.0**
M2C	18	1.5	2	11.1	**16**	**88.9**
M2E	18	1.5	5	27.8	**13**	**72.2**
M1E	2	0.2	0	0.0	**2**	**100.0**
INP	1	0.1	0	0.0	**1**	**100.0**
Total	1184	100.0	942		242	

Looking at Column 2 it can be seen that the classes are imbalanced. Some classes contain many instances and others contain very few instances. This class imbalance does affect algorithm performance, as in the case of the J48 algorithm we used earlier to predict the overall, first year outcome. For example, where a particular class contains many instances, the majority of these instances will be correctly classified by the algorithm as members of that class, which is the case with classes PCD, MBR, RET and MBZ (Column 4 and Column 5). Only a few instances will be classified incorrectly. On the other hand, if a class contains very few instances, the majority of these instances tend to be classified incorrectly, the algorithm seemingly being unable to classify them as belonging to the class. This can be seen for classes M1C to INP, in which very high percentages of incorrectly classified instances are evident (Column 6 and Column 7). Most, if not all of the instances are classified incorrectly. A graph such as Figure 4.18 can help us to visualize, for each class, the percentage of correctly classified and incorrectly classified instances.

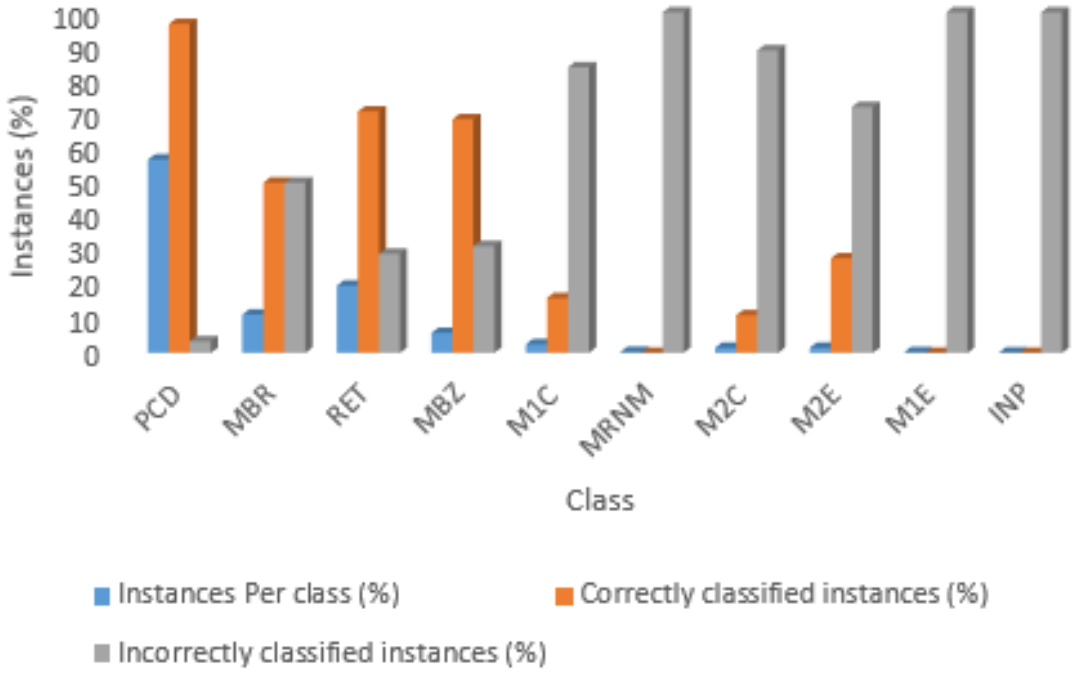


FIGURE 4.18: Correctly Classified and Incorrectly Instances

Looking at this graph we notice that when a class contains many instances, the majority of the instances will be correctly predicted by the algorithm as belonging to that class. For example, close to 60% of the instances in the dataset of 1184 records are true PCD records. The graph shows that 96.6% of these records are classified correctly by the J48 algorithm as PCD records. Very few records (3.4%) are classified incorrectly. Similarly, we can see that for the classes MBR, RET and MBZ, the majority of their records are classified correctly (MBR=50%, RET=70.9%, MBZ=68.6%). Looking at the graph, we notice that for those classes containing very few records (M1C to INP), in each case

a high percentage of records is classified incorrectly. For example, the MRNM class contains 4 records, none of which is classified as MRNM by J48. In this case 100% of the MRNM records are classified incorrectly.

In spite of our discussion above regarding some of the disadvantages associated with the performance of the J48 algorithm, the decision we took was to proceed and use this algorithm during the Evaluation Phase (Section 4.9). The high accuracy (80%), strong Kappa Statistic value (0.66) and low error rates obtained during the training phase were reason enough for us to decide to evaluate the J48 algorithm model (created during the training phase) against our test dataset. We stored the model in a file named *MathFirstYearPredict.model*. The decision tree (Figure 3) produced by J48 during the training phase can be viewed in the Appendix section, in Subsection D.1.

4.8.3 Predicting Degree Attainment

In Subsection 4.6.2 we used the minimum requirement subjects (APS, ENA, ENB, MAT and PSC) to predict degree achievement. In this section we present our findings regarding Mathematics I (Algebra and Calculus) as a predictor of degree achievement. The *.arff* file we prepared to perform this task contained 691 student records and the following attributes: Year (of first registration), FundingStatus, MATH1036 (Calculus), MATH1034 (Algebra), Progress_outcome and OUTCOME (qualification status). We used the following data mining algorithms to predict the outcome: JRIP, J48, NB and MLP.

Table 4.26 shows the output produced by the four algorithms.

TABLE 4.26: Mathematics I - Predicting Degree Achievement

	JRIP	**J48**	**NB**	**MLP**
Correctly Classified (%)	88.13	88.13	83.93	**88.56**
Incorrectly Classified (%)	11.86	11.86	16.06	**11.43**
Kappa Statistic	0.27	0.22	0.23	**0.36**
Mean Absolute Error	0.20	0.16	0.21	**0.14**
Root Mean Squared Error	0.32	0.28	0.33	**0.28**

Amongst the four algorithms, the MLP algorithm produced the best results, achieving the highest success rate and Kappa Statistic values, and committing the least errors. Below (Figure 4.19) is the MLP confusion matrix.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 584  12 |   a = Q
  67  28 |   b = NQ
```

FIGURE 4.19: Degree Attainment Confusion Matrix (MLP)

This algorithm has achieved a high accuracy rate of 88.6%. A total of 596 records belong to the Q class. Out of this number 584 (98%) records are classified correctly as Q records. This means that these 12 students, who have actually qualified for their degree, are being considered as having failed. A total of 95 records belong to the class NQ. Only 28 (29.5%) of these were correctly identified as belonging to this class. The remaining 67 (70.5%) of these records were incorrectly identified as class Q records. In this case, a very high percentage of non-qualified students are considered to have qualified and obtained their degrees. This is a very serious error, with serious implications for both students and the institution. We saved the MLP model in a file named *MathDegreePredict.model*, to be used during the Evaluation Phase (Section 4.9).

4.9 Evaluation

In this section we present our findings on our assessment of the different models we created during the Modeling Phase of the methodology we followed in our research. In the current, Evaluation Phase of the methodology, we evaluated the effectiveness of the models by applying them against different test datasets. Our findings are presented in Subsections 4.9.1, 4.9.2, 4.9.3, 4.9.4, 4.9.5, 4.9.6 and 4.9.7.

We used the following metrics to assess the models we created during the training phase (Larose, 2014) (Du, 2010): *accuracy, overall error rate, false negative rate* and *false positive rate*. In order to use these measures in evaluating the models against the test datasets, we decided that it would be more simpler and convenient to work with only two classes, namely Pass and Fail. In a case where a record's actual class label is a Pass and the record is also classified as a Pass, we refer to this as a *true positive* (TP) classification (Vipin et al., 2006) (Larose, 2014). When a Pass record is incorrectly classified as a Fail, we refer to this instance as a *false negative* (FN) classification (Vipin

et al., 2006) (Larose, 2014). A record whose class label is a Fail and which is also classified as a Fail is an example of a *true negative* (TN) classification (Vipin et al., 2006) (Larose, 2014). On the other hand, a record whose class label is a Fail and which is incorrectly classified as a Pass is an example of a *false positive* (FP) (Vipin et al., 2006) (Larose, 2014). The class labels we used in our test set records are listed in Table 4.20. The accuracy metric determines the success rate of an algorithm in terms of the number of correct classifications relative to the total number of predictions (Vipin et al., 2006).

Using a table known as a *confusion matrix* (Larose, 2014), we can determine an algorithm's performance, by looking at the number of correct classifications and the number of the incorrect classifications. Figure 4.20 shows the matrix for a model in which two classes are used.

		Predicted	
		Pass	**Fail**
Actual	**Pass**	TP	FN
	Fail	FP	TN

FIGURE 4.20: Binary Classification Confusion Matrix

The matrix can be used to derive a number of model performance metrics, which include the following (Larose, 2014) (Vipin et al., 2006): *accuracy* (Acc), *overall error rate* (OER), *true positive rate* (TPR), *false positive rate* (FPR), *true negative rate* (TNR), *false negative rate* (FNR). These metrics are defined in Equations 4.2 to 4.7.

- Accuracy (Acc)

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.2}$$

- Overall Error Rate (OER)

$$OER = \frac{FN + FP}{TP + FP + TN + FN} \tag{4.3}$$

- True Positive Rate (TPR)

$$TPR = \frac{TP}{TP + FN} \tag{4.4}$$

- False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN} \tag{4.5}$$

- True Negative Rate (TNR)

$$TNR = \frac{TN}{FP + TN} \tag{4.6}$$

- False Negative Rate (FNR)

$$FNR = \frac{FN}{FN + TP} \tag{4.7}$$

4.9.1 Influence of NSC STEM Predictor Attributes

The task we had to perform in Subsection 4.4.2 was to determine the significance of selected NSC Grade 12 STEM subjects as predictors of the APS. For this purpose, we applied a WEKA-based multiple linear regression algorithm against a training set of 38231 NSC Grade 12 student records. We stored the resulting regression model in a file named *RQ3_LinReg.model*. We then applied this model against a test dataset consisting of 116 NSC Grade 12 student records, collected in 2020. Figure 4.21 shows some of the test set records whose APS was estimated using the NSC STEM subjects as predictor attributes.

In this figure the predicted APS values appear just before the last column. We assessed our regression model by comparing the predicted APS scores with the actual APS scores obtained by the students. For the six test set records shown in Figure 4.21 and a few records selected from the test set, Table 4.27 shows the predicted and actual APS scores, as well as the difference between them.

TABLE 4.27: Predicted versus Actual Admission Point Score Values

Record	**Predicted**	**Actual**	**Difference Scores (Predicted-Actual)**
1.	43	45	-2
2.	43	43	0
3.	43	44	-1
4.	44	45	-1
5.	43	43	0
6.	43	41	2
18.	44	50	-6
23.	48	46	2
69.	49	49	0
99.	45	52	-7
109.	45	45	0
114.	44	46	-2

```
@attribute LSC numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute GEG numeric
@attribute MP3 numeric
@attribute INT numeric
@attribute CAT numeric
@attribute ENA numeric
@attribute ENB numeric
@attribute APM numeric
@attribute AGS numeric
@attribute predictedAPS numeric
@attribute APS numeric

@data
77,75,81,73,85,-1,-1,-1,-1,-1,-1,43.00851,?
-1,-1,74,75,86,70,63,-1,-1,-1,-1,42.911169,?
-1,-1,74,68,85,-1,83,75,-1,-1,-1,43.496883,?
-1,-1,75,80,86,-1,70,73,-1,-1,-1,43.713938,?
74,-1,-1,82,72,-1,-1,-1,-1,-1,-1,43.300922,?
75,-1,80,69,73,-1,-1,-1,-1,-1,-1,42.657841,?
```

FIGURE 4.21: Predicting the Admission Point Score (Test Set))

Looking at the sample output in this table, it can be seen that the predicted and actual APS values are on the whole similar. We found this to be the case when considering the complete output from the 116 test dataset records. Our result in this evaluation experiment shows that the multiple linear regression model created earlier, using selected NSC STEM subjects as predictor attributes, was very successful and effective in estimating the APS. To confirm our conclusion, we first plotted a graph to investigate the association between the actual and predicted APS values. The plot we obtained is shown in Figure 4.22.

From this plot we obtain the following information:

- There is a linear connection between the actual and predicted APS values. The graph has a positive slope.
- The trend line passes through the majority of the points, with only one outlier located at the bottom of the plot.
- The actual and predicted APS values are positively correlated (R=$\sqrt{R^2}$=0.51).
- The regression equation can be used, given actual APS values (x), to predict APS values.

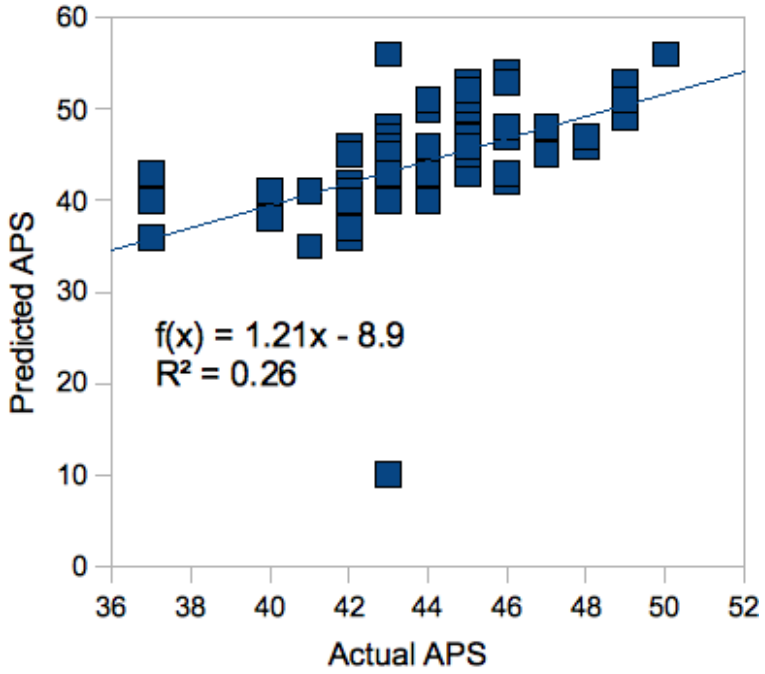


FIGURE 4.22: Actual Admission Point Score versus Predicted Admission Point Score (Test Set)

Having established that the actual and predicted APS values had a linear association, we next conducted a T-test experiment to analyse the sample of scores obtained by subtracting actual from predicted APS scores. We used these scores as our data to perform the T-test, based on the following hypotheses:

$H_0 : \mu_D = \mu_{predicted} - \mu_{actual} = 0$ (mean of difference scores equals 0)
$H_1 : \mu_D \neq 0$ (mean of difference scores is not 0)

To test the hypothesis we used the following T-test equation:

$$t = \frac{\overline{D} - 0}{\frac{s_D}{\sqrt{N}}} \tag{4.8}$$

where

- $\overline{D}$ is the mean score. In the dataset, $\overline{D}$ = -0.21,
- s_D is the standard deviation. In the dataset $s_D = 4.57$, and
- N is the total number of scores. In the test dataset, $N = 116$.

Using these results and Equation 4.8, we performed the following calculation to obtain the value of the test statistic (t):

$$\text{t}=\frac{-0.21-0}{\frac{4.57}{\sqrt{116}}} = \frac{-0.21}{0.42} = -0.50$$

The calculated value is -0.50. Since there are 116 difference scores, we have 115 (N-1) degrees of freedom. Using a t Distribution table, we obtained the following t value at the level of significance, α=0.05: $t_{.025}(115) = \pm 1.960$. The value we obtained using Equation 4.8, namely t(-0.50), is *less* than 1.960. We do not reject the null hypothesis (H_0). Our conclusion is that there is no difference between the means of predicted and actual APS values. Based on these findings, we consider our linear regression model to have been successful and effective in predicting the APS, using selected NSC STEM subjects as predictor attributes.

Focusing on the research question we posed in Section 4.4, regarding the relationship between NSC STEM attributes and the APS, a summary of our findings is as follows:

- The different feature selection experiments we conducted identified the following NSC STEM subjects as having the strongest influence in relation to the APS: Mathematics (MAT), Physical Sciences (PSC), Geography (GEG), Information Technology (INT), Advanced Programme Mathematics (APM) and Agricultural Sciences (AGS).
- The NSC STEM predictor attributes we used are all significant in their relationship with the APS. Our results also show the strength of the relationship of each attribute to the APS.
- The difference between the estimated APS, obtained using NSC STEM predictor attributes, and the actual APS, is not significant.

4.9.2 Predicting Computer Science I (BCO, IAP)

During the Modeling Phase we created a regression model (Subsection 4.5.2) which we used to predict the overall, Computer Science I score. For this purpose we used two first-semester Computer Science I courses, namely BCO (COMS1015) and IAP (COMS1018), as predictor attributes. We named the model *linreg4B.model*. Given the problems encountered by the faculty, and indeed the whole university, regarding student attendance at the height of the pandemic in 2020, we were only able to secure a small

number of test set records that we could use to evaluate the model. Figure 4.23 shows a sample of the output we obtained after applying the regression model against a test dataset consisting of 27 Computer Science I student records.

```
@relation CS1Q44_predicted

@attribute COMS1015 numeric
@attribute COMS1018 numeric
@attribute predictedCS1 numeric
@attribute CS1 numeric

@data
96,96,92.830964,?
92,96,91.251184,?
76,76,73.651015,?
64,48,53.118208,?
88,68,73.877938,?
90,91,87.641032,?
82,81,78.840948,?
68,43,51.877727,?
```

FIGURE 4.23: Predicting Computer Science I (Basic Computer Organisation, Introduction to Algorithms and Programming)

Table 4.28 shows a sample of predicted and actual APS values, including the difference between them.

TABLE 4.28: Predicted versus Actual Computer Science I Scores (Basic Computer Organisation and Introduction to Algorithms and Programming)

Record	**Predicted**	**Actual**	**Difference Scores (Predicted-Actual)**
1	93	95	-2
3	74	67	7
8	52	54	-2
16	8	4	4
17	72	79	-7
19	61	51	10
24	91	89	2
26	93	92	1

Figure 4.24 is the graph we plotted to see whether a linear relationship existed between the predicted and actual Computer Science I scores.

The predicted and actual scores have a linear relationship with one another, with the slope of the graph being very close to 1. Almost all the points are gathered around the trend line. The linear correlation between these scores is quite high, with $R=\sqrt{0.93}=0.96$. This is a near-perfect correlation. However, we do need to stress the

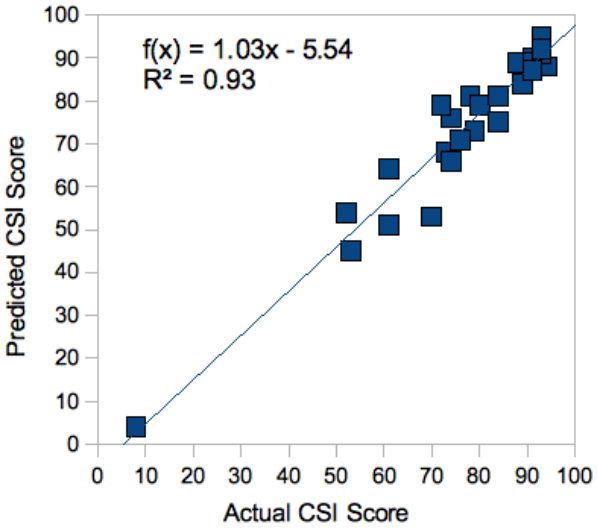


FIGURE 4.24: Predicting Computer Science I (Basic Computer Organisation, Introduction to Algorithms and Programming)

fact that we obtained this result using a small number of test dataset records that were available to us.

As in the previous section we used the difference scores as our data, from which the mean and standard deviation values were calculated. In this case the

- number of difference scores, N, is 27,
- mean score, $\overline{D}$, is 3.07, and
- standard deviation, s_D, is 5.08

Using these values and Equation 4.8, we calculated the t-statistic as follows:

$$\text{t}=\frac{3.07-0}{\frac{5.08}{\sqrt{27}}} = \frac{3.07}{0.98} = 3.13$$

The calculated value is 3.13. Using α=0.05 and 26 degrees of freedom, we obtained the following result from the t Distribution table: $t_{.025}(26) = \pm 2.06$. Since $|t_{calculated}| > t_{critical}$ $(3.13 > 2.06)$, we conclude that the difference between the actual and predicted Computer Science I scores is significant.

Our findings on predicting Computer Science I show that

- the multiple linear regression model developed during the Modeling Phase can be used to predict Computer Science I scores in test dataset records,

- in the test case, the Computer Science I scores predicted by our linear regression model correlate very strongly with the actual Computer Science I scores obtained by first year Computer Science students,

4.9.3 Predicting First Year Outcome (BCO,IAP)

In Subsection 4.5.3 we identified the J48 algorithm as the best predictor of the overall outcome amongst first time first year BSc. students. As predictor attributes, we used two first semester courses, namely BCO and IAP. In this subsection we assess the J48 model created during the experiments we conducted in Subsection 4.5.3. We used this model to classify 75 test set records, which were collected during the period 2017-2018. Figure 4.25 is an example of the *.arff* test file we used in our experiment. The unknown class labels of the records are marked using the question mark symbol ("?").

```
@relation TESTQ4

@attribute BCO numeric
@attribute IAP numeric
@attribute Progress_outcome {M2C,M2E,MBR,MRNM,PCD,RET}

@data
72,90,?
67,82,?
70,91,?
55,66,?
55,63,?
56,73,?
78,92,?
70,97,?
70,92,?
50,93,?
```

FIGURE 4.25: First Year Test Set Records (Basic Computer Organisation, Introduction to Algorithms and Programming)

Using the WEKA toolkit, we applied our J48 model against the test dataset, and obtained the output shown in Figure 4.26.

Figure 4.26 contains the following, new information:

- In the first line, the *_predicted* phrase is added to the relation name, *TESTQ4*, resulting in a new relation name *TESTQ4_predicted*.
- There is an extra attribute named *predictedProgress_outcome* whose values are as follows: *M2C,M2E,MBR,MRNM,PCD,RET*.
- In the data section, each test set record is now assigned a class label.

```
@relation TESTQ4_predicted

@attribute BCO numeric
@attribute IAP numeric
@attribute predictedProgress_outcome {M2C,M2E,MBR,MRNM,PCD,RET}
@attribute Progress_outcome {M2C,M2E,MBR,MRNM,PCD,RET}

@data
72,90,PCD,?
67,82,PCD,?
70,91,PCD,?
55,66,PCD,?
55,63,PCD,?
56,73,PCD,?
78,92,PCD,?
70,97,PCD,?
70,92,PCD,?
50,93,PCD,?
44,39,MBR,?
73,70,PCD,?
43,38,MBR,?
46,41,MBR,?
40,37,M2C,?
50,65,PCD,?
```

FIGURE 4.26: Classifying Tesd Dataset Records

Our next task was to evaluate the J48 model's effectiveness in predicting the overall, first year outcome. This we did by comparing the outcomes predicted by our model (Figure 4.26), and the actual outcomes, obtained from student records.

Figure 4.27 is the binary classification confusion matrix, which shows how the test dataset records were classified, in terms of passes and failures.

		Predicted	
		Pass	Fail
Actual	Pass	47	3
	Fail	14	11

FIGURE 4.27: First Year Outcome Binary Classification (Basic Computer Organisation and Introduction to Algorithms and Programming)

Using the confusion matrix in Figure 4.27, we calculated the following metrics:

- The *accuracy* (Acc) of the model is 77.3%, which is slightly lower compared to the 82.9% accuracy result obtained during the Modeling Phase. However, we consider the 77.3% accuracy result to be acceptable. The *overall error rate* (OER) is 22.7%, which is slightly higher compared to the 17.1% error rate value obtained during the Modeling Phase.
- The *true positive rate* (TPR) is 0.94, and the *false positive rate* (FPR) is 0.56.
- The *true negative rate* (TNR) is 0.44, and the *false negative rate* (TNR) is 0.06.

A model's performance can also be evaluated by using a *Receiver Operating Characteristics* (ROC) graph, which depicts the relationship between the true positive rate and the false positive rate. Figure 4.28 is the ROC graph indicating the performance of our J48 model against the test dataset. The graph shows the relationship between the true positive rate and the false positive rate.

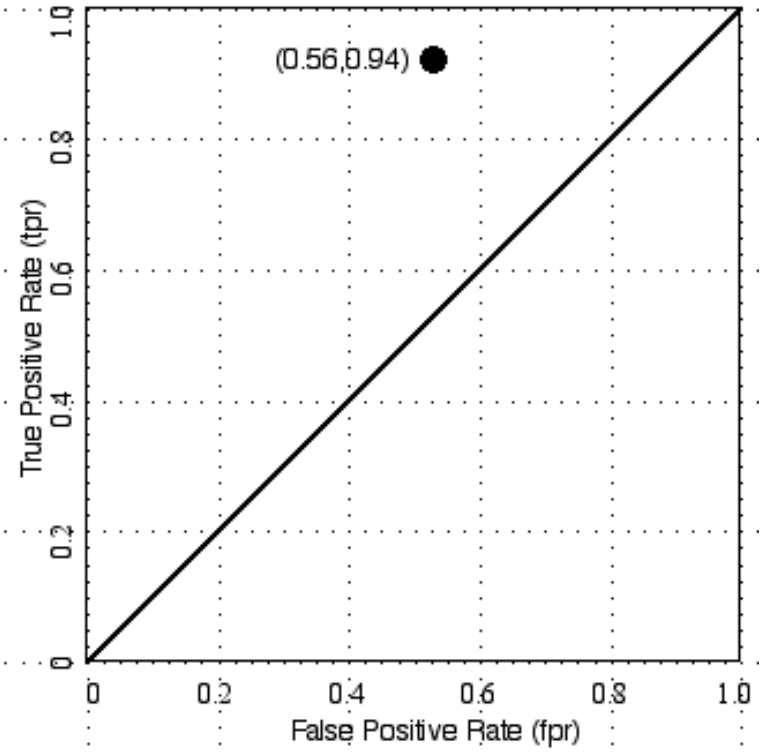


FIGURE 4.28: First Year Outcome Receiver Operating Characteristics Graph (Basic Computer Organisation and Introduction to Algorithms and Programming)

A point on the graph represents a pair of values, namely a false positive and a true positive value. Such a point indicates the performance of a classifier. For a good classifier, its point is situated closer to the top left corner of the graph. A point that is situated along the diagonal line represents a random guessing classifier. If a point is situated below the diagonal line, the classifier's performance is worse compared to that of a random guessing classifier. Figure 4.28 shows that the J48 classifier we used to predict the first year outcome has performed much better than a random guessing classifier. Based on its successful performance, the model could be used in the future to predict the overall outcome of first year STEM students who have completed the two first-semester Computer Science I courses, namely BCO and IAP.

Our findings regarding BCO and IAP as predictors of the overall, first year outcome, show that

- during the Modeling Phase, the J48 decision tree algorithm outperformed the other three algorithms (JRIP, NB, and MLP) in predicting the overall first year outcome, using BCO and IAP as predictor attributes,
- during the Evaluation Phase, the J48 decision tree model performed successfully, using BCO and IAP as predictor attributes, in predicting the overall outcome of test dataset records,
- the two first semester Computer Science I subjects, BCO and IAP, can be used in our context to predict the overall, first year outcome.

4.9.4 Predicting First Year Outcome (Minimum Requirements)

Our task in Subsection 4.6.1 was to predict the overall first year outcome using the following Wits Faculty of Science minimum admission requirement subjects as predictor attributes: the APS, NSC Mathematics, NSC Physics and NSC English. Following the experiments we conducted during the Modeling Phase, two models were obtained after applying the J48 decision tree algorithm against two different training datasets.

- The first training set consisted of 2123 records belonging to English Home Language (ENA) students. In classifying the records in this dataset, the J48 algorithm was 73.4% accurate. We stored the resulting model in a file named *ENAQ5Training.model*.
- The second training set consisted of 2462 records belonging to English First Additional Language (ENB) students. In this case the J48 algorithm was 65.4% accurate. We stored the resulting model in the file *ENBQ5Training.model*.

We evaluated each model by applying it against a corresponding test set. The ENA test set contained 483 records and the ENB test set contained 579 records. The sample *.arff* files of the two test sets can be found in the Appendix chapter, Subsection E.1 (Figure 4 and Figure 5).

First Year Outcome Prediction (ENA Students)

Figure 4.29 shows some of the ENA records whose class labels were predicted by the ENA model (*ENAQ5Training.model*).

```
@relation TESTQ5ENA_predicted

@attribute APS numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute ENA numeric
@attribute predictedProgress_outcome {INP,M1C,M1E,M2C,M2E,MBR,MBZ,MRNM,PCD,RET}
@attribute Progress_outcome {INP,M1C,M1E,M2C,M2E,MBR,MBZ,MRNM,PCD,RET}

@data
48,87,83,70,PCD,?
48,74,79,84,PCD,?
40,64,66,66,PCD,?
50,93,86,84,PCD,?
41,89,73,70,PCD,?
40,83,63,72,PCD,?
41,70,54,71,PCD,?
51,96,83,85,PCD,?
42,85,77,63,PCD,?
44,70,-1,70,PCD,?
42,71,71,80,PCD,?
46,84,83,75,PCD,?
43,60,67,72,PCD,?
50,79,90,84,PCD,?
```

FIGURE 4.29: First Year Outcome Prediction Output (English Home Language)

To obtain a much clearer picture of the J48 model's performance we decided to simplify the problem, from a multiple classification to a binary classification problem involving only two classes, namely Pass and Fail. We created the binary classification matrix shown in Figure 4.30, in which the records are classified as either Pass or Fail.

		Predicted	
		Pass	Fail
Actual	Pass	326	13
	Fail	129	15

FIGURE 4.30: First Year Outcome Prediction Binary Confusion Matrix (English Home Language Test Set)

Using the entries in Figure 4.30, we evaluated the performance of our J48 model by obtaining the following metric values:

- The *accuracy* (Acc) of the model is 70.6%, a result that compares favourably to the 73.4% accuracy result obtained during the Modeling Phase. The *overall error rate* (OER) is 29.4%. This value does not differ very much from the 26.6% error rate result obtained during the Modeling Phase.
- The *true positive rate* (TPR) is 0.9617 and the *false positive rate* (FPR) is 0.8598.
- The *true negative rate* (TNR) is 0.1042 and the *false negative rate* (FNR) is 0.0383.

We further evaluated the performance of our J48 model by constructing a ROC graph. On a ROC graph, the true positive rate (tpr) is plotted on the vertical, y axis, and the false positive rate (fpr) is plotted on the horizontal, x axis. A point,(x,y), situated on the graph represents the performance of a single classifier. A classifier whose point is located along the diagonal line joining the points (0,0) and (1,1) represents a random classifier. The ROC graph contains four critical points, described in Table 4.29.

TABLE 4.29: Receiver Operating Characteristics Critical Points

tpr	**fpr**	**Interpretation**
0	0	Model classifies every instance as negative
1	1	Model classifies every instance as positive
1	0	The ideal model
0	1	Worst classification of instance

The single point on the ROC graph in Figure 4.31 shows the performance of the model we used to predict the overall, first year outcome.

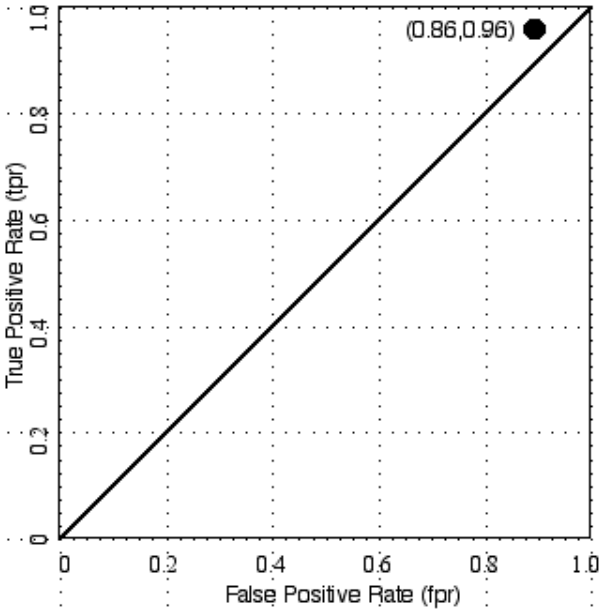


FIGURE 4.31: English Home Language Receiver Operating Characteristics Graph

Points that are located within the upper left-hand triangle, such as the point (0.86,0.96) in the figure, correspond to classifiers whose performance is considered to be better than random classifiers. On the other hand, classifiers whose points are located within the lower right-hand triangle, their performance is considered to be worse than random classifiers. Our conclusion from this result is that using the minimum admission requirements subjects, the J48 model we created during the Modelling Phase was successful in predicting the first year outcome of the majority of the test dataset records belonging

to students who took ENA as their high school English subject. It must be noted, however, that the model may need to be improved in order for it to deal with the problem of incorrectly classifying minority class records.

First Year Outcome Prediction - ENB Students

Figure 4.32 shows some of the ENB records whose class labels were predicted by the J48 model (*ENBQ5Training.model*) that was applied against a test dataset of 579 records belonging to first year students who took ENB as their high school English subject.

```
@relation TESTQ5ENB_predicted

@attribute APS numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute ENB numeric
@attribute predictedProgress_outcome {INP,M1C,M1E,M2C,M2E,MBJ,MBR,MBZ,MRNM,PCD,RET}
@attribute Progress_outcome {INP,M1C,M1E,M2C,M2E,MBJ,MBR,MBZ,MRNM,PCD,RET}

@data
43,80,96,84,RET,?
46,63,60,85,MBR,?
43,70,52,79,RET,?
40,75,79,75,MBZ,?
43,65,72,80,PCD,?
40,80,82,68,RET,?
47,83,83,78,PCD,?
43,67,70,75,PCD,?
48,83,87,73,PCD,?
43,80,83,62,PCD,?
42,60,86,75,PCD,?
49,79,90,78,PCD,?
```

FIGURE 4.32: First Year Outcome Prediction Output (English First Additional Language)

The binary classification matrix in Figure 4.33 shows how the 579 test dataset records were classified by the J48 model.

		Predicted	
		Pass	Fail
Actual	Pass	212	66
	Fail	179	122

FIGURE 4.33: First Year Outcome Prediction Binary Confusion Matrix (English First Additional Language Test Set)

The following measures were calculated using the information contained in Figure 4.33.

- The model is 57.7% accurate. This performance is lower compared to the 65.4% accuracy result obtained during the Modeling Phase. At 42.2%, the *overall error rate* (OER) is somewhat higher, compared to the error rate of 34.6% obtained during the Modeling Phase.
- The *true positive rate* (TPR) is 0.7626 and the *false positive rate* (FPR) is 0.5947.

- The *true negative rate* (TNR) is 0.4053 and the *false negative rate* (FNR) is 0.2374.

In order to determine how well the J48 model performed in predicting the overall, first year outcome, we used the true positive and false positive rates to plot the ROC graph shown in Figure 4.34.

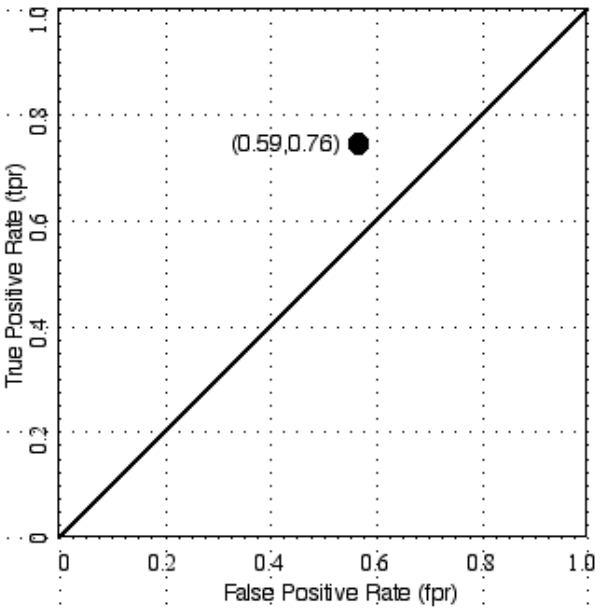


FIGURE 4.34: English First Additional Language Receiver Operating Characteristics Graph

The point on this ROC graph is located above the diagonal, indicating that the model performed better than a random classifier. We conclude from this result that the J48 model we used was successful in predicting the first year outcome when applied against a test dataset of 579 records belonging to students who took the ENB subject in high school. In other words, the minimum requirements subjects are suitable predictors of the overall, first year outcome.

Regarding the use of the minimum requirement subjects as predictors of the overall, first year outcome, our Modeling Phase and Evaluation Phase experiments revealed the following information:

- The two J48 models created during the Modeling Phase (*ENAQ5Training.model* and *ENBQ5Training.model*), performed successfully against test datasets in predicting the overall, first year result.
- The ROC graphs we plotted demonstrated the success of the two J48 models in predicting the overall first year outcome.

4.9.5 Predicting Degree Attainment (Using Minimum Requirements)

Our task in Subsection 4.6.2 was to determine the effectiveness of the Wits minimum admission requirement subjects stipulated by the Faculty of Science as predictors of degree attainment by first year students enrolled in STEM disciplines. The training set we used during the Modeling Phase consisted of 1370 first year student records, which the J48 algorithm was 91.3% accurate in classifying. To evaluate the model produced by the algorithm, we applied it against a test dataset comprising 432 records with unknown class labels (see Subsection E.2, Figure 6). A sample of the output produced by the model is shown in Figure 4.35.

```
@relation TrainQ5degAtt_predicted

@attribute FUNDING_STATUS {F,NF}
@attribute APS numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute ENA numeric
@attribute ENB numeric
@attribute 'prediction margin' numeric
@attribute 'predicted OUTCOME' {Q,NQ}
@attribute OUTCOME {Q,NQ}

@data
F,48,86,88,71,-1,0.892308,Q,?
F,53,92,97,-1,83,0.737705,Q,?
F,52,96,93,69,-1,0.892308,Q,?
F,50,92,90,78,-1,0.892308,Q,?
F,55,97,91,90,-1,0.892308,Q,?
```

FIGURE 4.35: Degree Prediction Output

One of the objectives of this study was to determine whether, for first year students, their using minimum admission requirements subjects could be used to predict whether they would qualify and obtain their degrees (Q) in record time, failing which they would not qualify and obtain their degrees in record time (NQ). Following the application of our J48 model against the test dataset, we obtained the following confusion matrix (Figure 4.36):

Using the entries in this confusion matrix, we calculated values for different measures. The model's success rate is 80.1% and its overall error rate is 19.9%. The true positive and false positive rates are quite high (TPR=0.9971, FPR=1.0000). On the other hand,

a	b	<--classified as
346	1	a=Q
85	0	b=NQ

FIGURE 4.36: Degree Attainment Confusion Matrix (using Minimum Requirements)

the true negative and false negative rates are very low (TNR=0.000, FNR=0.0029). These results show that the J48 model classifies almost all the test set records as belonging to class Q, meaning that almost all the students, except only one, have qualified and obtained their degrees. All the 85 students who have actually failed (NQ) are incorrectly classified as having qualified (Q). We conclude from this result that the J48 model we used to predict degree achievement, using the minimum admission requirement subjects, is not suitable for this purpose. This result shows that in order to evaluate the model's performance in predicting degree attainment, we cannot rely solely on the accuracy measure, however high it might be. The fact that there are 85 records which represent students who have not qualified (NQ) but who are predicted as having qualified (Q) raises a serious concern. Another issue is the highly skewed nature of the data in favour of the majority class (Q). The test dataset contains 432 records, of which 347 (80.3%) belong to the majority class and 85 (19.7%) are minority class records (NQ). However, due to the huge class imbalance, the model was not effective in identifying records of the minority, negative class (NQ). It is perhaps appropriate at this point to sound a cautionary note. Given a dataset with a huge class imbalance the accuracy metric, however high its value, can be misleading and should not be relied upon as a criterion for judging a classification model's performance.

Following our Modeling Phase and Evaluation Phase experiments our findings, regarding the use of minimum requirement subjects as predictors of degree achievement, revealed the following information:

- The training and test datasets are highly imbalanced.
- Despite the fact that high accuracy results were obtained during our Modeling and Evaluation Phase experiments, all of the 85 minority class records were incorrectly classified.
- Our Evaluation Phase result shows that, despite the 80.1% accuracy rate, the model we created during the Modeling Phase (*DegreeJ48Training.model*), was not successful in predicting the degree qualification status of first time first year STEM students.

4.9.6 Predicting First Year Outcome (using Mathematics I)

The critical role played by Mathematics I (Algebra and Calculus) in our first year curriculum cannot be overemphasized. The subject features prominently in the different STEM curricula amongst which students can select their degree programme of choice. The task we performed in Subsection 4.8.2 was to determine the effectiveness of Mathematics I as a predictor of the overall outcome at the end of first year. Following the

data mining experiments we conducted during the Modelling Phase to investigate this aspect of our research question, we found the J48 decision tree algorithm to be the best and most effective predictor of the overall, first year result. We then decided to use the model produced by the J48 algorithm *MathFirstYearPredict.model* to classify the 474 first year records contained in out test dataset (see Subsection E.3, Figure 7 in the Appendix chapter). Following our application of the model against the test dataset, Figure 4.37 is a sample of the output obtained after the model was applied against the test dataset.

```
@relation TrainQ7outcome_predicted

@attribute Year numeric
@attribute FUNDING_STATUS {F,NF}
@attribute MATH1036 numeric
@attribute MATH1034 numeric
@attribute 'prediction margin' numeric
@attribute 'predicted Progress_outcome' {PCD,MBR,RET,MBZ,M1C,MRNM,M2C,M2E,M1E,MBJ,INP}
@attribute Progress_outcome {PCD,MBR,RET,MBZ,M1C,MRNM,M2C,M2E,M1E,MBJ,INP}

@data
2017,NF,35,31,-0.990795,MBR,?
2017,F,27,50,-0.990795,RET,?
2017,F,50,55,0.996913,PCD,?
2017,F,61,80,0.990795,PCD,?
2017,F,50,50,0.598887,PCD,?
2017,F,43,37,-0.990795,RET,?
2017,F,24,31,-0.990795,M2E,?
2017,F,36,35,-0.497688,RET,?
2017,F,40,52,-0.990795,RET,?
2017,F,29,41,-0.990795,RET,?
2017,NF,0,0,-0.582882,MBZ,?
2017,F,75,64,0.995376,PCD,?
```

FIGURE 4.37: First Year Outcome Prediction Output (Mathematics I)

The confusion matrix in Figure 4.38 shows how the test dataset records were classified by our J48 model (*MathFirstYearPredict.model*).

a	b	c	d	e	f	g	h	i.	j	k	← Classified as
0											a=INP
	0										b=M1C
		0									c=M1E
			0								d=M2C
	1			0		1	2				e=M2E
					0					1	f=MBJ
	4		6	7		50	16	6	4	36	g=MBR
							2				h=MBZ
	2		1	1		3	7	0		1	i=MRNM
			1	1		3	2		227	10	j=PCD
			6	2		19	7		15	30	k=RET

FIGURE 4.38: First Year Outcome Prediction (Using Mathematics I)

Figure 4.39 is the binary confusion matrix we created by reducing the multiple classes in Figure 4.38 to two classes, namely Pass and Fail.

		Predicted	
		Pass	Fail
Actual	Pass	227	17
	Fail	19	211

FIGURE 4.39: First Year Outcome Prediction Binary Confusion Matrix (Using Mathematics I)

Using the entries in Figure 4.39, we obtained values for several metrics. The J48 model is 92.4 % accurate, a higher value compared to the 79.6% accuracy value obtained during the Modelling Phase. The overall error rate is 7.6%, much lower compared to the 20.4% value obtained during the Modelling Phase. The true positive and false positive rate values are, respectively, 0.9303 and 0.0826. The true negative and false negative rate values are, respectively 0.9174 and 0.0697.

We used the true positive and false positive rate values to plot the ROC graph in Figure 4.40.

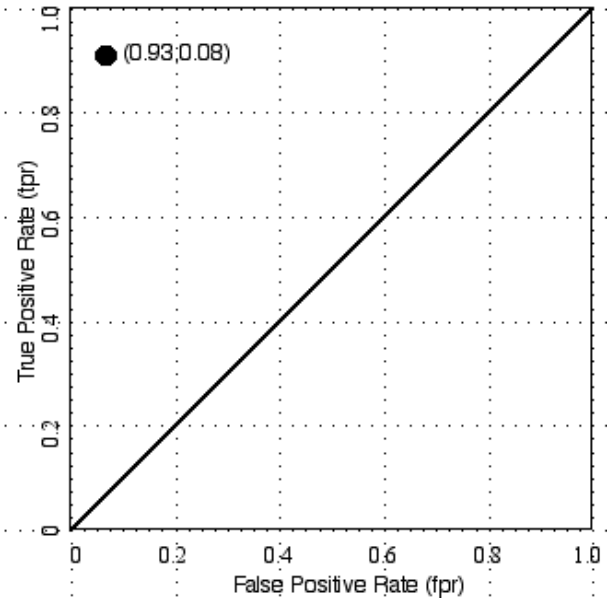


FIGURE 4.40: Mathematics I ROC Graph

The single point on this graph is located very close to the top left-hand corner, indicating that the model is very successful in predicting the first year, overall outcome. We can therefore say that in our context, Mathematics I can be used to predict the overall outcome of first time first year BSc. students enrolled in STEM disciplines.

4.9.7 Predicting Degree Attainment (Using Mathematics I)

The experiment we conducted in Subsection 4.8.3 was to determine the effectiveness of Mathematics I as a predictor of degree attainment. We stored the MultilayerPerceptron (MLP) model produced during the experiment in a file named *MathDegreePredict.model*. Here, we present our findings on the performance of the model in predicting degree attainment. We applied the model against a test dataset of 205 first year student records. Figure 4.41 shows a sample of the prediction results we obtained after applying our model against the test dataset.

```
@relation TrainQ7degAtt_predicted

@attribute Year numeric
@attribute FUNDING_STATUS {F,NF}
@attribute FINAL_YEAR numeric
@attribute MATH1036 numeric
@attribute MATH1034 numeric
@attribute Progress_outcome {PCD,RET,MBR,M1C,M2C,MBZ}
@attribute 'prediction margin' numeric
@attribute 'predicted OUTCOME' {Q,NQ}
@attribute OUTCOME {Q,NQ}

@data
2017,F,2019,83,85,PCD,0.990233,Q,?
2017,F,2019,47,67,PCD,0.973221,Q,?
2017,F,2019,38,50,RET,0.982783,Q,?
2017,F,2019,64,66,PCD,0.986534,Q,?
2017,F,2019,60,55,PCD,0.986293,Q,?
2017,F,2019,56,58,RET,0.99462,Q,?
2017,F,2019,55,61,PCD,0.982963,Q,?
2017,F,2019,50,50,PCD,0.980963,Q,?
2017,F,2019,44,62,PCD,0.969507,Q,?
2017,F,2019,72,75,PCD,0.987976,Q,?
2017,F,2019,50,53,PCD,0.980454,Q,?
2017,F,2019,52,63,PCD,0.980347,Q,?
```

FIGURE 4.41: Degree Attainment Prediction Test Set (Using Mathematics I)

Figure 4.42 shows the confusion matrix produced by the model.

		Predicted	
		Q	NQ
Actual	Q	152	0
	NQ	53	0

FIGURE 4.42: Degree Attainment Prediction Confusion Matrix (Using Mathematics I)

All 152 records that actually belong to class Q are correctly classified as Q. On the other hand all 53 records that actually belong to class NQ are classified incorrectly as belonging to class Q. Using the entries in this matrix, we computed different metric values. The model is 74.1% accurate in predicting degree achievement, using Mathematics I as a predictor attribute. This value is much lower compared to 91.3%, the value we obtained during the Modeling Phase. The overall error rate of 25.9% is much higher compared to the 8.7% value obtained during the Modelling Phase. The true positive and false positive rate values are the same (1.0000). Likewise, the true negative and false negative rate values are the same (0.0000). The model is predicting every instance in the test dataset to be of class Q. In other words, all the students represented by the records in the dataset have qualified and obtained their degrees. This is not correct. Our Mathematics-based model is, in this case, clearly not suitable as a predictor of degree achievement, despite the fact that it was very successful in predicting the overall, first year outcome.

We made important findings on the question we needed to answer regarding the effectiveness of Mathematics I as a predictor of the overall performance in first year, and degree achievement. Our findings showed, among other things, that

- Algebra I and Calculus I have a positive, strong and significant relationship with each other,
- Algebra I and Calculus I each has positive, weak and significant relationship with the NSC Mathematics (MAT) subject,
- the correlation between Algebra I and the NSC Mathematics Paper 3 (MP3) is negative, very weak and not significant. On the other hand, the correlation between Calculus I and MP3 is positive, very weak and not significant,
- the J48 decision tree model (*MathFirstYearPredict.model*), in which Mathematics I was used as a predictor attribute, was highly successful in predicting the first year, overall result,
- the MLP model (*MathDegreePredict.model*), in which Mathematics I was used as a predictor attribute, was not successful in predicting degree attainment.

4.10 Chapter Summary

In Chapter 3 we discussed in detail the six-phase CRISP-DM methodology we followed in conducting the research. During the Modeling Phase we identified the statistical and data mining techniques we were going to use in our experiments. A list of these techniques is provided in Table 3.9, Subsection 3.7.4. The tasks we performed are divided into three parts. First, beginning with the Modeling Phase, we applied different

data mining algorithms on the training datasets, resulting in several prediction models. Next, these models were used to classify test datasets records. Finally, in the CRISP-DM Evaluation Phase, we used different measures to evaluate the models' performance against the test datasets.

During the Modeling Phase we performed statistical analysis and data mining experiments, and obtained results which enabled us to answer research questions about first time first year students enrolled in STEM disciplines offered by our Faculty of Science. The factorial ANOVA experiments we conducted to determine the relationships amongst gender, race, funding status and the APS produced some interesting results. Whilst student gender and race together had no significant effect on the APS, the race factor on its own had a significant effect on the APS. We also found race and funding status together, and also individually, to have a significant effect on the APS. Regarding the APS obtained by English Home Language (ENA) and English First Additional Language (ENB) students, the statistical analysis test we conducted to compare these two groups of students showed that the mean APS obtained by ENA students was significantly higher compared to the mean APS obtained by ENB students.

For high school students intending to pursue STEM-related degree programmes, several issues need to be taken into account. One of these is to identify which amongst the Grade 12 STEM subjects contribute significantly towards the APS. Using feature selection and linear regression techniques we identified several Grade 12 STEM subjects as strong predictors of the APS, amongst which the following were significant predictors: Advanced Programming Mathematics (APM), Agricultural Sciences (AGS), Computer Applications Technology (CAT), Geography (GEG), Life Sciences (LSC), Mathematics (MAT), Mathematics Paper 3 (MP3) and Physical Sciences (PSC). Regarding prior programming experience (PPE), first year students who took Information Technology (INT) as one of their Grade 12 subjects, performed significantly better in the two first-semester Computer Science I courses (BCO and IAP) compared to their peers who did not take Information Technology as one of their Grade 12 subjects. We also used BCO and IAP to predict performance in Computer Science I and the overall result in first year. In both cases, the results showed IAP to be the strongest predictor, compared to BCO.

The experiments we conducted using the J48 algorithm and the Wits minimum admission requirements highlighted the potential of the Admission Point Score, NSC Mathematics, NSC Physics and NSC English as predictors of the overall, first year outcome. Cluster analysis and association rule mining techniques have shown that student characteristics can be used to obtain insightful information about first time first year students in our faculty. The results also highlighted the key role played by Mathematics I in the various options of the STEM degree curricula offered by the Faculty of Science. The strong correlation between NSC Grade 12 Mathematics and Mathematics I may be an

indication of a smooth link in the content offered between the two subjects. Mathematics I also emerged as a strong predictor of the overall first year result. In Chapter 5 we discuss the results we obtained after performing different statistical and data mining experiments. Our findings will be discussed within the context of results obtained in studies similar to ours. Furthermore, in Chapter 5 we will discuss some of the limitations we identified in connection with this study.

Chapter 5

Discussion

In Chapter 4 we performed detailed data mining and statistical analysis experiments, whose purpose was to determine answers to our research questions. In the Evaluation Phase of the CRISP-DM methodology we followed in conducting this data mining-based study, a detailed performance evaluation of the data mining algorithms used during the Modeling Phase was carried out. The evaluation was done using different test datasets and selected performance metrics.

In this chapter we present and discuss our findings, which we compare and contrast with findings obtained in studies similar to ours. This study sought to answer the following main research question:

> *How can data mining and statistical techniques be leveraged to obtain useful and meaningful information from the historical data belonging to cohorts of first-time first year students enrolled in STEM degree disciplines at a South African university?*

This question we answered by posing several research sub-questions, to which answers were sought by conducting several experiments.

This chapter is organized as follows. In Section 5.1 our findings on the question of student gender, ethnicity and funding status are discussed. In Section 5.2 we discuss the role of English on the performance of first time first year BSc. students enrolled in STEM disciplines in our faculty. The first year students who are the subjects of this study will have taken a variety of STEM subjects at high school, including during their final year (Grade 12). These subjects include Engineering Graphics & Design, Computer Applications Technology, Information Technology, Mathematics, Life Sciences and Physical Sciences. In Section 5.3 our findings on the influence of some of these subjects on achievement of the Admission Point Score (APS) and academic achievement

in first year are discussed. Some of the first year students take Computer Science I as one of their first year courses. The course involves extensive computer programming. In Section 5.4, our findings on the question of prior computer programming experience and its relationship to performance in Computer Science I and the academic outcome at the end of first year are discussed. Furthermore, in Section 5.5, the findings on the Wits minimum admission requirements, which constitute the criteria the faculty uses for accepting prospective first year students into different STEM disciplines are discussed. Students' characteristics can provide useful and meaningful insights about them. These include gender, race, economic background, academic performance, and funding status. In Section 5.6 we discuss our findings on this question. In Section 5.7 our findings on the important role and influence of Mathematics I across the first year STEM disciplines are discussed. In Section 5.8 the discussion focuses on some of the limitations associated with this study.

5.1 Gender, Race and Funding Status

Insightful information can be obtained from a study of different student characteristics. One of several things we might want to know is which characteristics impact or influence academic performance. In this study we investigated selected student characteristics to determine their effect on the first year students' high school performance, which is determined using the Admission Point Score (APS). Two-way ANOVA experiments were performed. In the first experiment, the aim was to determine the relationship of gender and race, on the one hand, and the APS, on the other. In the second experiment, we wanted to determine the relationship of race and funding status, with the APS.

We found the interaction between the gender and race factors not to be significant. Similarly, (Vandenberg et al., 2021) did not find the interaction effect between gender and race to be significant, regarding elementary students' attitudes toward Computer Science. In this study, gender did not have a significant effect on the APS. Similarly, the two-way ANOVA analysis performed by (Vitale et al., 2010) found that gender had no significant affect on the academic achievement of agricultural economics students. Likewise, (Keengwe, 2007) performed a two-way ANOVA analysis and found that gender did not significantly affect students' perceptions on computer use as a way of improving their learning. However, contrary to these findings, a much earlier two-way ANOVA study conducted by (Carroll and Garavalia, 2002) found gender to have a significant effect on the admission criteria of pharmacy students. Similarly, (Vandenberg et al., 2021) found the difference between male and female students to be significant, regarding their attitudes toward Computer Science. Male students had significantly higher attitudes towards Computer Science compared to their female counterparts.

In this study we found the race factor to have a significant impact on the APS. Upon further examination, several findings were made. Amongst Chinese, Indian and White students their performance was not significantly different. However, Indian and White students obtained significantly higher APS scores compared to Black and Coloured students, whose performance was not significantly different from that of Chinese students. According to (Essop, 2020) white students, for a number of reasons, produce the best results. In (Vandenberg et al., 2021)'s study, race was not a significant factor in elementary students' attitudes toward Computer Science.

This study's finding on students' ethnicity and academic performance was not unexpected. For example, there are several factors which impact negatively on Black and Coloured students' academic achievement. In our social and economic context, many of these students come from poor families, who live under the most depressed social and economic circumstances. In most of these families, the parents are not highly educated, and good role models within and outside of family units are very few. The township and rural schools they attend are severely under-resourced, both in terms of material resources and qualified teachers and administrators. As if all this was not enough, some or most of these schools do not have experienced teachers whose content knowledge in critical subjects is on par with highly qualified teachers from well-resourced, predominantly Indian and White schools.

This study also looked at the relationship between race and funding type, on the one hand, and the APS, on the other. Both factors were significantly related to the APS. There are a number of ways in which this result can be used in the schools and faculty in proactively providing academic and other forms of support to funded and non-funded students with borderline APS results. For example, the mean APS results of Bursary (B), NSFAS (NUG) and Non-funded (NF) students are lower compared to the results obtained by students who receive other forms of funding. For example, Coloured and Indian Bursary students have the lowest mean APS result compared to Black and White students. For Black, Chinese and Coloured NSFAS students the mean APS results are lower than Indian and White students. Non-funded students of all races have obtained similar, low mean APS results. It may be necessary in the future for schools to flag "weaker" students from within these groups as potential "at risk" cases, whose academic progress should be consistently monitored, and various actions taken to remedy problematic academic situations.

From the significant relationship between funding status and the APS, there are a few points worth noting. The best performing students are those who receive Merit Awards (SUG), followed by Corporate Bursary (CB), External Bursary (EB) and Internal Bursary Undergraduate (IBUG) students. First year merit award students, based on their excellent high school results, receive merit bursaries. In our context, and indeed elsewhere, these students have been shown to obtain the best first year results compared to

their peers (Naidoo and McKay, 2018). However, there was a negative and significant correlation to the GPA of NSFAS-funded students, leading (Naidoo and McKay, 2018) to question the effectiveness of NSFAS bursaries. In our study, NSFAS-funded (NUG) students were only better compared to non-funded (NF) students, who obtained the lowest average APS score.

Contrary to (Naidoo and McKay, 2018)'s finding regarding the performance of NSFAS-funded students, (Wildschut et al., 2020) discovered a positive, significant relationship between NSFAS funding and academic achievement. In a study conducted earlier in our school, (Modise, 2016) found that NSFAS funding did not significantly affect the performance of first year Computer Science students. It is clear that this matter is not yet settled, given the contradictory findings emerging from these different studies. (Wildschut et al., 2020) has also found that on average, NSFAS-funded students are outperformed by students who receive funding from other sources. This finding is consistent with (Sommerville and Singaram, 2018)'s, who found a significant performance difference between externally funded (EB) and NSFAS-funded (NUG) students. We are in agreement with (Wildschut et al., 2020) and (Sommerville and Singaram, 2018)'s findings. In this study, significant differences were found in the average APS results obtained by different students, based on the type of funding they received. As in (Sommerville and Singaram, 2018), External Bursary (EB) students outperformed NSFAS-funded (NUG) students.

5.2 English Language Proficiency

Most South African higher education institutions use English as the medium of instruction. Not all students who attend tertiary institutions come from English home language backgrounds. The majority of students, including those who study in historically White universities, come from families in which the home language is not English. Most of these students come from families in which indigenous African languages are spoken. In a higher education institution such as ours, which uses English to deliver content, what is the impact of this policy on English second language students? The academic prospects of students who have difficulty comprehending content delivered in English are limited (Stephen et al., 2004).

In a similar vein, (McGhie, 2012) refers to the "disadvantage" faced by Black students who, as English second language speakers, have to use English, a language in which they are not competent, in order "to learn and construct meaning". (Wedderburn, 2005), whose study was conducted amongst first year Computer Science students, states that the needs of English second language students must be considered when teaching and setting assignments, with greater emphasis being placed for example, on explaining the meaning of words and how they are used in a specific context. According to

(Wedderburn, 2005), the problem not only affects English second language Computer Science students. Students in other fields of study are likewise affected. We cannot but agree with all these studies that, for English second language speakers, using English as a language of teaching poses a formidable challenge.

At high school, the first year students in our faculty will have done either the English Home Language (ENA) or the English First Additional Language (ENB) subjects. One of the objectives of this study was to discover whether the first time first year students in our faculty who did either the ENA or ENB subject in high school obtained significantly different APS results. We found a significant difference between the mean APS scores of ENA and ENB students, with ENA students outperforming ENB students. We speculate that ENA students carry this advantage with them when they enter university, which places them in a better position to succeed in first year compared to their ENB counterparts. Our preliminary findings indicate that across different schools in our faculty, ENA students outperform their ENB counterparts.

In a study conducted earlier in our school, it was found that high school English at the higher grade (first language) correlated better with achievement in first year Computer Science, compared with high school Mathematics (Rauchas et al., 2006). A result similar to (Rauchas et al., 2006) was obtained in an earlier study by (Pillay and Jugoo, 2005), in which English first language students significantly outperformed English second language students in a procedural programming examination. In medicine, (Sommerville and Singaram, 2018) found that proficiency in English amongst medical undergraduates had a significant influence on their academic performance.

In a first year optometry degree programme, and in contrast with findings from other studies, (Mashige et al., 2014) found the correlation between high school English and first year optometry module scores to be weak. However, it is not clear from the study whether the English ENA and ENB subjects were used separately or combined to determine the correlations. In a much earlier study, (Wong and Chia, 1996) stated that a high proficiency level in English would have a positive effect on proficiency in mathematics, resulting in successful academic performance amongst first year financial accounting students. Similarly, the study by (Dlomo et al., 2011) found that students who obtained a high ENA score were successful in their studies, whilst students from English second language backgrounds performed the worst.

In some studies, the lack of English first language status has been found not to be detrimental to English second language students. In their study (Lee and Mallik, 2015) refer to evidence which suggests that using English first language for teaching does not, in any significant way, disadvantage English second language students in their performance.

5.3 Pre-University STEM Subjects

One of the aims in this study was to identify which STEM subjects, taken by Grade 12 NSC students, had the most influence as predictors of the aggregate Admission Point Score (APS). The feature selection and multiple linear regression methods we used produced similar results, identifying the following Grade 12 NSC subjects as the strongest and significant predictors of the APS: Advanced Programming Mathematics (APM), Computer Applications Technology (CAT), Geography (GEG), Life Sciences (LSC), Mathematics (MAT), Mathematics Paper 3 (MP3) and Physical Sciences (PSC). For different stakeholders, these results have practical implications. For example, to improve the chances of obtaining a good APS, students on the science track can be advised and encouraged to select STEM subjects that are known to have a strong and significant influence on the APS. Our universities use the APS as one of the criteria for admitting students into their preferred fields of study. For STEM students, a high APS and good performance in high school STEM subjects guarantees acceptance in a number of STEM-based degree disciplines.

In contrast with our study, in which we used high school STEM subjects to predict the APS, (Blackmore et al., 2021) investigated the relationship between secondary school performance and the undergraduate grade point average (GPA) scores obtained by undergraduate health sciences students. (Blackmore et al., 2021) discovered a link between three secondary school STEM subjects (Advanced Mathematics, Physics and Chemistry) and a high first year GPA. Only Advanced Mathematics emerged as the best predictor of overall academic performance. In (Wang, 2013)'s study, it was shown that Advanced Mathematics and Science high school subjects placed students in a better position, in terms of attaining a STEM-based degree. In this study, high school Mathematics (MAT) was the strongest predictor of the APS, followed by Advanced Programming Mathematics (APM). Due to the fact that only a few high school students took Advanced Mathematics (ADV), this subject could not be used to determine its contribution towards the APS.

In the South African context, universities use high school marks together with the APS to decide whether or not to admit students who apply for specific academic programmes. In addition to these admission criteria, some universities also include the National Benchmark Test (NBT) scores (Hunt et al., 2009). The NBT comprises three components, namely academic literacy, quantitative literacy and mathematics. Studies similar to ours have been conducted to determine the effect on college performance of high school subjects, high school Grade Point Average (HSGPA) and standardized placement tests such as the Scholastic Aptitude Test (SAT) and American College Testing (ACT). Results from these studies can provide, in our context, some understanding and insights

into how high school STEM subjects, the APS and the NBT could be used, not only for the purpose of student placement, but also to predict first year academic performance.

Similar to South Africa's NBT, the ACCUPLACER and COMPASS placement tests used in (Belfield and Crosta, 2012)'s study each contain quantitative and literacy components. There was a weak, positive correlation between the results from these tests and college GPA. On the other hand, the high school GPA (HSGPA) was strongly correlated to college GPA, leading to (Belfield and Crosta, 2012) suggesting that colleges could consider using the HSGPA as a criterion for placement decision-making. Since it appears to encapsulate the predictive power of all the high school subjects, the authors suggest that the HSGPA can, on its own, be used to predict college performance. (Belfield and Crosta, 2012)'s finding is in agreement with (Geiser and Santelices, 2007), who also found that high school grades were more powerful as a college admissions criterion compared to standardized tests such as the SAT. According to (Geiser and Santelices, 2007), high school grades, despite being regarded by some researchers as being unreliable, "remain the best available indicator with which to hazard predictions of student success in college". These and similar studies seem to confirm the correctness of the admissions approach taken by our faculty, where appropriate high school STEM subjects and the APS, which plays a role similar to the HSGPA, are used as a criterion for admitting students to first year STEM degree studies.

5.4 Prior Computer Programming Experience

In this part of the study we wanted to do the following: investigate the relationship between prior computer programming experience and performance in two, first semester Computer Science I courses (BCO and IAP); determine the effectiveness of BCO and IAP in predicting overall achievement in Computer Science I and determine the effectiveness of BCO and IAP as predictors of the overall, first year outcome.

Regarding prior computer programming experience and performance in BCO and IAP, we found significant differences in both courses between students with prior programming and their peers who lacked such experience. This finding is in agreement with the result of an earlier study conducted in the school by (Modise, 2016), who found significant differences in BCO and IAP between the two groups of students. Other studies have reported findings similar to these, regarding the importance of prior computer programming experience. (Kori et al., 2016) found that the first semester weighted average grades of students with programming experience were higher compared to grades obtained by students who lacked programming experience. We concur with (Kori et al., 2016)'s view that universities should consider prior programming experience when crafting their ICT-related curricula, including providing adequate support to students, to

prevent them from dropping out of their studies. Another study, by (Wilcox and Lionelle, 2018), found that prior computer programming experience had significant advantages, with gender playing an influential role. Female students with prior programming experience outperformed their male counterparts in all respects. Similar to (Wilcox and Lionelle, 2018)'s findings, (Byrne and Lyons, 2001)'s earlier study showed that female students with prior programming experience outperformed all other students in the first year, Programming and Logical Methods course. However, given the fact that in a class of 110 students only three female students had prior programming experience, a significant conclusion could not be drawn from this result. We are not in full agreement with (Wilcox and Lionelle, 2018), that the large performance gap in introductory courses may be difficult to close between students with computer programming skills and those who lack such skills. We believe that given adequate support, experienced mentoring and extra programming lessons, inexperienced students may reach the same level as their experienced peers. Contrary to expectation, in (Ayalew et al., 2018)'s study the high school computer studies course, which includes programming, did not have a significant relationship with the first year programming subject. On the contrary, (Ayalew et al., 2018) found the high school mathematics and physics subjects to be significantly correlated with the first year programming subject.

There is evidence from our training phase results that using multiple linear regression, the two first semester courses (BCO and IAP) can together successfully predict achievement in Computer Science I (CSI). Due to COVID-19 restrictions, which affected student attendance, we were unable to obtain a sufficient number of test dataset records to meaningfully evaluate our linear regression model. In the regression model the highest contribution towards the CSI score came from IAP, whose regression coefficient was greater compared to that of BCO. In this respect the results are in agreement with findings from previous studies conducted in our school by (Matsime, 2015) and (Sandlesh, 2015), which also showed that BCO and IAP together were good predictors of performance in CSI. In contrast to (Matsime, 2015) and (Sandlesh, 2015), who used multiple linear regression to predict CSI, (Mashiloane, 2016) used the J48 decision tree algorithm and the mean score of two first semester courses to predict CSI performance, in which the algorithm achieved a high accuracy value of 92.54%. These findings have several implications for our students and other institutional stakeholders. Students who do not do well in BCO and IAP face the risk of failing CSI at the end of the second semester. Those who do not perform better in the second semester CSI subjects, may decide to drop out of the entire course. For academics and administrators, students identified to be at risk of failure require various forms of support such as counseling, confidence building, extra tutorials and classes. Based on their performance in BCO and IAP, (Sandlesh, 2015)'s Academic Recommender System (ARS) was able to classify student records into high, medium and low risk categories, in terms of predicting their performance in CSI. The tool provides medium or high risk students with appropriate

recommendations, to enable them to succeed in the second semester. Low risk students, who will have passed BCO and IAP, are encouraged to continue with their good work in the second semester. In another study within our School, (Telford, 2016) used a web-based recommender system for a similar purpose.

The results obtained during the Modeling Phase and Evaluation Phase have shown that the J48 decision tree algorithm, in conjunction with BCO and IAP as predictor attributes can be used, to predict the overall achievement of first year BSc. students in our faculty. In contrast, (Matsime, 2015) used the Naïve Bayes (NB) algorithm to predict the overall first year outcome. Compared to the 82.92% success rate obtained by the J48 model we used in our study, (Matsime, 2015)'s Naïve Bayes algorithm was only 58.59% accurate. We are in full agreement with (Matsime, 2015)'s observation regarding the imbalanced nature of educational datasets, and the tendency of classifiers being biased more towards the majority class. This is also what we observed when we evaluated the performance of the J48 algorithm, where majority class instances were mostly correctly classified, with some of the minority class instances being misclassified. One of the serious implications resulting from a misclassification error may be to predict that a student who ends up failing is going to pass. As a consequence, the student is not identified as being "at risk", and does not receive the required counseling and academic support they need, had their status been correctly identified.

5.5 Wits Minimum Admission Requirements

In our faculty, first year students are admitted on the basis of their achievement of their high school Mathematics, Physical Sciences, English and the APS. These are the minimum admission requirement subjects specified by the faculty. We conducted two experiments. Our first task was to determine whether these requirements were suitable predictors of the first year outcome. Secondly, we wanted to determine if they could also be used to predict degree attainment. During the Modeling Phase and Evaluation Phase two datasets were used, one consisting of records belonging to first year students who took the English Home Language (ENA) subject at high school, and the other consisting of records belonging to students who took the English First Additional Language (ENB) subject.

During the Evaluation Phase, in which ENA and ENB test datasets were used, the J48 algorithm was successful in predicting the first year overall outcome. However, for the algorithm to work at its best, some improvement is required, to enable it to correctly classify records across all classes. Second, the imbalanced class distribution of the records posed another challenge to the classifier. The majority class (PCD in this case) was highly dominant, whilst the other classes were vastly underrepresented, containing a small number of records. The high false positive errors also raise a serious concern,

in that students who have actually failed may incorrectly be considered to have passed, a situation that again raises serious implications for students, administrators and the faculty. Admission requirements similar to those used in our study were shown to be suitable predictors of engineering students' performance (Louw et al., 2013). Given our findings regarding the performance of the J48 algorithm, we cannot but agree with (Lemnaru, 2012) that the J48 algorithm does not perform well against imbalanced datasets. Amongst all the classifiers used by (Lemnaru, 2012) the Multilayer Perceptron (MLP) was found to be the most robust in handling the class imbalance problem. However, this did not prove to be the case in one of our experiments, where the MLP classifier was outperformed by three other classifiers (J48, JRIP and NB).

Regarding the prediction of degree attainment, our results were not encouraging. Whilst the evaluation of the MLP classifier yielded a good accuracy result (80.1%), other evaluation metrics were not that good. The false positive rate was very high, indicating the total failure of the classifier to correctly identify students who have not qualified (NQ), by incorrectly categorizing them as having qualified (Q). What would be the consequence, legal or otherwise, if a student who has not qualified were to be notified that they have qualified? We are in agreement with (He and Garcia, 2009) that a classifier is required "that will provide high accuracy for the minority class without severely jeopardizing the accuracy of the majority class". Perhaps we need to sound a cautionary note at this point. Our findings indicate very clearly that classification accuracy, in the absence of other metrics, cannot be used as the sole criterion for evaluating the performance of a classifier. For a comprehensive assessment of a classifier's performance against imbalanced data other metrics are required, in addition to accuracy. Due to the imbalanced nature of our training and test datasets, the MLP classifier yielded a highly skewed degree of accuracy in favour of the majority class (Q). In contrast with (Lemnaru, 2012)'s results, the MLP classifier failed dismally in our case to demonstrate its robustness against our datasets, both during the Modeling Phase and Evaluation Phase. In the absence of a plausible explanation of this result, our speculation is that the type of data and the domain from which the data is obtained, may affect the way in which it is analyzed by an algorithm.

5.6 Student Characteristics

A number of studies have used student characteristics to investigate, amongst others, their influence on students' academic performance. In a study conducted at a South African tertiary institution, (Pillay and Jugoo, 2005) wanted to identify characteristics which negatively affected student performance in a programming course. The characteristics included gender, learning style and previous computer experience. In (Cochran et al., 2014), the authors wanted to identify characteristics that were associated with students withdrawing from online classes. The characteristics included student age, gender and race. Our study used the following characteristics to determine their relationship with the performance of first time first year students enrolled in STEM disciplines in our faculty: funding status, gender, race, home language, APS, NSC Mathematics, NSC Physical Sciences, NSC English, first year outcome and academic plan. In order to better understand the role and influence of these characteristics on academic achievement, we used clustering and association rule data mining techniques.

Our clustering results indicate that home language and the high school English subject (ENA or ENB) contribute towards the success or failure of our first year students. In (Jacobs, 2010)'s study students from an African home language background were identified as being at a higher risk for failure, when compared to their counterparts from Afrikaans and English home language backgrounds. In (Pillay and Jugoo, 2005)'s study, for example, the programming and examination marks obtained by students from English home language backgrounds were significantly higher compared to those obtained by students from English second language backgrounds. In (Rauchas et al., 2006)'s study, first year Computer Science students who studied Higher Grade English at high school outperformed their counterparts who took Standard Grade English. A different view is held by (Mashige et al., 2014) who, based on their findings, concluded that English proficiency was not a critical factor of success in first year optometry modules.

Our clustering results have brought to the fore some interesting issues regarding the role of English in our high school curriculum. From our results, it would seem that students whose English First Additional Language (ENB) score is at least 65%, are most likely to succeed in first year, by obtaining a PCD overall outcome. On the other hand, it would seem that students who obtain ENB scores below 65% do not pass their first year, obtaining either RET or MBR overall outcomes. However, we need to stress the fact that these are only preliminary findings. There is a need for a detailed investigation in the future, to determine whether there is indeed a pattern that indicates failure when students obtain ENB scores that fall below 65%. The implication, should this be the case, would be for the faculty to look into the possibility of introducing an ENB threshold of 65% or more as part of the minimum admission requirements. This would help to prevent this potentially "at risk" group of students from failing their first year.

Given (Mashige et al., 2014)'s finding regarding the weak influence of ENB on student performance, our stance is all the more strengthened that this subject be looked at in some depth to determine its contribution or impact towards first year students' academic performance.

Again, to further stress the important role of English, both as home and teaching language, we identified from our clusters two groups of funded students enrolled in the biological sciences (BIOL). In the first group the home language was not English and in the second English was used as the home language. In both groups the students did the English Home Language (ENA) subject at high school. The high school performance between these two groups was very different, in terms of the APS, Mathematics, Physical Sciences and English. Perhaps unsurprisingly, the students who spoke English at home were overall the best performers. However, both groups were successful at the end of the first year, due in part to the advantage of having done ENA at high school. In contrast with these two groups, another group of non-funded BIOL students were unsuccessful in first year. A closer look at their profile showed that their home language was not English. Furthermore, they took English First Additional Language (ENB) at high school. Their high school APS, Mathematics, Physical Sciences and English scores were also lower, compared to the first group of funded, successful BIOL students.

Our results show that funded students in STEM disciplines are, on the whole, successful in their studies. This may be the consequence of them having been selected based on their high school results, academic potential and capabilities, and not necessarily based on the fact that they come from poor households. (Naidoo and McKay, 2018) caution against using "neediness" as a criterion for funding. In their study, no relationship was found between funding and academic achievement. (Naidoo and McKay, 2018) also argue for the need for further research to be conducted to correctly identify first year students who deserve to be funded, basing their argument on the fact that high school results may not necessarily be suitable indicators of tertiary-level academic success.

On the question of gender, (Wildschut et al., 2020)'s study found the performance of funded and self-funded male students to be significantly and negatively related with average performance, particularly for NSFAS-funded students. In our study, there were four male-dominated clusters, all representing students who received different types of funding, including NSFAS-funding. In three of these clusters, the students were successful in first year, obtaining a PCD overall result.

In addition to cluster analysis, we also used the association rule mining technique to investigate the relationship between our first year students' characteristics and their academic achievement. Several insights were obtained from the association rules that were generated from our experiments. To assess the relevance of these rules, we used the *confidence* and *lift* metrics. As opposed to the clustering results, which showed a limited

number of STEM fields, more STEM fields were covered by the association rules. As in (Hämäläinen, 2006), the majority of the rules were related to successful students.

The rules provide some insights into the different fields of study in which the first year STEM students are enrolled. Taking confidence and lift values into account, Mathematical Statistics (STAT) students are ranked highest, as indicated by the following rule:

```
PlanCode=STAT==>Outcome=PCD <conf=1.0, lift=1.57>
```

The confidence of this rule is very high (100%), indicating that of the students who enroll in STAT, all of them are successful in first year. The lift for this rule is greater than 1, indicating that a student who enrolls in STAT is 1.57 times likely to pass in first year. In other words, the antecedent and consequent of this rule are positively correlated. They occur more often together. This result is hardly surprising. Students who enroll in STAT are accepted only on condition that they achieve very high scores in Grade 12 ($\geq$80% Mathematics, $\geq$80% Physical Sciences, $\geq$80% English and $\geq$40 points). Being high achievers, the vast majority of these students end up passing their first year. Similar, interesting rules were found in the following disciplines: Microbiology and Development Biology (MICR), Biological Sciences (BIOL), Human Physiology (HPHS), Physics (PHYS), Genetics and Development Biology (GENE).

The cluster analysis results showed that across different disciplines, funded students were, on the whole, successful in their studies. This finding is confirmed by the association rules that were generated by the *apriori* algorithm. Funded students were successful in, among other fields, the Mathematical Sciences (STAT), Biological Sciences (BIOL), Actuarial Sciences (ACSI) and Human Physiology (HPHS). The confidence and lift values of the association rules for funded students were both slightly higher compared to non-funded students. For example, the following rules were generated for Human Physiology (HPHS):

```
FStatus=Funded,PlanCode=HPHS==>Outcome=PCD <conf=0.97, lift=1.53>
FStatus=non-funded,PlanCode=HPHS==>Outcome=PCD <conf=0.94,lift=1.48>
```

The cluster analysis and association rule mining results indicate that funded first time first year STEM students are, like their non-funded counterparts, academically successful in their studies. The cluster analysis and association mining results obtained in this study can be used within our faculty to recommend to individual students, based on their academic potential and characteristics, the most suitable BSc. degree options from which to select their field of study.

5.7 Role and Influence of Mathematics I

The majority of first year students in our faculty are required to take Mathematics I as a compulsory subject. This subject features prominently in the different options of the BSc. degree offered by our faculty. Given its role and influence in our STEM curricula, we conducted an in-depth investigation to determine its impact on academic achievement in first year, including its suitability as a predictor of degree attainment. We were also interested to discover how strongly the NSC Grade 12 Mathematics subjects (MAT and MP3) were related to the two Mathematics I subjects, Math1034 (Algebra) and Math1036 (Calculus).

Our study produced several findings regarding high school mathematics subjects and their relationship to Mathematics I. The correlation between MAT and each of the two first year Mathematics subjects (Algebra I and Calculus I) was positive, weak and significant. Other studies where high school mathematics subjects have also been found to correlate significantly with first year mathematics subjects include (Mashige et al., 2014), (Louw et al., 2013) and (Kizito et al., 2016). Contrary to expectation, our study found the high school MAT and MP3 subjects to be weakly and insignificantly related to each other. At some point in the faculty, for students whose MAT score fell just below the 70% admission threshold, the MP3 score would be used as part of the admission requirements to decide whether to accept or reject them. The poor relationship between MAT and MP3 brings into question the decision that was taken to use MP3 as a complementary subject in admitting students with below-threshold MAT scores. We also found the MP3 and Mathematics I subjects to be weakly and insignificantly correlated, with MP3 also having a negative correlation with Algebra I.

Whilst the correlation between MAT and the APS is weak and significant, MP3 on the other hand correlates even more weakly and insignificantly with the APS. These findings bring into question the relevance and usefulness of MP3 within the high school Mathematics curriculum, and its potential as a bridging subject to first year Mathematics. Compared to MAT the APS correlated moderately and significantly with Algebra I and Calculus I. One would have expected that the MAT subject would correlate better than the APS, with Algebra I and Calculus I. We can only speculate on this result for now, by saying that the APS, being an aggregate score, carries the contribution of scores from different subjects, resulting in it being able to correlate better with Algebra I and Calculus I, compared to the single MAT score.

The J48 algorithm which used Mathematics I during the Evaluation Phase to predict the overall outcome in first year, was highly accurate (92.4%). In a number of studies the argument has been made against relying on accuracy as the only performance evaluation metric, when using datasets containing class imbalances (Chawla, 2009), (He and Garcia, 2009), (López et al., 2013).

Since we did not want to rely solely on accuracy to evaluate the J48 model we were using during the Evaluation Phase to predict the overall first year outcome, we constructed a ROC graph in order to obtain a more reliable performance evaluation result. The result we obtained from the ROC graph showed our J48 model to be highly effective and successful in predicting the overall first year result with Mathematics I (Algebra I and Calculus I) being used as a predictor attribute. Not only in our faculty does first year Mathematics receive greater emphasis and attention. At an Irish university, where for Science students Mathematics is also a compulsory subject, the Mathematics Support Centre (MSC) established by that university has had a positive impact on the Mathematics performance of students who attend the centre, in particular students who are weak in Mathematics (Bhaird et al., 2009). In our context an entity similar to the MSC could be established faculty-wide where students, in particular those at risk for failure, can drop in for one-on-one remedial sessions with their tutors or lecturers.

The Multilayer Perceptron (MLP) algorithm we used to predict degree attainment was hardly successful in performing this task. None of the 53 minority class (NQ) test set records were classified correctly as class NQ records. All of the 152 test set records of the majority class (Q) were classified correctly as belonging to class Q. This result demonstrated the inability of our MLP model to handle imbalanced data, despite being considered to be a robust solution to the imbalanced data problem ((Lemnaru, 2012); (He and Garcia, 2009)). Given our result, there is a need in a future study to address this challenging, class imbalance problem. Since our study was limited in scope, and given the fact that this problem requires detailed attention, we decided not to explore the various solutions available to address it. This would have been time-consuming and very much beyond the scope of this study. Possible solutions to the problem include different sampling methods, cost-sensitive techniques, kernel-based techniques and ensemble-based techniques (Chawla, 2009), (He and Garcia, 2009),(López et al., 2013).

5.8 Study Limitations

In this section we highlight some of the limitations of this study. Firstly, the COVID-19 pandemic had a serious effect on student attendance, particularly in 2020, where physical presence on campus was severely restricted, with students having to attend their classes online. We realized the negative impact of this situation when we had to obtain the 2020 dataset of first year student records, which were going to be used as a test dataset

during the model performance Evaluation Phase. We obtained very few test set records that we could use to evaluate the regression model we created during the Modelling Phase to predict the final, Computer Science I score. Secondly, during the course of our experiments, we realized that due to the imbalanced nature of our datasets and the multiple class distribution, some of the data mining algorithms we used were limited in their ability to assign minority class records to the actual class to which they belonged. This problem requires detailed investigation and different experiments and techniques to address it. Given the limited scope of this study, this work can be done in a separate, future study. Lastly, we conducted this study at a single South African university. Our results may therefore not be applicable to other, higher education institutions, locally and globally. Having said this, we do believe that this study and its findings contain useful insights which can be of benefit to and inform the work of educational researchers, including those engaged in research in the field of Educational Data Mining (EDM).

Chapter 6

Conclusion and Future Work

Universities collect and amass vast amounts of data about different stakeholders, including students. Most of the data, after being collected, is stored in several repositories, with little prospect of ever being used again, for any purpose. As a result these repositories resemble "data tombs" (Kaufinann, 2006), archives that are only occasionally accessed, as and when the need arises to obtain some information from them. The universities and other entities which amass huge volumes of data and only use a tiny amount thereof, are best described as "data rich but information poor" environments (Kaufinann, 2006). In South Africa, and indeed elsewhere, there is a dire shortage of professionals and graduates who possess the much-needed STEM qualifications. In our own context, the data our university and faculty collect about first year STEM students can be leveraged to provide insightful and meaningful information about them.

In order to conduct this study in a more focused and systematic manner, we posed several research questions, in order to obtain information about the first time first year students enrolled in STEM degree programmes in our faculty. From this study, several interesting results were obtained.

Gender and race are amongst the factors this study sought to investigate, to determine their relationship with the APS. Between gender and race, their interaction was not significant. The interesting information from our data is that differences in the APS of students from different ethnic groups does not depend on their gender. Gender did not have a significant effect on the APS. Regarding the race factor, the statistically significant result shows Indian and White students outperforming Black and Coloured students. The difference between Indian and White students was not significant. This finding reflects the current reality of the South African high school educational environment, in which Indian and White students are taught mostly in well-resourced and well-administered high schools, by effective and highly qualified teaching staff. There is no doubt that in such an environment students are best placed to excel in their studies.

To remedy the dire situation that many of the township and rural schools find themselves in, several actions need to be taken. These schools must be well-resourced, and be administered efficiently and effectively. Highly qualified and "content-knowledgeable" or "content-capable" teachers must be employed, discipline amongst teachers and students must be fostered in the interests of creating an enabling and supportive educational environment. Furthermore, parents cannot be spectators in the education of their children. Through school governing bodies (SGBs), they need to be involved and play a day-to-day, active role in their children's education.

Different types of funding are provided to academically capable first year students, the majority of whom come from disadvantaged socioeconomic backgrounds. This study has revealed some interesting information about the relationship between the type of funding received by these students and their prior performance, as indicated by the APS. For example, our results confirm the academic excellence of scholarship or merit award (SUG) students of all races. Their mean APS results were the highest, compared to the APS results obtained by students who received other types of funding. However, even within this group, Indian and White students and to some extent Chinese students, obtain higher APS results compared to their Black and Coloured counterparts. This finding is statistically significant. Our speculation is that the best performing members of this group may be students who attended independent, private high schools and colleges, with much better resources and highly qualified, motivated teachers.

Our results show that NSFAS-funded (NUG) students, with the exception of non-funded (NF) and bursary (B) students, enter university with low APS results compared to students who receive other types of funding. The NSFAS spends billions of rand in funding needy students who qualify for tertiary studies. It is only reasonable, therefore, given the huge amount of government's and taxpayers' money spent on these students, to expect the majority of them to complete their studies in regulation time. Completing their studies on time will afford other, equally deserving and capable students the opportunity to obtain funding for their own studies. Our findings also show certain categories of NSFAS-funded students, including all categories of non-funded students, to be entering first year with rather low APS scores. To ensure these students' academic progression, their progress needs to be constantly monitored, through counselling and academic support, to prevent them from failing their studies. These are students who may be at risk for failure. There may therefore be a need for faculty-wide support to be provided to such students, to prevent them from failing.

Regarding the use of English as a language for teaching, both at high school and university, our preliminary findings show that students who took the English Home Language (ENA) subject at high school are more successful in first year compared to students who took the English First Additional Language (ENB) subject. An interesting finding we made is that students who spoke an indigenous language at home, but who took the

ENA subject in high school, were also most likely to succeed in first year compared to their counterparts from an indigenous background, and who took the ENB subject in high school.

In order to improve the chances of academic success for English second language students, there may be a need for intervention at two levels, namely at high schools and tertiary institutions. Our findings on the issue of English as a teaching language may be pointing to a gap that exists between ENA and ENB, and probably to the English language disadvantage ENB students carry with them when entering university. At high school level, this situation may perhaps be remedied by introducing a single-stream English subject. This should be coupled with the necessary remedial support to English second language students, to elevate their level of English proficiency to a standard suitable for the comprehension and understanding of the first year subject content.

We found the minimum admission requirement subjects to be successful predictors of the overall, first year result. The usefulness of this result is that we now have a model that can be used to predict, right at the beginning of their enrollment in our faculty, whether or not an individual student will be successful at the end of their first year of study. Students who are identified and predicted to be at risk for failure, can be provided with the necessary support and counselling during the course of the year, to prevent them from failing. We did not find the minimum admission requirement subjects to be suitable or successful predictors of degree achievement, a situation that calls for further research to be conducted to study this issue in greater detail.

Cluster analysis and association rule mining techniques provided useful information about student characteristics and their relationship to academic achievement. It would seem from our preliminary results that students who took ENB as their English subject in high school and speak an indigenous language at home, will most likely repeat their first year if their ENB score is lower than 65%. We need to stress that this is only a preliminary finding. Further research is required to determine whether this finding does indeed point towards a particular trend. There may be a need, across the faculty, for these students to be provided with the necessary language support, to improve their English language ability. Another interesting finding relates to English second language students who use English as their home language and who take ENA as their English subject in high school, and are enrolled in the same first year degree programme as their ENB counterparts. These students are more likely to proceed to second year, even when their ENA scores fall below the ENB scores of their counterparts. Again, further research is required to determine whether this result points to an enduring trend.

From our data the information we get is that on the whole, funded first year STEM students perform successfully in their studies. This is very encouraging. Given what seems to be a good indication of successful academic achievement, funding should be provided to capable and deserving, first time first year STEM students who require

financial support to begin their higher education studies. Our cluster analysis results seem to indicate that generally, funded and non-funded female students are successful in their first year studies. STEM disciplines are characterized by a huge gender imbalance, with females being underrepresented. Given the encouraging performance of first year female students in our context, our faculty should do more to recruit and encourage female students with appropriate high school subjects to enrol in STEM degree disciplines. In order to ensure that they succeed in their first year studies, courses can be recommended for them to select, which could be identified using data mining techniques such as association rule mining.

Given its pivotal role in our curriculum, Mathematics I is a critical gatekeeper subject. For different stakeholders in our faculty, several implications arise from the fact that the subject has been shown to be a very strong predictor of overall performance in first year. Students who do not perform satisfactorily in Mathematics in the first semester are potentially at risk for failure. Once these "at-risk" students have been identified the relevant schools, in particular the School of Mathematics, need to take proactive measures to prevent them from dropping out or failing. Such measures may include compulsory tutorial sessions, consultations with lecturers, mathematics study techniques, and counselling. There is a need for further research to be conducted to determine other, more effective ways in which Mathematics I can be better leveraged to predict degree achievement.

There is a positive, weak and significant correlation between the NSC Grade 12 Mathematics (MAT) subject and Mathematics I (Algebra I and Calculus I). This result indicates the relevance of the high school MAT subject to our Mathematics I curriculum. The same cannot be said about the NSC Grade 12 Mathematics Paper 3 (MP3) subject. It has a positive, weak and insignificant correlation with Calculus I; a negative, weak and insignificant correlation with Algebra I; and a positive, weak and insignificant correlation with NSC Grade 12 Mathematics (MAT). In its current form the MP3 subject does not seem to be relevant to the Mathematics I curriculum offered in our faculty. There is is no good reason why it should or could be considered as part of the criteria the faculty should use to accept first time first year students into different options of the BSc., STEM disciplines.

The relatively new discipline of Educational Data Mining has begun to address the challenge facing many higher education institutions, which is how to extract meaningful and useful information from the vast amount of data they collect and store in their repositories, about their students. In this study we used data mining and statistical techniques to leverage and obtain information from the data collected by our university, faculty and schools about the first time first year students enrolled in STEM disciplines in the Faculty of Science. The primary aim of this data mining based study was to obtain a better understanding of first time first year students enrolled in different STEM

disciplines in our faculty. To achieve this aim, we used different statistical and data mining techniques, with the intention of obtaining answers to the different research questions that we posed to direct this study. It is hoped that the results and insights emanating from this research will be of use to different stakeholders in our institution, for them to understand how best to engage with the first year STEM students at their service. In the South African educational context, research in Educational Data Mining is still in its infancy. Apart from contributing to the body of research that is beginning to emerge in this field, we believe that the work conducted in this study and the results thereof can provide valuable insights to educational researchers elsewhere.

6.1 Future Work

This study raised a few issues of interest which, due to its limited scope, the study could not address. We have identified some interesting topics for future investigation and research, whose detailed study may contribute useful information and knowledge in various areas of the increasingly important field of Educational Data Mining.

Some of the topics that could be studied as part of future work include the following:

- Some of the datasets in this study were highly imbalanced, a situation that affected the performance of some of the algorithms we used during the Modeling and Evaluation Phases. This problem requires a detailed investigation, in which different data mining techniques can be used and compared to determine the most effective technique that could be implemented to deal better with the problem. In a future study, some of the methods that could be used to address the problem include the following: sampling techniques (under-sampling and over-sampling), cost-sensitive techniques, kernel-based and active learning techniques, and ensemble-based techniques.

- On of the ways in which the performance of a classification task can be improved is by adopting a hybrid approach. Instead of using a single classifier, this approach combines different classifiers in order to benefit from their individual strengths and to improve the effectiveness and reliability of the classification models. In a future study, different hybrid approaches could be implemented and their performance compared against the same task.

- Regarding the English language background of our first year BSc. students, our preliminary results seem to suggest that students who did the English First Additional Language (ENB) in high school and obtain a score lower than 65%, may not proceed to second year. Research in the future could be conducted to investigate

this issue to determine whether a ENB score lower than 65% does indeed predict or indicate failure to proceed to second year.

It is perhaps appropriate at this point, as we conclude this chapter and document, to again highlight the most important findings emanating from our study, as follows:

- Regarding the relationship amongst gender, ethnicity and academic achievement (APS), we found
  - the main effect for gender on the APS not to be statistically significant,
  - a statistically significant main effect for ethnicity on the APS,
  - that the interaction between gender and ethnicity on the APS was not significant.
- In the relationship involving ethnicity, funding status and academic performance we found a statistically significant
  - main effect for ethnicity on the APS,
  - main effect for funding status on the APS, and
  - interaction between ethnicity and funding status on the APS.
- The APS obtained by first year STEM students who took the ENA subject in high school is higher and significantly different compared to the A PS obtained by first year STEM students who took the ENB subject in high school.
- Amongst the NSC Grade 12 STEM subjects, Mathematics (MAT) has the strongest and statistically significant contribution to the APS,
- Using two first-semester Computer Science I courses (BCO and IAP), we were able to show that
  - first year Computer Science I (CSI) students with prior computer programming experience obtain significantly higher scores in BCO and IAP compared to their peers who lack such experience,
  - BCO and IAP are together successful in predicting the overall, first year outcome
- The Faculty of Science minimum admission requirements subjects (Mathematics, Physical Sciences, English, APS) are together successful predictors of the overall, first year outcome. However, these subjects were not successful predictors of degree attainment.
- Mathematics I is the best and successful predictor of the overall, first year outcome. However, the subject failed to predict degree attainment successfully.

Some of the limitations associated with this study were identified, including several recommendations which different stakeholders in our university could implement. Finally, a few ideas were presented as the basis for future investigation and study. It is hoped that this research and its findings will lead to a better understanding of the first time first year STEM students in our university, their academic challenges and achievements in a very demanding academic environment such as ours.

Appendix

A Feature Selection and Regression Analysis Results

Table 1: Feature Selection Results

Algorithm	**Search Method**	**Selected Attributes**
CfsSubsetEval	GreedyStepwise	MAT, APM
	GeneticSearch	MAT, APM, CAT
	LinearForwardSelection	MAT, APM, CAT
	RandomSearch	MAT, APM
	ScatterSearchVI	MAT, APM, CAT
ClassifierSubsetEval	GeneticSearch	PSC
	RandomSearch	MAT, PSC, GEG, INT, AGS
ReliefAttributeEval	Ranker	CAT, INT, ENA
WrapperSubsetEval	GeneticSearch	PSC
	RandomSearch	MAT, PSC, GEG, INT, AGS

Table 2: Linear Regression Results

Predictor Variable	**Regression Coefficient**	**Significance (Yes/No)**
MAT	0.08088	Yes
PSC	0.01874	Yes
LSC	0.00405	Yes
APM	0.04999	Yes
ENB	0.06182	Yes
MP3	0.00967	Yes
ENA	0.05366	Yes
INT	-0.00414	Yes
GEG	-0.00493	Yes
AGS	-0.00467	Yes
CAT	-0.01117	Yes

B Student Characteristics

Table 3: Student Characteristics

Characteristic	**Description**	**Value**
Ethnicity	Ethnic Background	African Chinese Coloured Indian White
Gender		Male (M), Female (F)
Funding	Funding Status	Funded (F) non-funded (NF)
Admission Year	Year of Registration	2015 - 2019
Plan Code	Applied Chemistry Applied Computing Actuarial Science Animal, Plant and Environmental Sciences Astronomy Astrophysics Biological Sciences Computational and Applied Mathematics Chemistry Chemistry with Chemical Engineering Computer Science Geography and Archaeology Sciences General Geology Mathematics Mathematics of Finance Mathematical Sciences A Materials Science Nuclear Sciences and Engineering Physical Science Physics I (Major)	ACHE ACOM ACSI APES ASTR BIOL CAMS CHEM CHEM11 COMM GAES GEN GEOL MATH MFI MS1A MSCI NUCS PHSC PHYS

Table 3: Student Characteristics Continued

Characteristic	Description	Value
Home Language	Afrikaans	AF
	Bilingual English/Afrikaans	BIL
	Chinese (Cantonese)	CC
	Chinese (Mandarin)	CM
	Chinese (Other)	CH
	English	EN
	French	FR
	German	GE
	Greek	GR
	Hindi	HI
	Ndebele	NDB
	North Sotho/Pedi	NST/ PED
	Portuguese	PO
	Shona	SHO
	South Sotho/Sesotho	SST/SES
	Swahil/Kiswahilii	SI/KI
	Swazi/Siswati	SWA/SIS
	Tsonga/Shangaan	TSS
	Tswana/Setswana	TWA/SET
	Turkish	TU
	Urdu (Pakistan)	UR
	Tshivenda	VEN
	Xhosa	XHA
	Zulu	ZUL
APS	Admission Point Score	Numeric
English Type	English Home Language	ENA
	English First Additional Language	ENB
ENA	English Home Language	Numeric
ENB	English First Additional Language	Numeric
MAT	NSC Mathematics	Numeric
PSC	NSC Physical Science	Numeric
Outcome	Registration renewed with conditions (Wits Readmissions Committee-1)	M1C
	Registration renewed with conditions (Wits Readmissions Committee-2)	M2C
	Registration renewal refused (Wits Readmissions Committee-2)	M2E
	Exclusion waived-must return to the same year of study	MBR
	Failed minimum requirements - Needs permission to re-register in same Faculty	MBZ
	Permitted to Proceed	PCD
	Must return to complete requirements for year of study	RET
	Minimum requirements not met	MRNM

C Clustering and Association Rule Mining

C.1 Clusters

Table 4: k-means Cluster Analysis

Number of Clusters (k)	Within Groups Sum of Squares (wss)
1	24512.98
2	21775.43
3	19215.15
4	16680.15
5	16983.17
6	16152.78
7	15661.91
8	14330.59
9	15003.77
10	14662.08
11	14064.15
12	13813.49
13	13737.49
14	13334.41
15	13197.74
16	12836.25
17	12619.27
18	12553.50
19	12380.75
20	12319.43

Cluster centroids:

		Cluster#									
Attribute	Full Data	0	1	2	3	4	5	6	7	8	9	10
	(6700)	(500)	(239)	(522)	(636)	(1250)	(558)	(707)	(446)	(477)	(637)	(728)
Year	2016.7355	2017.742	2016.7029	2016.023	2016.0943	2017.0496	2016.1434	2017.3819	2017.0538	2016.3061	2016.4867	2016.717
FUNDING_STATUS	Funded	Funded	NonFunded	NonFunded	Funded	Funded	Funded	Funded	Funded	Funded	NonFunded	Funded
Gender	M	F	F	F	F	M	M	F	F	M	F	M
Race	BL	BL	BL	BL	BL	BL	BL	BL	IN	BL	WH	IN
Language	EN	SO	XH	EN	SE	SO	ZU	ZU	EN	SE	EN	EN
APS	44.3921	44.332	41.7448	41.5517	45.3428	45.124	43.7079	43.2984	46.7623	42.2746	42.6907	48.2624
MAT	71.7413	74.58	60.7573	43.7107	71.4371	78.5944	68.3961	71.1938	76.8789	62.7107	74.4662	85.4766
PSC	69.3072	74.07	57.8326	38.4272	71.0676	78.9344	65.2276	72.2065	72.0538	60.4486	66.0832	81.1305
Progress_outcome	PCD	PCD	MBR	PCD	PCD	PCD	PCD	RET	PCD	RET	PCD	PCD
ENGLISH_TYPE	ENB	ENA	ENB	ENA	ENB	ENB	ENA	ENB	ENA	ENB	ENA	ENA
ENGLISH	67.7676	71.532	63.8912	43.7126	69.2186	69.1912	60.9373	72.3663	74.9978	57.6981	74.3752	77.147
PLANCODE	GEN0	GEN0	BIOL	GEN0	GEOL	GEN0	BIOL	GEN0	BIOL	GEN0	GEN0	ACSI

Figure 1: Using Eleven Clusters (k=11)

C.2 Association Rules

```
Best rules found:

 1. PLANCODE=STAT 167 ==> Progress_outcome=PCD 167    <conf:(1)> lift:(1.57) lev:(0.01) [60] conv:(60.76)
 2. FUNDING_STATUS=Funded PLANCODE=STAT 162 ==> Progress_outcome=PCD 162    <conf:(1)> lift:(1.57) lev:(0.01) [58] conv:(58.94)
 3. Gender=M PLANCODE=STAT 119 ==> Progress_outcome=PCD 119    <conf:(1)> lift:(1.57) lev:(0.01) [43] conv:(43.29)
 4. FUNDING_STATUS=Funded Gender=M PLANCODE=STAT 114 ==> Progress_outcome=PCD 114    <conf:(1)> lift:(1.57) lev:(0.01) [41] conv:(41.47)
 5. Language=EN PLANCODE=STAT 99 ==> Progress_outcome=PCD 99    <conf:(1)> lift:(1.57) lev:(0.01) [36] conv:(36.02)
 6. FUNDING_STATUS=Funded Language=EN PLANCODE=STAT 95 ==> Progress_outcome=PCD 95    <conf:(1)> lift:(1.57) lev:(0.01) [34] conv:(34.56)
 7. Gender=M Language=EN PLANCODE=STAT 70 ==> Progress_outcome=PCD 70    <conf:(1)> lift:(1.57) lev:(0) [25] conv:(25.47)
 8. FUNDING_STATUS=Funded Gender=M Language=EN PLANCODE=STAT 66 ==> Progress_outcome=PCD 66    <conf:(1)> lift:(1.57) lev:(0) [24] conv:(24.01)
 9. FUNDING_STATUS=Funded PLANCODE=BIO1 86 ==> Progress_outcome=PCD 84    <conf:(0.98)> lift:(1.54) lev:(0) [29] conv:(10.43)
10. Language=ZU PLANCODE=ACSI 83 ==> FUNDING_STATUS=Funded 81    <conf:(0.98)> lift:(1.37) lev:(0) [21] conv:(7.91)
11. Gender=F Progress_outcome=PCD PLANCODE=ACSI 163 ==> FUNDING_STATUS=Funded 159    <conf:(0.98)> lift:(1.37) lev:(0.01) [42] conv:(9.32)
12. FUNDING_STATUS=Funded PLANCODE=HPHS 69 ==> Progress_outcome=PCD 67    <conf:(0.97)> lift:(1.53) lev:(0) [23] conv:(8.37)
13. PLANCODE=STAT 167 ==> FUNDING_STATUS=Funded 162    <conf:(0.97)> lift:(1.36) lev:(0.01) [42] conv:(7.96)
14. PLANCODE=STAT 167 ==> FUNDING_STATUS=Funded Progress_outcome=PCD 162    <conf:(0.97)> lift:(2.09) lev:(0.01) [84] conv:(14.88)
15. Progress_outcome=PCD PLANCODE=STAT 167 ==> FUNDING_STATUS=Funded 162    <conf:(0.97)> lift:(1.36) lev:(0.01) [42] conv:(7.96)
16. Gender=M Language=ZU PLANCODE=ACSI 64 ==> FUNDING_STATUS=Funded 62    <conf:(0.97)> lift:(1.36) lev:(0) [16] conv:(6.1)
17. Gender=F PLANCODE=ACSI 186 ==> FUNDING_STATUS=Funded 180    <conf:(0.97)> lift:(1.36) lev:(0.01) [47] conv:(7.6)
18. Gender=F Language=EN PLANCODE=ACSI 84 ==> FUNDING_STATUS=Funded 81    <conf:(0.96)> lift:(1.35) lev:(0) [21] conv:(6.01)
19. Gender=F Language=EN Progress_outcome=PCD PLANCODE=ACSI 80 ==> FUNDING_STATUS=Funded 77    <conf:(0.96)> lift:(1.35) lev:(0) [19] conv:(5.72)
20. PLANCODE=MICR 77 ==> Progress_outcome=PCD 74    <conf:(0.96)> lift:(1.51) lev:(0) [25] conv:(7)
21. Progress_outcome=PCD PLANCODE=ACSI 476 ==> FUNDING_STATUS=Funded 457    <conf:(0.96)> lift:(1.34) lev:(0.02) [117] conv:(6.81)
22. Language=EN PLANCODE=STAT 99 ==> FUNDING_STATUS=Funded 95    <conf:(0.96)> lift:(1.34) lev:(0) [24] conv:(5.66)
23. Language=EN PLANCODE=STAT 99 ==> FUNDING_STATUS=Funded Progress_outcome=PCD 95    <conf:(0.96)> lift:(2.06) lev:(0.01) [48] conv:(10.59)
24. Language=EN Progress_outcome=PCD PLANCODE=STAT 99 ==> FUNDING_STATUS=Funded 95    <conf:(0.96)> lift:(1.34) lev:(0) [24] conv:(5.66)
25. PLANCODE=ACSI 553 ==> FUNDING_STATUS=Funded 530    <conf:(0.96)> lift:(1.34) lev:(0.02) [135] conv:(6.59)
26. Gender=M PLANCODE=STAT 119 ==> FUNDING_STATUS=Funded 114    <conf:(0.96)> lift:(1.34) lev:(0) [29] conv:(5.67)
27. Gender=M PLANCODE=STAT 119 ==> FUNDING_STATUS=Funded Progress_outcome=PCD 114    <conf:(0.96)> lift:(2.06) lev:(0.01) [58] conv:(10.61)
28. Gender=M Progress_outcome=PCD PLANCODE=STAT 119 ==> FUNDING_STATUS=Funded 114    <conf:(0.96)> lift:(1.34) lev:(0) [29] conv:(5.67)
29. PLANCODE=BIO1 159 ==> Progress_outcome=PCD 152    <conf:(0.96)> lift:(1.5) lev:(0.01) [50] conv:(7.23)
30. PLANCODE=HPHS 153 ==> Progress_outcome=PCD 146    <conf:(0.95)> lift:(1.5) lev:(0.01) [48] conv:(6.96)
31. Gender=M PLANCODE=ACSI 367 ==> FUNDING_STATUS=Funded 350    <conf:(0.95)> lift:(1.34) lev:(0.01) [87] conv:(5.83)
32. Gender=F Language=EN PLANCODE=ACSI 84 ==> Progress_outcome=PCD 80    <conf:(0.95)> lift:(1.5) lev:(0) [26] conv:(6.11)
33. Gender=M Progress_outcome=PCD PLANCODE=ACSI 313 ==> FUNDING_STATUS=Funded 298    <conf:(0.95)> lift:(1.33) lev:(0.01) [74] conv:(5.6)
34. FUNDING_STATUS=Funded Gender=F Language=EN PLANCODE=ACSI 81 ==> Progress_outcome=PCD 77    <conf:(0.95)> lift:(1.49) lev:(0) [25] conv:(5.89)
35. Gender=F PLANCODE=HPHS 117 ==> Progress_outcome=PCD 111    <conf:(0.95)> lift:(1.49) lev:(0.01) [36] conv:(6.08)
36. FUNDING_STATUS=Funded Language=EN PLANCODE=ACSI 229 ==> Progress_outcome=PCD 217    <conf:(0.95)> lift:(1.49) lev:(0.01) [71] conv:(6.41)
37. Language=EN PLANCODE=ACSI 245 ==> Progress_outcome=PCD 232    <conf:(0.95)> lift:(1.49) lev:(0.01) [76] conv:(6.37)
38. FUNDING_STATUS=Funded Gender=M Language=EN PLANCODE=ACSI 148 ==> Progress_outcome=PCD 140    <conf:(0.95)> lift:(1.49) lev:(0.01) [45] conv:(5.98)
39. Gender=F PLANCODE=BIO1 108 ==> Progress_outcome=PCD 102    <conf:(0.94)> lift:(1.48) lev:(0.01) [33] conv:(5.61)
40. Gender=M Language=EN PLANCODE=ACSI 161 ==> Progress_outcome=PCD 152    <conf:(0.94)> lift:(1.48) lev:(0.01) [49] conv:(5.86)
41. PLANCODE=PHYS 89 ==> Progress_outcome=PCD 84    <conf:(0.94)> lift:(1.48) lev:(0) [27] conv:(5.4)
42. Gender=M PLANCODE=PHYS 71 ==> Progress_outcome=PCD 67    <conf:(0.94)> lift:(1.48) lev:(0) [21] conv:(5.17)
43. Gender=M Language=EN PLANCODE=STAT 70 ==> FUNDING_STATUS=Funded 66    <conf:(0.94)> lift:(1.32) lev:(0) [16] conv:(4)
44. Gender=M Language=EN PLANCODE=STAT 70 ==> FUNDING_STATUS=Funded Progress_outcome=PCD 66    <conf:(0.94)> lift:(2.03) lev:(0.01) [33] conv:(7.49)
45. Gender=M Language=EN Progress_outcome=PCD PLANCODE=STAT 70 ==> FUNDING_STATUS=Funded 66    <conf:(0.94)> lift:(1.32) lev:(0) [16] conv:(4)
46. FUNDING_STATUS=NonFunded PLANCODE=HPHS 84 ==> Progress_outcome=PCD 79    <conf:(0.94)> lift:(1.48) lev:(0) [25] conv:(5.09)
47. Language=EN Progress_outcome=PCD PLANCODE=ACSI 232 ==> FUNDING_STATUS=Funded 217    <conf:(0.94)> lift:(1.31) lev:(0.01) [51] conv:(4.15)
48. Language=EN PLANCODE=ACSI 245 ==> FUNDING_STATUS=Funded 229    <conf:(0.93)> lift:(1.31) lev:(0.01) [54] conv:(4.12)
49. PLANCODE=GENE 89 ==> Progress_outcome=PCD 83    <conf:(0.93)> lift:(1.47) lev:(0) [26] conv:(4.63)
50. FUNDING_STATUS=NonFunded PLANCODE=BIO1 73 ==> Progress_outcome=PCD 68    <conf:(0.93)> lift:(1.46) lev:(0) [21] conv:(4.43)
51. Gender=M Language=ZU Progress_outcome=MBR 100 ==> FUNDING_STATUS=Funded 93    <conf:(0.93)> lift:(1.3) lev:(0) [21] conv:(3.58)
52. FUNDING_STATUS=Funded PLANCODE=PHYS 70 ==> Progress_outcome=PCD 65    <conf:(0.93)> lift:(1.46) lev:(0) [20] conv:(4.24)
53. Gender=M Language=EN Progress_outcome=PCD PLANCODE=ACSI 152 ==> FUNDING_STATUS=Funded 140    <conf:(0.92)> lift:(1.29) lev:(0.01) [31] conv:(3.34)
54. Gender=M Language=EN PLANCODE=ACSI 161 ==> FUNDING_STATUS=Funded 148    <conf:(0.92)> lift:(1.29) lev:(0.01) [33] conv:(3.29)
55. Gender=F Language=EN PLANCODE=ACSI 84 ==> FUNDING_STATUS=Funded Progress_outcome=PCD 77    <conf:(0.92)> lift:(1.97) lev:(0.01) [37] conv:(5.62)
56. Gender=F PLANCODE=GENE 70 ==> Progress_outcome=PCD 64    <conf:(0.91)> lift:(1.44) lev:(0) [19] conv:(3.64)
57. Language=ZU Progress_outcome=PCD PLANCODE=GEN0 111 ==> FUNDING_STATUS=Funded 101    <conf:(0.91)> lift:(1.27) lev:(0) [21] conv:(2.89)
58. Gender=M Language=ZU Progress_outcome=PCD 322 ==> FUNDING_STATUS=Funded 290    <conf:(0.9)> lift:(1.26) lev:(0.01) [60] conv:(2.79)
```

FIGURE 2: Association Rules (Support=0.01, Confidence=0.9)

D First Year and Degree Attainment Prediction

D.1 Mathematics - Predicting First Year Performance

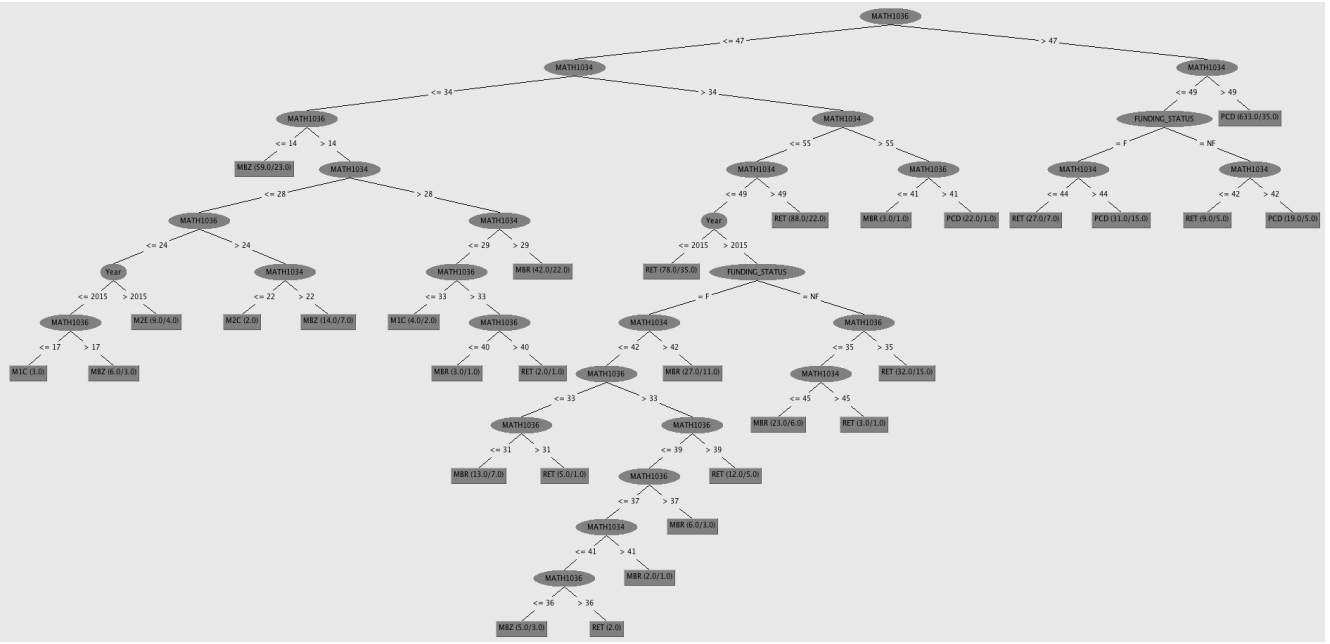


FIGURE 3: J48 Decision Tree: Predicting First Year Outcome

D.2 Mathematics - Predicting Degree Outcome

E Model Evaluation

E.1 Predicting First Year Outcome (Minimum Requirements)

```
@relation TESTQ5ENA

@attribute APS numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute ENA numeric
@attribute Progress_outcome {INP,M1C,M1E,M2C,M2E,MBR,MBZ,MRNM,PCD,RET}

@data
48,87,83,70,?
48,74,79,84,?
40,64,66,66,?
50,93,86,84,?
41,89,73,70,?
40,83,63,72,?
41,70,54,71,?
51,96,83,85,?
42,85,77,63,?
44,70,-1,70,?
42,71,71,80,?
46,84,83,75,?
```

FIGURE 4: English Home Language Test Set: Predicting First Year Outcome (483 records)

```
@relation TESTQ5ENB

@attribute APS numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute ENB numeric
@attribute Progress_outcome {INP,M1C,M1E,M2C,M2E,MBJ,MBR,MBZ,MRNM,PCD,RET}

@data
43,80,96,84,?
46,63,60,85,?
43,70,52,79,?
40,75,79,75,?
43,65,72,80,?
40,80,82,68,?
47,83,83,78,?
43,67,70,75,?
48,83,87,73,?
43,80,83,62,?
42,60,86,75,?
49,79,90,78,?
47,93,91,86,?
```

FIGURE 5: English First Additional Language Test Set: Predicting First Year Outcome (579 records)

E.2 Predicting Degree Outcome (Minimum Requirements)

```
@relation TestQ5degAtt


@attribute FUNDING_STATUS {F,NF}
@attribute APS numeric
@attribute MAT numeric
@attribute PSC numeric
@attribute ENA numeric
@attribute ENB numeric
@attribute OUTCOME {Q,NQ}

@data



F,48,86,88,71,-1,?
F,53,92,97,-1,83,?
F,52,96,93,69,-1,?
F,50,92,90,78,-1,?
F,55,97,91,90,-1,?
F,40,78,67,69,-1,?
NF,46,79,80,62,-1,?
NF,44,79,-1,75,-1,?
F,46,70,74,81,-1,?
F,45,82,88,63,-1,?
```

FIGURE 6: English Home Language and English First Additional Language Test Set: Predicting Degree Outcome (432 records)

E.3 Predicting First Year Outcome (Mathematics I)

```
@relation TestQ7outcome

@attribute Year numeric
@attribute FUNDING_STATUS {F,NF}
@attribute MATH1036 numeric
@attribute MATH1034 numeric
@attribute Progress_outcome {PCD,MBR,RET,MBZ,M1C,MRNM,M2C,M2E,M1E,MBJ,INP}

@data
2017,NF,35,31,?
2017,F,27,50,?
2017,F,50,55,?
2017,F,61,80,?
2017,F,50,50,?
2017,F,43,37,?
2017,F,24,31,?
2017,F,36,35,?
2017,F,40,52,?
2017,F,29,41,?
2017,NF,0,0,?
2017,F,75,64,?
2017,F,62,54,?
```

FIGURE 7: Mathematics I Test Set: Predicting First Year Outcome (474 records)

Bibliography

Abraham, I. O. and Nosa, A. L. (2018). Unemployment and output growth: Evidence from upper-middle-income countries in sub-saharan africa. *American Economic & Social Review*, 3(1):32–43.

Adeyemo, A. and Kuyoro, S. (2013). Investigating the effect of students socioeconomic/-family background on students' academic performance in tertiary institutions using decision tree algorithm. *Acta Phytotaxonomica et Geobotanica*, 4(2):61–78.

Agarwal, S., Pandey, G., and Tiwari, M. (2012). Data mining in education: data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 2(2):140.

Ajith, P., Sai, M., and Tejaswi, B. (2013). Evaluation of student performance: an outlier detection perspective. *Int. J. Innov. Technol. Explor. Eng*, 2(2):40–44.

Algarni, A. (2016). Data mining in education. *IJACSA) International Journal of Advanced Computer Science and Applications*, 7(6).

Ayalew, Y., Tshukudu, E., and Lefoane, M. (2018). Factors affecting programming performance of first year students at a university in botswana. *African Journal of Research in Mathematics, Science and Technology Education*, 22(3):363–373.

Azhar, M., Nadeem, S., Naz, F., Perveen, F., and Sameen, A. (2014). Impact of parental education and socio-economic status on academic achievements of university students. *European Journal of Psychological Research*, 1(1).

Babad, E. (2001). Students' course selection: Differential considerations for first and last course. *Research in Higher Education*, 42(4):469–492.

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Baker, R. S., Wang, Y., Paquette, L., Aleven, V., Popescu, O., Sewall, J., Rosé, C., Tomar, G. S., Ferschke, O., Zhang, J., et al. (2007). Educational data mining. *Data Mining and Learning Analytics: Applications in Educational Research*, pages 55–66.

Baloyi, B. and Isaacs, G. (2015). South africa's 'fees must fall'protests are about more than tuition costs. *CNN.com*.

Barlow-Jones, G., van der Westhuizen, D., and Coetzee, C. (2014). An investigation into the performance of first year programming students in relation to their grade 12 computer subject results. In *EdMedia+ Innovate Learning*, pages 77–83. Association for the Advancement of Computing in Education (AACE).

Belfield, C. R. and Crosta, P. M. (2012). Predicting success in college: The importance of placement tests and high school transcripts. ccrc working paper no. 42. *Community College Research Center, Columbia University*.

Bennedsen, J. and Caspersen, M. E. (2007). Failure rates in introductory programming. *AcM SIGcSE Bulletin*, 39(2):32–36.

Bhaird, C. M. A., Morgan, T., and O'Shea, A. (2009). The impact of the mathematics support centre on the grades of first year students at the national university of ireland maynooth. *Teaching Mathematics and its Applications: An International Journal of the IMA*, 28(3):117–122.

Bhambri, V. (2011). Application of data mining in banking sector. *IJCST*, 2(2):199–202.

Bhise, R., Thorat, S., and Supekar, A. (2013). Importance of data mining in higher education system. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, 6(6):18–21.

Bienkowski, M., Feng, M., and Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *Office of Educational Technology, US Department of Education*.

Bigus, J. P. and Bigus, J. (2001). *Constructing intelligent agents using Java*. Wiley New York.

Blackmore, C., Hird, K., and Anderton, R. S. (2021). An investigation of secondary school stem subjects as predictors of academic performance in tertiary level health sciences programs. *International Journal of Higher Education*, 10(1):76–84.

Blanchet, R. (2016). *Predicting First-Semester College Student Success at a Small Technology College*. PhD thesis, School of Education St. John Fisher College.

Bøe, M. V., Henriksen, E. K., Lyons, T., and Schreiner, C. (2011). Participation in science and technology: young people's achievement-related choices in late-modern societies. *Studies in Science Education*, 47(1):37–72.

Bovo, A., Sanchez, S., Héguy, O., and Duthen, Y. (2013). Clustering moodle data as a tool for profiling students. In *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE)*, pages 121–126. IEEE.

Bramer, M. (2007). *Principles of data mining*, volume 180. Springer.

Byrne, P. and Lyons, G. (2001). The effect of student attributes on success in programming. In *Proceedings of the 6th annual conference on Innovation and technology in computer science education*, pages 49–52.

Carpenter, E. E. (2013). First to second year retention based on financial aid package: A quantitative study. Master's thesis, Faculty of The Graduate College, University of Nebraska.

Carroll, C. A. and Garavalia, L. S. (2002). Gender and racial differences in select determinants of student success. *American Journal of Pharmaceutical Education*, 66(4):382–387.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9:13.

Chaudhari, B. and Parikh, M. (2012). A comparative study of clustering algorithms using weka tools. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 1(2):154–158.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886.

Chen, X. (2013). Stem attrition: College students' paths into and out of stem fields. statistical analysis report. nces 2014-001. *National Center for Education Statistics*.

Cochran, J. D., Campbell, S. M., Baker, H. M., and Leeds, E. M. (2014). The role of student characteristics in predicting retention in online courses. *Research in Higher Education*, 55(1):27–48.

Cohen, W. (1995). Fast effective rule induction in proc. of the 12th international conference on machine learning (a. prieditis and s. russell, eds.),(tahoe city, ca).

Crisp, G., Nora, A., and Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a stem degree: An analysis of students attending a hispanic serving institution. *American Educational Research Journal*, 46(4):924–942.

Dancey, C. and Reidy, J. (2008). Statistics without maths for psychology: Using spss for windows.

De Winter, J. and Dodou, D. (2011). Predicting academic performance in engineering using high school exam scores. *International Journal of Engineering Education*, 27(6):1343.

Der, G. and Everitt, B. S. (2008). *A handbook of statistical analyses using SAS*. CRC Press.

Dlomo, Z., Moses, M., Yu, D., and Jansen, A. (2011). Investigating the significance of the 2008 matric curriculum on first-year economics performance. *South African Journal of Higher Education*, 25(4):696–709.

Du, H. (2010). *Data mining techniques and applications: an introduction*. Cengage Learning Boston.

Durairaj, M. and Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies*, 5(4):5987–5991.

El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., and El Allioui, Y. (2019). A multiple linear regression-based approach to predict student performance. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 9–23. Springer.

Erdem, H. E. (2013). A cross-sectional survey in progress on factors affecting students' academic performance at a turkish university. *Procedia-Social and Behavioral Sciences*, 70:691–695.

Erdogdu, F. and Erdogdu, E. (2015). The impact of access to ict, student background and school/home environment on academic success of students in turkey: An international comparative analysis. *Computers & Education*, 82:26–49.

Essop, A. (2020). The changing size and shape of the higher education system in south africa, 2005–2017. *Johannesburg: University of Johannesburg*.

for Educators, S. A. C. (2010). A review of teacher demand and supply: Identifying research gaps and the role of sace.

Frederickson, N. and Petrides, K. (2008). Ethnic, gender, and socio-economic group differences in academic performance and secondary school selection: A longitudinal analysis. *Learning and individual differences*, 18(2):144–151.

Geiser, S. and Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. research & occasional paper series: Cshe. 6.07. *Center for studies in higher education*.

Great Britain. Department for Business, I. and (BIS), S. (2013). The benefits of higher education participation for individuals and society: key findings and reports:'the quadrants'. BIS Research Paper No. 146.

Group, C. . G. (2015). The economic benefits of vocational education and training in south africa. https://www.cityandguildsgroup.com/ /media/CGG Case Study.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hämäläinen, W. (2006). Descriptive and predictive modelling techniques for educational technology. *Licentiate thesis, Department of Computer Science, University of Joensuu.*

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

He, L., Zhou, G., Salinitri, G., and Xu, L. (2020). Female underrepresentation in stem subjects: An exploratory study of female high school students in china. *EURASIA Journal of Mathematics, Science and Technology Education*, 16(1).

Hegland, M. (2001). Data mining techniques. *Acta numerica*, 10:313–355.

Holden, E. and Weeden, E. (2003). The impact of prior experience in an information technology programming course sequence. In *Proceedings of the 4th conference on Information technology curriculum*, pages 41–46.

Holmes, K., Gore, J., Smith, M., and Lloyd, A. (2018). An integrated analysis of school students' aspirations for stem careers: Which student and school factors are most predictive? *International Journal of Science and Mathematics Education*, 16(4):655–675.

Huebner, R. A. (2013). A survey of educational data-mining research. *Research in higher education journal*, 19.

Hunt, K., Rankin, N. A., Schöer, V., Nthuli, M., and Sebastiao, C. (2009). Blind admission? the ability of nsc maths to signal competence in university commerce courses as compared to the former sc higher grade maths. https://mpra.ub.uni-muenchen.de/18075/.

Islam, M. M. and Al-Ghassani, A. (2015). Predicting college math success: Do high school performance and gender matter? evidence from sultan qaboos university in oman. *International journal of higher Education*, 4(2):67–80.

Jackson, R. (2002). The national student financial aid scheme of south africa (nsfas): How and why it works. *The Welsh Journal of Education*, 11(1):82–94.

Jacobs, M. (2010). *A framework for the placement of university students in science programmes.* PhD thesis, University of the Free State.

Jacobs, M., Williamson, J., and Pretorius, E. (2015). The relationship between selected grade 12 subjects and academic achievement of first year life science education students. In *ISTE INTERNATIONAL CONFERENCE ON MATHEMATICS,*

SCIENCE AND TECHNOLOGY EDUCATION, pages 2–11, Mopani Camp, Kruger Park. Unisa Press. This is a sample entry for a paper in conference proceedings.

Jindal, R. and Borah, M. D. (2013). A survey on educational data mining and research trends. *International Journal of Database Management Systems*, 5(3):53.

Joshi, S. and Shetty, S. P. (2015). Performance analysis of different classification methods in data mining for diabetes dataset using weka tool. *International Journal on Recent and Innovation Trends in Computing and Communication*, 3(3):1168–1173.

Joyce, C., Hine, G., and Anderton, R. (2017). The association between secondary mathematics and first year university performance in health sciences. *Issues in Educational Research*, 27(4):770–783.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72.

Kaleva, S., Pursiainen, J., Hakola, M., Rusanen, J., and Muukkonen, H. (2019). Students' reasons for stem choices and the relationship of mathematics choice to university admission. *International Journal of STEM Education*, 6(1):43.

Kaufinann, M. (2006). Data mining: Concepts and techniques. *Mining, What Is Data*, 10:559–569.

Kavitha, A., Kavitha, R., and Viji Gripsy, J. (2012). Empirical evaluation of feature selection technique in educational data mining.. *ARPN Journal of Science and Technology, 2*, 11.

Keengwe, J. (2007). Faculty integration of technology into instruction and students' perceptions of computer technology to improve student learning. *Journal of Information Technology Education: Research*, 6(1):169–180.

Khan, Y. A., Ahamad, Z., and Kousar, S. (2013). Factors influencing academic failure of university students. *International Journal of Educational Administration and Policy Studies*, 5(5):79–84.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540.

Kizito, R., Munyakazi, J., and Basuayi, C. (2016). Factors affecting student success in a first-year mathematics course: a south african experience. *International Journal of Mathematical Education in Science and Technology*, 47(1):100–119.

Koh, H. C., Tan, G., et al. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2):65.

Kori, K., Pedaste, M., Leijen, Ä., and Tõnisson, E. (2016). The role of programming experience in ict students' learning motivation and academic achievement. *International Journal of Information and Education Technology*, 6(5):331.

Kovačić, Z. J. and Green, J. S. (2010). Predictive working tool for early identification of 'at risk'students. *National Centre for Tertiary Teaching Excllence, New Zealand. Available at https://ako. ac. nz/assets/Knowledge-centre/RHPF-c11-tool-forearly-identification-of-at-risk-students/RESEARCH-REPORT-Predictive-Working-Toolfor-Early-Identification-of-At-Risk-Students. pdf.*

Kruger, N. (2016). Africa has a skills shortage. how do we fix it? https://www.weforum.org/agenda/2016/05/south-africa-skills-shortage-how-do-we-fix-it/.

Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Lee, C. L. and Mallik, G. (2015). The impact of student characteristics on academic achievement: Findings from an online undergraduate property program. *Pacific Rim Property Research Journal*, 21(1):3–14.

Lemnaru, C. (2012). Strategies for dealing with real world classification problems. *Unpublished doctoral dissertation, Faculty of Computer Science and Automation, Tehnical University of Cluj Napoca-Romania*.

Letseka, M. and Maile, S. (2008). *High university drop-out rates: A threat to South Africa's future*. Human Sciences Research Council Pretoria.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141.

Louw, C., Hofman, W., and Van Wyk, B. (2013). Mathematics: A powerful pre-and post-admission variable to predict success in engineering programmes at a university of technology. *Perspectives in Education*, 31(4):114–128.

Luan, J. (2002). Data mining and knowledge management in higher education-potential applications. In *in Proc. Workshop Assoc. Inst. Res. Int. Conf*, pages 1–18, Toronto, ON, Canada. ERIC.

Ma, J., Pender, M., and Welch, M. (2016). Education pays 2016: The benefits of higher education for individuals and society. trends in higher education series. *College Board*.

Manson, T. A. (2015). A relationship between matriculation english results and academic performance in nursing students at the kwa-zulu natal college of nursing. Master's thesis, University Of Kwazulu Natal.

Marnewick, C. (2012). The mystery of student selection: are there any selection criteria? *Educational Studies*, 38(2):123–137.

Mashige, K. P., Rampersad, N., and Venkatas, I. S. (2014). Do national senior certificate results predict first-year optometry students' academic performance at university? *South African Journal of Higher Education*, 28(2):550–563.

Mashiloane, L. (2016). Educational data mining (edm) in a south african university: a longitudinal study of factors that affect the academic performance of computer science i students. Master's thesis, University Of Witwatersrand.

Matsime, T. (2015). Data mining techniques and performance in a computer science i class. Technical report, Honours Research Report, School of Computer Science and Applied Mathematics (CSAM), University of the Witwatersrand, Johannesburg, South Africa.

McGhie, V. F. (2012). *Factors impacting on first-year students' academic progress at a South African university.* PhD thesis, Stellenbosch: Stellenbosch University.

Miller-Cotto, D. and Byrnes, J. P. (2016). Ethnic/racial identity and academic achievement: A meta-analytic review. *Developmental Review*, 41:51–70.

Modise, S. (2016). Comfort level, tutorial participation, financial status and academic performance in a computer science i classrooms. Technical report, Honours Research Report, School of Computer Science and Applied Mathematics (CSAM), University of the Witwatersrand, Johannesburg, South Africa.

Moodley, P. and Singh, R. J. (2015). Addressing student dropout rates at south african universities. *Alternation (Durban).*

Mthimunye, K. D. and Daniels, F. M. (2017). Performance in grade 12 mathematics and science predicts student nurses' performance in first year science modules at a university in the western cape. *Curationis*, 40(1):1–6.

Muremela, G., Kutame, A., Kapueja, I., and Adigun, O. T. (2023). Retaining scarce skills teachers in a south african rural community: an exploration of associated issues. *African Identities*, 21(4):743–759.

Murray, J. (2013). The factors that influence mathematics achievement at the berbice campus. *International journal of business and social science*, 4(10).

Naidoo, A. and McKay, T. (2018). Student funding and student success: A case study of a south african university. *South African Journal of Higher Education*, 32(5):158–172.

Naidoo, U., Flack, P. S., Naidoo, I., and Essack, S. Y. (2014). Secondary school factors relating to academic success in first year health science students. *South African Journal of Higher Education*, 28(4):1332–1343.

Nowak, A. and Dahal, G. (2016). he contribution of education to economic growth: Evidence from nepal. *International Journal of Economic Sciences*, 5(2):22–41.

OCDE, O. (2016). *OECD Employment Outlook 2016*. Organisation for Economic Co-operation and Development.

OECD, P. (2013). What are the social benefits of education? *Education indicators in focus*, pages 1–4.

Oosthuizen, A. (2019). Youth unemployment: statistics & solutions. *HSRC Review*, 17(4):3–8.

Örkcü, H. H. (2013). Subset selection in multiple linear regression models: A hybrid of genetic and simulated annealing algorithms. *Applied Mathematics and Computation*, 219(23):11018–11028.

Palmer, T.-A., Burke, P. F., and Aubusson, P. (2017). Why school students choose and reject science: A study of the factors that students consider when selecting subjects. *International Journal of Science Education*, 39(6):645–662.

Park, H. M. (2009). Comparing group means: t-tests and one-way anova using stata, sas, r, and spss. *The University Information Techology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University, IN, USA*.

Pather, S. and Dorasamy, N. (2018). The mismatch between first-year students' expectations and experience alongside university access and success: A south african university case study. *Journal of Student Affairs in Africa*, 6(1):49–64.

Pavlik Jr, P. I., Cen, H., and Koedinger, K. R. (2009). Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In *Educational Data Mining 2009*.

Petre, R.-S. et al. (2012). Data mining in cloud computing. *Database Systems Journal*, 3(3):67–71.

Piatek-Jimenez, K., Cribbs, J., and Gill, N. (2018). College students' perceptions of gender stereotypes: Making connections to the underrepresentation of women in stem fields. *International Journal of Science Education*, 40(12):1432–1454.

Pillay, N. and Jugoo, V. R. (2005). An investigation into student characteristics affecting novice programming performance. *ACM SIGCSE Bulletin*, 37(4):107–110.

Punlumjeak, W. and Rachburee, N. (2015). A comparative study of feature selection techniques for classify student performance. In *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 425–429. IEEE.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Radcliffe, B. (2017). How education and training affect the economy. http://www.investopedia.com/articles/economics/09/education-training-advantages.asp. Accessed: 2017-06-27.

Rahman, L., Setiawan, N. A., and Permanasari, A. E. (2017). Feature selection methods in improving accuracy of classifying students' academic performance. In *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 267–271. IEEE.

Rankin, N., Schöer, V., Sebastiao, C., and van Walbeek, C. (2012). Predictors of academic performance: National senior certificate versus national benchmark test. *South African Journal of Higher Education (SAJHE)*, 26(3):564–585.

Rauchas, S., Rosman, B., Konidaris, G., and Sanders, I. (2006). Language performance at high school and success in first year computer science. *ACM SIGCSE Bulletin*, 38(1):398–402.

Rojewski, J., Lee, I. H., and Gemici, S. (2012). Use of t-test and anova in career-technical education research. *Career and Technical Education Research*, 37(3):263–275.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Romero, C., Zafra, A., Luna, J. M., and Ventura, S. (2013). Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Systems*, 30(2):162–172.

Rylands, L. J. and Coady, C. (2009). Performance of students with weak mathematics in first-year mathematics and science. *International Journal of Mathematical Education in Science and Technology*, 40(6):741–753.

Sandlesh, K. (2015). Data mining algorithms: Course selection, recommendation and performance in first year, bachelor of science (bsc.). Technical report, Unpublished Honours dissertation, School of Computer Science and Applied Mathematics (CSAM), University of the Witwatersrand, Johannesburg, South Africa.

Schoer, V., Cliff, A., and Fleisch, B. (2015). When signals are lost in aggregation: A comparison of language marks and competencies of first-year university students. *South African Journal of Higher Education*, 29(5):156–178.

Sibanda, L., Iwu, C. G., and Benedict, O. H. (2015). Factors influencing academic performance of university students. *Demography and Social Economy*, 2(2):103–115. https://doi.org/10.15407/dse2015.02.103.

Siemens, G. and Baker, R. S. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254.

Singh, S., Malik, S., and Singh, P. (2016). Research paper factors affecting academic performance of students. *Indian Journal of Research*, 5(4):176–178.

Sommerville, T. and Singaram, V. (2018). Exploring demographic influences on students' academic performance over a five-year programme. *South African Journal of Higher Education*, 32(2):273–287.

Song, Y.-Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Spector, P. (2003). An introduction to the sas system. *Department of Statistics, University of California, Berkeley. URL: http://www. stat. berkeley. edu/classes/s100/sas. pd f.*

Stephen, D., Welman, J., and Jordaan, W. (2004). English language proficiency as an indicator of academic performance at a tertiary institution. *SA journal of human resource management*, 2(3):42–53.

Tafliovich, A., Campbell, J., and Petersen, A. (2013). A student perspective on prior experience in cs1. In *Proceeding of the 44th ACM technical symposium on Computer science education*, pages 239–244.

Tamilselvi, R. and Kalaiselvi, S. (2013). An overview of data mining techniques and applications. *International Journal of Science and Research (IJSR), India Online ISSN*, 2(2):2319–7064.

Tan, P.-N. et al. (2006). *Introduction to data mining*. Pearson Education India.

Telford, N. (2016). A web-based prototype course recommender system using apache mahout. Technical report, Unpublished Honours dissertation, School of Computer Science and Applied Mathematics (CSAM), University of the Witwatersrand, Johannesburg, South Africa.

Tewari, D. (2014). Is matric math a good predictor of student's performance in the first year of university degree? a case study of faculty of management studies, university of kwazulu-natal, south africa. *Int J Edu Sci*, 8(1):233–237.

Thiele, T., Singleton, A., Pope, D., and Stanistreet, D. (2016). Predicting students' academic performance based on school and socio-demographic characteristics. *Studies in Higher Education*, 41(8):1424–1446.

Tshabalala, T. and Ncube, A. C. (2014). Teachers'perceptions on the causes and effects of high rate teacher-turnover of mathematics and science teachers in rural secondary schools-a case study of nkayi district. *International Journal of Innovation and Applied Studies*, 8(1):149.

Van der Westhuizen, D. and Barlow-Jones, G. (2015). High school mathematics marks as an admission criterion for entry into programming courses at a south african university. *The Independent Journal of Teaching and Learning*, 10(1):37–50.

Van Rooy, B. and Coetzee-Van Rooy, S. (2015). The language issue and academic performance at a south african university. *Southern African Linguistics and Applied Language Studies*, 33(1):31–46.

Van Walbeek, C., Rankin, N., Sebastiao, C., and Schoer, V. (2012). Predictors of academic performance: National senior certificate versus national benchmark test. *South African Journal of Higher Education*, 26(3):564–585.

Vandenberg, J., Rachmatullah, A., Lynch, C., Boyer, K. E., and Wiebe, E. (2021). Interaction effects of race and gender in elementary cs attitudes: A validation and cross-sectional study. *International Journal of Child-Computer Interaction*, 29:100293.

Velmurugan, T. and Anuradha, C. (2016). Performance evaluation of feature selection algorithms in educational data mining. *Performance Evaluation*, 5(02).

Venkatas, I., Rampersad, N., and Mashige, K. (2014). Do national senior certificate results predict first-year optometry students' academic performance at university? *South African Journal of Higher Education*, 28(2):550–563.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., and Ortigosa, Á. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User modeling and user-adapted interaction*, 21(1):217–248.

Vipin, P.-N. T. M. S. et al. (2006). Introduction to data mining.

Vitale, J., Wanger, S., and Adams, D. (2010). Explaining student performance in an undergraduate agricultural economics classroom. *NACTA Journal*, 54(1):2–9.

Wang, M.-T. and Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (stem): Current knowledge, implications for practice, policy, and future directions. *Educational psychology review*, 29(1):119–140.

Wang, X. (2013). Why students choose stem majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, 50(5):1081–1121.

Wedderburn, L. A. (2005). The learning styles and approaches of students studying the fundamental algorithmic concepts course at the university of the witwatersrand. Master's thesis, School of Computer Science, University of the Witwatersrand.

Weiss, G. M. and Davison, B. D. (2010). Data mining. handbook of technology management, h. bidgoli.

White, G. and Sivitanides, M. (2003). An empirical investigation of the relationship between success in mathematics and visual programming courses. *Journal of information systems education*, 14(4):409.

Wilcox, C. and Lionelle, A. (2018). Quantifying the benefits of prior programming experience in an introductory computer science course. In *Proceedings of the 49th acm technical symposium on computer science education*, pages 80–85.

Wildschut, A., Megbowon, E., and Miselo, A. (2020). Impact of funding on academic performance: An exploration of two south african universities. *Journal of Education (University of KwaZulu-Natal)*, pages 29–49.

Wilson-Strydom, M. (2012). Using the nbts to inform institutional understandings of'under-preparedness': Implications for admissions criteria. *South African Journal of Higher Education*, 26(1):136–151.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39.

Wong, D. S. and Chia, Y.-M. (1996). English language, mathematics and first-year financial accounting performance: a research note. *Accounting education*, 5(2):183–189.

Yang, S. J., Lu, O. H., Huang, A. Y., Huang, J. C., Ogata, H., and Lin, A. J. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing*, 26:170–176.

Zaffar, M., Hashmani, M. A., Savita, K., and Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students academic performance. *International Journal of Advanced Computer Science and Applications*, 9(5).

Zatari, T. (2015). Data mining in social media. *Int. J. Sci. Eng. Res*, 6(7).

Zou, K. H., Tuncali, K., and Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3):617–628.