Investigating the Influence of First-Year Expectations and Experiences on Student Academic Performance

by

Elizabeth Mmapholo Booi

(35009500)

Submitted in accordance with the requirements of

Master of Science (Operations Research)

in the Department of Decision Sciences
University of South Africa

Supervisor: Prof. K.M. Malan and Dr. M.T. MaseTshaba

January 2024

Investigating the Influence of First-Year Expectations and Experiences on Student Academic Performance

by

Elizabeth Mmapholo Booi

(35009500)

E-mail: 35009500@mylife.unisa.ac.za

Abstract

This study examines the impact of first-year university students' expectations and experiences on their academic performance to enable early strategic interventions. The research is grounded in various theoretical frameworks, including Astin's theory of student involvement, Gardner's transition theory, Tinto's theory of student departure and Lizzio's framework of five senses of success, providing a comprehensive understanding of the transition of students to university life. The study follows a six-step Cross-Industry Standard Process for Data Mining (CRISP-DM). It involves data profiling of student demographics, academic attributes, expectations and experiences from a sample dataset of 2 054 records at the University of the Western Cape, South Africa.

Key findings reveal differences in academic performance across different demographics, with financial support significantly affecting outcomes. Factor analysis identified latent factors such as *effective learning, social well-being, academic support, and access to information*. The study found that the student performance models were not sufficiently robust for accurate predictions, with F1-scores below 60%. In contrast, academic outcome models, especially the random forest model, showed more promise, with F1-scores above 70%. Recommendations focus on targeted interventions, comprehensive orientation, enhanced academic support, and fostering an environment for social well-being. The study highlights the need for a multifaceted approach to student support, emphasising regular monitoring, evaluation and adaptability in interventions to create a supportive academic environment.

Keywords: educational data mining; random forest; exploratory factor analysis; student academic performance; first-year experience; first-year expectation; transition theory; student involvement theory; theory of student departure; five senses of success framework

Supervisor : Prof. K.M. Malan and Dr. M.T. MaseTshaba

Department : Department of Decision Sciences, University of South Africa

Degree : Master of Science (Operations Research)

'n Ondersoek na die effek van eerstejaarverwagtinge en ervarings op studente se akademiese prestasie

deur

Elizabeth Mmapholo Booi

(35009500)

E-pos: 35009500@mylife.unisa.ac.za

Opsomming

Hierdie studie doen ondersoek na die impak van eerstejaaruniversiteit-studente se verwagtinge en ervarings op hul akademiese prestasie ten einde tydige strategiese ingrypings te aktiveer. Die navorsing is op verskeie teoretiese raamwerke gegrond, wat insluit Astin se teorie van studentebetrokkenheid, Gardner se oorgangsteorie, Tinto se teorie oor studente wat opskop en Lizzio se suksesraamwerk van vyf gewaarwordinge. Die navorsing bied dus 'n omvattende begrip van studente se oorgang tot universiteitslewe. Die studie volg 'n ses-stap, kruisindustrie standaard proses vir dataontginning (Cross-Industry Standard Process for Data Mining [CRISP-DM]). Dit behels die datasamestelling van studentedemografie, akademiese eienskappe, verwagtinge en ervarings uit 'n steekproefdatastel van 2 054 rekords van studente aan die Universiteit van Wes-Kaapland, Suid-Afrika.

Deurslaggewende bevindinge dui op verskille in akademiese prestasie oor verskillende demografieë heen, met finansiële steun wat die uitkomste beduidend affekteer. Faktoranalise het latente faktore soos *effektiewe leer, maatskaplike welstand, akademiese ondersteuning en toegang tot inligting geïdentifiseer*. Die studie het bevind dat studenteprestasiemodelle nie sterk genoeg was vir akkurate voorspellings nie. F1-tellings was laer as 60%. Daarteenoor was die akademiese-uitkomstemodelle, veral die ewekansige bosmodel met F1-tellings van hoër as 70% meer belowend. Aanbevelings fokus op gerigte ingrypings, omvattende oriëntasie, verhoogde akademiese ondersteuning en die kweek van 'n omgewing wat maatskaplike welstand bevorder. Die studie vestig die aandag op die behoefte aan 'n veelvlakkige benadering tot studenteondersteuning en lê klem op gereelde monitering, evaluering en plooibaarheid van ingrypings ten einde n ondersteunende akademiese omgewing daar te stel.

Go dira dipatlisiso ka tlhotlheletso ya ditsholofelo tsa ngwaga wa ntlha wa dithuto gammogo le maitemogelo a tiragatso ya seakademiki a baithuti

ka ga

Elizabeth Mmapholo Booi

(35009500)

Imeili: 35009500@mylife.unisa.ac.za

Tshobokanyo

Patlisiso e e tlhatlhoba tshusumetso ya ditsholofelo tsa ngwaga wa ntlha wa dithuto wa baithuti ba yunibesithi gammogo le maitemogelo a bone a tiragatso ya seakademiki go ka ba kgontsha go dira ditogamaano go sa le gale. Patlisiso e e ikaegile mo matlhomesong a thuto a a mmalwa ao a akaretsang kgopolo ya Astin ya go nna le seabe ga baithuti, kgopolo ya phetogo ya Gardner, kgopolo ya Tinto ya go tsamaya ga baithuti, le letlhomeso la maikutlo a katlego a le matlhano la Lizzio, mme tsotlhe tse di tla tlamela ka kitso ya go tlhaloganya ka botlalo phetogo ya baithuti fa ba tsena mo botshelong jwa yunibesithi. Patlisiso e e latela magato a le marataro a Thulaganyo ya Tekanyetso ya Kgabaganyo-madirelo ya Kepadatha (CRISP-DM). E akaretsa go dira porofaele ya datha ya palo ya baithuti, dinonofo tsa seakademiki, ditsholofelo le maitemogelo go tswa seteng ya datha ya tseosekao ya direkoto di le 2 054 tsa kwa Yunibesithing ya Kapa Bophirima, Aforika Borwa.

Diphitlhelelo tse di botlhokwa di senola dipharologanyo magareng ga tiragatso ya seakademiki go kgabaganya dipalo tsa baithuti tse di farologanyeng, le tshegetso ya matlole eo e amang dipoelo segolo. Tlhatlhobo ya dintlha e supile dintlha tse di fitlhegileng jaaka *go ithuta sentle, boitekanelo jwa loago, tshegetso ya seakademiki, le phitlhelelo ya tshedimosetso*. Patlisiso e e fitlhetse gore dikao tsa tiragatso ya baithuti di ne di sa nonofela ponelopele e e tlhomameng, ka maduo a F1 a a ka tlase ga 60%. Mo pharologanyong, dikao tsa dipoelo tsa seakademiki, segolojang sekao sa random forest, se supile tshepiso e nngwe gape, ka maduo a F1 a a kwa godimo ga 70%. Dikatlanegiso di tsepamisa mogopolo mo ditsereganyong tse di lebilweng, molebo ka botlalo, tshegetso e e tokafaditsweng ya seakademiki, gammogo le kgodiso ya maemo a boitekanelo jwa loago. Patlisiso e tlhagisa tlhokego ya mokgwa wa dikarolo di le dintsi tsa tshegetso ya baithuti, go gatelela tekolo ya nako le nako, go tlhatlhoba le go fetofetoga ga ditsereganyo go tlhola maemo a tshegetso a seakademiki.

Declaration

I declare that *Investigating the Influence of First-Year Expectations and Experiences on Student Academic Performance* is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I submitted the thesis to originality checking software and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at the University of South Africa for another qualification or at any other higher education institution.

Signed:

Date:

Acknowledgements

I would like to acknowledge and express my special appreciation to my supervisors, Prof. Katherine Malan and Dr Mantepu MaseTshaba. Their mentorship, encouragement, and inspiration have been invaluable throughout my Master of Science journey. Prof. Malan, in particular, has provided unwavering support and has been instrumental in the progression of this research.

I also deeply appreciate my colleagues at the University of the Western Cape: Mr Larry Pokpas, Prof Sue Pather, Prof Liz Archer, Ms Lois Dippenaar and Dr Ian Johnson. Their invaluable support and assistance have significantly contributed to my research and academic journey.

My heartfelt gratitude extends to my family: Soyisile, Siyanda, and Lethabo Booi. Their unconditional support has been a constant source of strength and motivation throughout my studies. I dedicate this significant achievement to my parents, Sophia and Hendrick Rakgotho, as well as to my siblings, Mpho, Kgomotso, Marienkie, and Salphy Rakgotho, who have been my support pillars.

Contents

List of Figures iii

List of Tables v

1 Introduction 1

1.1 Problem Statement . . . 3
1.2 Research Objectives . . . 3
1.3 Significance of Study . . . 4
1.4 Limitation . . . 5
1.5 Thesis Outline . . . 5

2 Background 7

2.1 Introduction . . . 7
2.2 Measuring and Defining Student Academic Performance . . . 7
2.3 First Year Experience . . . 10
2.4 Exploratory Factor Analysis . . . 18
2.5 Data Mining Techniques . . . 24
2.6 Summary . . . 36

3 Literature Review 37

3.1 Understanding Features Affecting Student Performance . . . 37
3.2 Understanding First-Year Experience Features . . . 40
3.3 Data Mining Techniques . . . 44
3.4 Summary . . . 45

4 Methodology 47

4.1 Business Understanding . . . 47
4.2 Data Understanding . . . 48
4.3 Data Preparation . . . 50
4.4 Model Development . . . 61

4.5 Model Evaluation . . . 64
4.6 Deployment and Recommendation . . . 65
4.7 Ethical Consideration . . . 65
4.8 Summary . . . 66

5 Results **67**
5.1 Data Profiling . . . 67
5.2 Student Expectation - Predictive Models . . . 80
5.3 Student Experience - Predictive Models . . . 99
5.4 Expectation & Experience Gap - Models . . . 118
5.5 Summary . . . 139

6 Conclusion **143**
6.1 Introduction . . . 143
6.2 Analysis of the Results . . . 145
6.3 Recommendation . . . 152
6.4 Further Research . . . 156

References **158**

A Student Expectation and Experience Questions **170**

B Data Analysis Outputs **173**

C Ethical Clearance from the University of South Africa **179**

D Ethical Clearance from University of the Western Cape **182**

List of Figures

2.1 Tinto's Model of Student Integration [101, 102, 103] 11
2.2 Astin's Student theory of involvement postulates [3, 76] 14
2.3 Lizzio's Five Sense of Success framework [72] 15
2.4 Scree Plot 22
2.5 Phases of the CRISP-DM reference model [24] 25
2.6 Simplified CHAID Diagram 27
2.7 Classification and Regression Tree (CART) Example 27
2.8 C4.5 Algorithm Example 28
2.9 Confusion Matrix 35

4.1 Business Understanding 47
4.2 Data Understanding 48
4.3 Data Preparation 50
4.4 Missingness map of the Dataset 53
4.5 Scatterplot testing multivariate normality 56
4.6 Scree Plot 57
4.7 Model Development 61
4.8 Model Evaluation 64

5.1 Academic Performance and Outcome 68
5.2 Gender by Academic Performance and Outcome 68
5.3 First Generation by Academic Performance and Outcome 69
5.4 Population Group by Academic Performance and Outcome 69
5.5 Bursary by Academic Performance and Outcome 70
5.6 Residence by Academic Performance and Outcome 70
5.7 Programme by Academic Performance and Outcome 71
5.8 Scree plot incorporating the Kaiser criterion 78
5.9 Exploratory Factor Analysis Plot 79
5.10 Expectation Balanced Dataset Confusion Matrix 81

5.11 Expectation Balanced Dataset: Logistic Regression Model 83
5.12 Expectation Unbalanced Dataset Confusion Matrix 86
5.13 Expectation ROSE Dataset Confusion Matrix 88
5.14 Expectation Oversampling Dataset Confusion Matrix 92
5.15 Expectation Oversampling Dataset: Logistic Regression Model 94
5.16 Expectation Undersampling Dataset Confusion Matrix 97
5.17 Experience Balanced Dataset Confusion Matrix 100
5.18 Experience Balanced Dataset: Logistic Regression Model 102
5.19 Experience Unbalanced Dataset Confusion Matrix 104
5.20 Experience ROSE Dataset Confusion Matrix 107
5.21 Experience Oversampling Dataset Confusion Matrix 111
5.22 Experience Undersampling Dataset Confusion Matrix 114
5.23 Gap Balanced Dataset Confusion Matrix 119
5.24 Expectation and Experience Gap Balanced Dataset: Logistic Regression Model 121
5.25 Gap Unbalanced Dataset Confusion Matrix. 125
5.26 Gap ROSE Dataset Confusion Matrix. 127
5.27 Gap Oversampling Dataset Confusion Matrix 132
5.28 Gap Undersampling Dataset Confusion Matrix 135

6.1 Proposed First-Year Student Support Strategic Framework 154

List of Tables

2.1 Student Academic Performance Definition 10
2.2 Interpretation of Cronbach's Alpha 23

3.1 Factors and Features Used for Understanding Student Academic Performance 38
3.2 List of Data Mining Techniques Used to Predict an Outcome 46

4.1 Population and Sample 49
4.2 Description of Student Administration Dataset 49
4.3 Reliability Analysis for Student Expectation and Experience 55
4.4 Factor Selection 58
4.5 Reliability Analysis of Factors 58
4.6 Variables Coding 59

5.1 Summary Statistics Table for Interval and Ratio Variables by Performance 72
5.2 Summary Statistics Table for Interval and Ratio Variables by Outcome . . 73
5.3 Eigenvalues, Percentages of Variance, and Cumulative Percentages for Factors 79
5.4 Balanced Model Performance Measures for Student Expectation and Performance 80
5.5 Unbalanced Model Performance Measures for Student Expectation and Outcome 85
5.6 ROSE Model Performance Measures for Student Expectation and Outcome 87
5.7 Expectation ROSE Variable Importance in the Random Forest Model . . . 89
5.8 Oversampling Model Performance Measures for Student Expectation and Outcome 91
5.9 Expectation Oversampling Variable Importance in the Random Forest Model 93
5.10 Undersampling Model Performance Measures for Student Expectation and Outcome 96
5.11 Expectation Undersampling Variable Importance in the Random Forest Model 98

5.12 Balanced Model Performance Measures for Student Experience and Performance . . . 99
5.13 Unbalanced Model Performance Measures for Student Experience and Outcome . . . 103
5.14 ROSE Model Performance Measures for Student Experience and Outcome 106
5.15 Oversampling Model Performance Measures for Student Experience and Outcome . . . 109
5.16 Experience Oversampling Variable Importance in the Random Forest Model 112
5.17 Undersampling Model Performance Measures for Student Experience and Outcome . . . 113
5.18 Experience Undersampling Variable Importance in the Random Forest Model 117
5.19 Balanced Model Performance Measures for Gap and Student Performance . 118
5.20 GBM Variable Relevance Information . . . 123
5.21 Unbalanced Model Performance Measures for Gap and Academic Outcome 124
5.22 ROSE Model Performance Measures for GAP and Academic Outcome . . . 126
5.23 Gap ROSE Variable Importance in the Random Forest Model . . . 130
5.24 Oversampling Model Performance Measures for GAP and Academic Outcome 131
5.25 Gap Oversampling Variable Importance in the Random Forest Model . . . 133
5.26 Undersampling Model Performance Measures for GAP and Academic Outcome . . . 135
5.27 Gap Undersampling Variable Importance in the Random Forest Model . . . 138

6.1 Summary of the Student Performance (Above/Below Median) Models . . . 149
6.2 Summary of the Academic Outcome (pass/fail) Models Performance . . . 150

A.1 Description of Student Expectation and Experience Datasets. . . . 170
A.2 Description of Student Expectation and Experience Datasets. . . . 171
A.3 Description of Student Expectation and Experience Datasets. . . . 172

B.1 Summary Statistics of Student Expectation Profile . . . 173
B.2 Summary Statistics of Student Expectation Profile . . . 174
B.3 Summary Statistics of Student Experience Profile . . . 175
B.4 Summary Statistics of Student Experience Profile . . . 176
B.5 Student Expectation and Experience Gap Analysis . . . 177
B.6 Student Expectation and Experience Gap Analysis . . . 178

Chapter 1

Introduction

Transitioning from secondary education environment to a tertiary institution poses significant challenges for many students. This challenge is global; however, it is particularly pronounced in South Africa, where high poverty and inequality levels can worsen the difficulties experienced by first-year students [69]. This study examines the influence of first-year expectations and experiences on student academic performance. Many factors, including socio-economic status, previous educational achievement, psychological attributes, social integration into university life, and institutional support services, shape these expectations and experiences [63]. Understanding how these factors influence academic outcomes could provide essential insights for universities to improve their support initiatives for first-year students.

Numerous studies highlight the critical role that entry characteristics play in shaping a student's university experiences [62, 63, 65]. These entry characteristics include demographic factors such as age, gender, ethnicity, socio-economic background, high school educational achievement, individual attributes like self-esteem and motivation, and family background [102]. Upon matriculation, students' initial expectations are likely to be confronted by new experiences and both positive and negative experiences can influence their overall academic performance. The work done by Tinto [101] suggests that successful integration into academic and social life at college (university) are important determinants of student retention and success. Transitioning students need to form new social networks while balancing various non-academic responsibilities [102].

The gaps between initial expectation and actual experience during the first year in higher education are critical avenues that can significantly influence students' academic success and well-being. In this thesis, existing research served as foundational references to explore the multiple factors influencing first-year experiences in South African higher education. These studies used previously collected quantitative surveys administered at

two time points, capturing the beginning and end of the first year, and involved purposively sampled participants from diverse backgrounds such as those conducted in [82] and [83]. The insights from previous studies were expected to contribute valuable knowledge that will inform policy-making to improve student well-being and optimise student outcomes within South African higher education [82, 83].

Lizzio [72] studied the impact of student experiences and expectations on academic performance in Australia, emphasising the role of unmet expectations in contributing to dissatisfaction and poor academic outcomes. Braxton et al. [17] highlighted the potential limitations of research efforts, particularly in capturing variations in institutional cultures and socio-economic contexts. The present study aims to provide a theoretical framework for understanding the complexities of student experiences and expectations as they relate to academic performance in the South African context.

The comprehensive exploration of multiple factors influencing first-year experiences in South African higher education, as evidenced by the studies [82, 83], are crucial for addressing the structural inequalities and challenges that impact students' well-being and academic success. Integrating empirical evidence and insights from diverse social, economic and academic backgrounds can contribute to the broader goal of informing targeted student support to enhance student well-being and student outcomes.

The relationship between first-year students' expectations and experiences and their academic performance is a complex and pivotal factor in predicting their success in higher education. Kuh et al. [63] emphasise the importance of understanding the impact of students' expectations and experiences on their academic performance, whereas Pather et al. [82, 83] highlight the necessity of examining the gap between students' expectations and actual experiences to gain insights into student achievement and the effectiveness of academic institutions [102].

Furthermore, since the study aims to build a predictive model, the selection of statistical modelling techniques, such as logistic regression or decision tree is crucial for formulating predictive models [89]. The nature of the data and the degree of interpretability required should be considered when choosing these modelling techniques [17]. Proper application of these techniques can provide a more nuanced understanding of the influence of students' expectations and experiences on their academic performance, offering data-informed perspectives for decision-making within the higher education landscape [63].

1.1 Problem Statement

South African universities' throughput rates for three and four-year undergraduate degrees has been consistently low over the past years [31]. In 2019 study, among the entire cohort of enrolled students for the first time in 2000 in South Africa, only 44.3% graduated within five years. This shows a slight variation, with observed increases to 53.5% in 2006 and 59.1% in 2012, respectively [31]. Furthermore, the Department of Higher Education (DHET) noted between 2000 and 2006 a 30% increase in the number of students enrolling for a three-year degree in higher education institutions [31]. This enrollment increase has led to more entrants of first-year students who are either under-prepared for university-level studies or the universities not being well-equipped to support students [77].

Numerous studies have been undertaken to identify the factors influencing student success [81, 82, 83, 92, 97], and universities have implemented broad-based interventions. However, these measures have not resulted in a significant change in the number of students graduating within the prescribed minimum time, and there has been an increase in student attrition [31]. South African universities face the daunting task of addressing the issue of less than 30% of students graduating within the minimum time, and students who have been in higher education for a long time failing to complete their qualifications [31].

Furthermore, it is alarming that nearly half of the undergraduate students who enrol at a university never graduate. Hence, it is crucial to analyse the data of first-time entering students to understand the challenges they face. This analysis is important for implementing effective student interventions. Without appropriate targeted or holistic interventions to assist students in their transition through university, they will struggle to navigate the academic landscape [103] and, consequently, find it difficult to complete their qualifications.

The current study aims to build upon the existing research on student success and integrate it with research on first-year expectation and experience. This integration can potentially reveal new patterns or factors contributing to the existing body of research on predicting academic performance and success. It can provide a more comprehensive understanding of the determinants of academic performance, including the influence of academic support, social well-being, effective learning, access to information and the development of predictive models.

1.2 Research Objectives

The present study aims to investigate the influence of first-year expectations and experiences on the academic performance of first-time entering students in a formal qualification

at the University of the Western Cape, South Africa.
The specific objectives of this study are to:

i. Conduct a comprehensive data profiling of key areas: student demographics, academic attributes, expectations, and experiences.

ii. Implement exploratory factor analysis to determine the factors within first-year expectations and experiences that influence student academic performance.

iii. Investigate the feasibility of predicting student academic performance based on the factors identified through the analysis and demographic data.

iv. Formulate recommendations for developing a student intervention strategy based on the findings derived from the first three objectives.

1.3 Significance of Study

This study holds particular significance as it directly aligns with the institution's strategic focus on student experience and success. Furthermore, this research complements the shared goals of South African universities by providing data-driven recommendations. This study aligns seamlessly with the University of the Western Cape's goals to enhance the university experience for students, from initial registration to graduation. It seeks to achieve this through the deliberate creation of co-curricular activities, supportive services that are readily available, and cultivating an institutional ethos that encourages growth, development, and lifelong learning. Moreover, the study's significance extends to learning and teaching with the institution's commitment to providing high-quality, evidence-based learning and teaching opportunities, all rooted in responsive curricula and a rich diversity of learning, teaching, and assessment approaches. This comprehensive approach is instrumental in nurturing graduates with the knowledge, skills, and attributes necessary to excel in the dynamic world of work.

Furthermore, the insights derived from this study have the potential to inform and shape institutional strategies, thereby ensuring that students' academic journeys are not only rewarding but also result in timely graduation. By focusing on the first-year student's expectations and experiences, this study addresses a crucial gap in the existing research. It provides a nuanced understanding of the factors influencing academic performance, which is vital for developing effective student support interventions. Additionally, the outcomes of this study could assist the university in aligning its services and support structures with students' expectations and experiences, thereby fostering an environment

conducive to academic success. Therefore, this study holds significant implications for policy-making, institutional practices, and ultimately, student success in South African universities.

1.4 Limitation

This study's primary limitation lies in its reliance on self-reported Likert scale data, which risks response and social desirability biases. Participants might provide answers they perceive as expected or acceptable rather than their true opinions, impacting data accuracy and reliability. Additionally, self-reporting can result in missing data if participants skip items, potentially introducing non-random biases and affecting the completeness of predictive models.

The results obtained from this study may only apply to the specific population under investigation or similar populations with comparable characteristics. They should be cautiously extrapolated to others, considering cultural, demographic, or contextual differences.

The study's analysis depends on initial assumptions and chosen algorithms, which may bias results and affect their validity. While the developed models predict outcomes based on Likert scale responses, they do not establish causation between variables, limiting their interpretative scope to correlation.

1.5 Thesis Outline

The structure of the remaining chapters of this thesis is as follows:

- **Chapter 2** provides a comprehensive background on the concepts related to data mining techniques, the first-year experience, and various definitions of student academic performance.
- **Chapter 3** reviews the relevant literature, focusing on previous research on predicting student academic performance and the application of data mining methods in predicting student academic performance.
- **Chapter 4** outlines the study's methodology, including the a description of the data sources used, variables of interest, the data analysis plan, and ethical considerations.
- **Chapter 5** presents a discussion of the results and an evaluation of the predictive model developed to address the aim and objectives of the study. The chapter con-

cludes with a summary of the user evaluation of the predictive model's usability and usefulness.

- **Chapter 6** discusses how the study has addressed the research aim and objectives. It summarises the conclusions, discusses the limitations of the research, and outlines potential avenues for future work.

The appendices include the following:

- **Appendix A** student expectation and experience questions from the original study by Pather et al. [82]
- **Appendix B** data analysis outputs
- **Appendix C** ethical clearance from the University of South Africa (UNISA)
- **Appendix D** ethical clearance from the University of the Western Cape (UWC)

Chapter 2

Background

2.1 Introduction

This chapter 2 provides an introduction and definitions of key concepts for understanding the motivation for this study. The chapter begins by defining student academic performance (2.2), and explores the theoretical frameworks that underpin the first-year experience (2.3). Next, variable reduction using factor analysis is explained (2.4), along with relevant theoretical concepts of the data mining techniques used in the study to analyse the data (2.5).

2.2 Measuring and Defining Student Academic Performance

Student success and academic performance have been paramount concerns for higher education institutions [52, 69]. Universities are increasingly recognising the importance of using various measures and data sources, including academic performance, socio-demographics, and psycho-social behaviour, to monitor and enhance their programmes, modules, and student performance [53]. The next section explores student academic success in higher education.

2.2.1 Academic Performance Data

Academic performance data is crucial in assessing student success. This entails collecting and analysing students' performance data, such as grades, test scores, and completion rates. Such data allows universities to identify trends, patterns, and areas of concern regarding individual students or cohorts. By examining academic performance data, in-

stitutions can gain insights into the effectiveness of their teaching methods, curriculum design, and support systems. These quantitative indicators offer valuable insights into students' academic journeys and can help predict their future academic success.

2.2.2 Classifying Student Academic Profiles

In higher education, student success measures can be used to classify students into distinct profiles based on their academic performance, risk level, and engagement with the educational process. These classifications are valuable tools for institutions to identify students needing additional support and interventions. The following classifications are commonly used:

2.2.2.1 Note on Classification Overlap and Dynamics

It is important to note that these classifications are not mutually exclusive and can overlap. For example, a student may be both engaged and passing, or may move between classifications over time, such as from stopped to passing. The dynamic nature of student experiences necessitates a flexible approach to classification, recognising that student academic journeys can be complex and multifaceted. This understanding helps to develop more effective support strategies tailored to individual needs.

2.2.2.2 Passing Students

Passing students consistently meet or exceed the minimum academic requirements for their qualifications. They demonstrate satisfactory academic performance and are on track to complete their studies successfully. Identifying passing students is essential for acknowledging and reinforcing their achievements.

2.2.2.3 Graduating Students

Graduating students have completed their academic qualifications and are eligible for graduation. They would have met all their academic requirements and are prepared to enter the workforce or pursue further education. Understanding the characteristics and experiences of graduating students can inform strategies to enhance graduation rates.

2.2.2.4 Engaged Students

Engaged students are actively involved in their educational journey. They participate in extracurricular activities, engage in collaborative learning experiences, and demonstrate

a strong commitment to their studies. Engaged students tend to have a more enriching educational experience, which can contribute to their overall success.

2.2.2.5 Dropped-Out Students

Dropped-out students are those who have withdrawn from their academic qualifications before completion. Dropped-out students may face various challenges, such as academic difficulties, financial constraints, or personal issues, which hinder their progress. Identifying and understanding the reasons for student attrition is crucial for implementing retention strategies.

2.2.2.6 Stopped Students

Stopped students have temporarily interrupted their studies but have not officially withdrawn from their qualifications. They may take a leave of absence due to medical reasons, personal commitments, or other circumstances. Stopped students may eventually return to their studies, making it important to provide them with appropriate guidance and support.

2.2.2.7 Non-Graduating or Failing Students

Failing students or non-graduating students refers to students who do not meet the academic requirements necessary to complete a course, programme, or degree. These are students who have not achieved the minimum criteria for passing or earning a degree. The specific criteria for passing and graduating may vary depending on the university and the educational level (e.g., undergraduate or postgraduate). Generally, students who do not meet the required standards in terms of grades or coursework completion are considered failing students or non-graduating students.

2.2.3 The Grade Point Average (GPA)

A widely acknowledged indicator in higher education of student academic success is the Grade Point Average (GPA). The GPA is a quantitative measure of a student's cumulative academic performance, determined by allocating grade points to each course undertaken and subsequently computing their average. It is a valuable tool for evaluating a student's progress and determining whether they are eligible to advance to the next level of study. The assessment of academic performance often involves distinguishing between pass and fail outcomes or promotion statuses. Generally, a minimum module pass mark of 50% is

required for undergraduate progression, while postgraduate enrolment often necessitates an average above 60%.

2.2.4 Defining Student Academic Performance

In this study, student academic performance is defined through two distinct variables: Outcome and Performance (see table 2.1). The outcome relates to whether a student has successfully passed or failed their first year of studies. Performance is determined based on the GPA attained after the first academic year, categorised as either below (MB) or at least equal to (MA) the median GPA of the first-year cohort.

Table 2.1: Student Academic Performance Definition

Variables	**Data Values**	**Code**
Outcome	Pass (P)	1
	Fail (F)	0
Performance	Greater than or equal Median (MA)	1
	Less than Median (MB)	0

In higher education, student success is a multidimensional concept of academic performance data. Universities increasingly use these measures to classify students into profiles, including passing, graduating, engaged, dropped out, stopped, and non-graduating students. The GPA is a critical indicator of academic performance or progression. Understanding and defining student success is essential, as it enables universities to tailor support and intervention strategies to the unique needs of their student populations. By addressing these challenges and opportunities associated with student success, universities can foster a more inclusive and equitable learning environment that maximises the potential for all students to achieve their academic goals.

2.3 First Year Experience

The first-year experience in a university setting represents a pivotal stage in the academic journey of students. It plays a fundamental role in shaping their academic journey and future success. This initial phase is marked by a process of transition and the management of student expectations, necessitating a comprehensive understanding of their experiences to foster effective support mechanisms. In the pursuit of this understanding, various frameworks and research methodologies have been employed. Prominently, Tinto's Theory of Student Departure (2.3.1), Gardner's Framework for Understanding

First-Year Student Success (2.3.2), Astin's Student Theory of Involvement Model (2.3.3), and Lizzio's Five Senses of Success Framework (2.3.4) offer insightful perspectives. These frameworks collectively contribute to a comprehension of the first-year experience, guiding the development of initiatives aimed at enhancing student engagement and success during this critical period.

2.3.1 Tinto's Theory of Student Departure

Tinto's research has been instrumental in the difficulties faced by first-year students as they navigate the transition from high school environment to higher education [101, 102]. This transitional phase is marked by significant adjustments, including adapting to new academic expectations, unfamiliar learning environments, and establishing meaningful social connections within the university community. Tinto emphasises that the level of integration and participation experienced by students during this period significantly influences their academic persistence and success [103]. By fostering a sense of connection with peers, faculty, and the institution, universities can create a supportive and conducive environment for student satisfaction and achievement.

Tinto's framework underscores the importance of addressing key factors that impact students' experiences in their first year, including academic and social support, participation in campus activities, and the development of a sense of belonging [102, 103]. Figure 2.1 illustrates Tinto's Model of Student Integration. Essentially, Tinto's model explains that student success or persistence in higher education is largely determined by their level of integration. To facilitate these elements, institutions implement a range of strategies and initiatives, such as orientation, student advising, mentoring and extracurricular activities or programmes.

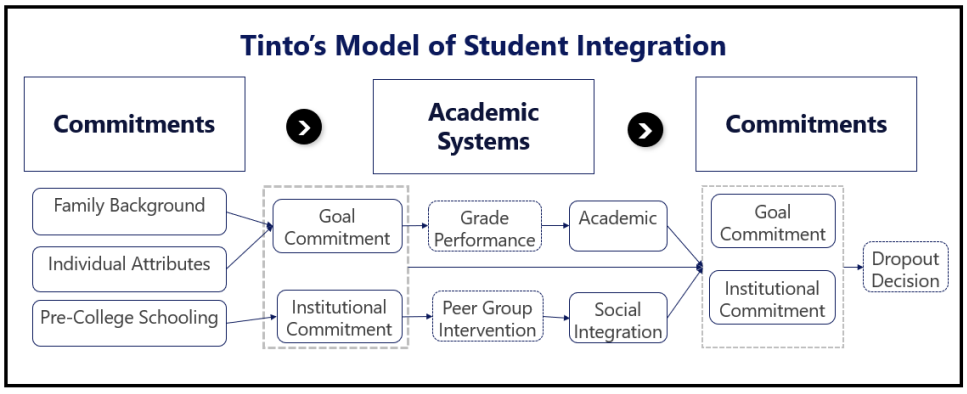


Figure 2.1: Tinto's Model of Student Integration [101, 102, 103]

Orientation programmes are designed to introduce students to the university's aca-

demic and social landscape, providing them with essential information and resources to navigate their new environment. Academic advising guides students through course selection, academic planning, and clarifying programme requirements. Mentoring initiatives, both formal and informal, connect first-year students with experienced peers, faculty, or staff members who can provide guidance and support. These programmes aim to create a supportive network and foster relationships that enhance students' sense of belonging and academic engagement. Additionally, universities offer many extracurricular opportunities such as clubs, societies, and community service initiatives, which promote students' engagement and forging social bonds beyond the academic setting.

By implementing comprehensive support structures and opportunities for participation, universities strive to enhance students' first-year experiences, promote their overall well-being, and increase their likelihood of completing their degrees successfully. These initiatives recognise the multifaceted nature of the first-year transition and seek to address the academic, social, and emotional dimensions of student life. By actively supporting students during this critical period, universities demonstrate their commitment to student success and provide a solid foundation for students to thrive throughout their academic journey.

In addition to Tinto's model 2.1, his theories explore the concepts of academic and social integration to elucidate the factors contributing to student success or attrition. Understanding these dynamics is crucial as students enter university with the expectation of completing their degrees. Students may have difficulty progressing academically when these expectations are not met. First-year students encounter various obstacles in their new university setting, such as adapting to intense academic expectations and responsibilities in their field. Arriving with preconceived ideas of university life and unmet expectations can heighten stress and anxiety for them [103].

Schlossberg's transition theory provides a valuable lens for understanding how students navigate the transition to university life. This theory focusses on the individual's ability to cope with change, considering factors such as the type of transition, the context, and the impact on the individual. Comparing this with Tinto's Theory of Student Departure can offer a more holistic view of the personal and situational variables affecting student retention and success [12, 90].

The notion of unmet expectations among first-year students can significantly impact their overall experience. Students may grapple with disillusionment and disengagement, hindering their academic progress and sense of belonging within the university community. Tinto's research underscores the importance of addressing these challenges and supporting students' transition into higher education. By fostering academic integration, universities

can create an environment that supports students in their academic pursuits, helping them navigate the academic demands and establish a sense of competence and mastery in their chosen field.

2.3.2 Gardner's Framework for Understanding First-Year Student Success

Gardner's Framework for Understanding First-Year Student Success outlines five critical dimensions essential for a smooth transition and effective engagement during the initial year of college or university [38]. These dimensions include intellectual, personal, social, emotional development, and physical well-being. This comprehensive approach acknowledges the complexity of the first-year experience, emphasising the need to provide academic, personal, and social requirements to strengthen their overall success. furthermore, social involvement is key to boosting student wellness and promoting a sense of community at university. Establishing substantial relationships with peers, faculty, and staff is crucial in building a supportive network that aids social and emotional development. These connections enable students to feel an integral part of the university community, nurturing a sense of identity and affiliation. This, in turn, plays a significant role in heightening their overall contentment and commitment to their academic journey.

Smith's comprehensive first-year engagement theory (2021) complements Gardner's framework by providing contemporary insight into factors that promote student retention and success [96]. Smith's theory underscores the importance of institutional support, student motivation, and engagement strategies. Comparing Smith's theory with Gardner's framework allows an evaluation of the evolving dynamics of student engagement and the effectiveness of various institutional practices. This comparison enriches understanding of first-year student success by highlighting the importance of tailored support mechanisms and proactive engagement strategies in fostering a conducive learning environment. These insights are vital for developing initiatives that improve student participation and success during this critical period [95, 96].

Universities can create a supportive environment that promotes academic and social integration by recognising the importance of students' expectations and proactively addressing their challenges. Implementing targeted interventions such as orientation programmes, academic advising, mentorship initiatives, and campus engagement opportunities can help alleviate stress and anxiety among first-year students. These initiatives provide the necessary resources and support networks to help students navigate their academic and social integration, fostering a sense of capability and facilitating a successful transition into higher education.

2.3.3 Astin's Student Theory of Involvement Model

In addition to Tinto's model, alternative theories like Astin's involvement framework consider both in-class and extracurricular elements in evaluating student achievement. Figure 2.2 shows Astin's Input-Environment-Outcome (I-E-O) Model which offers an in-depth framework for examining the various elements that affect student achievements [76].

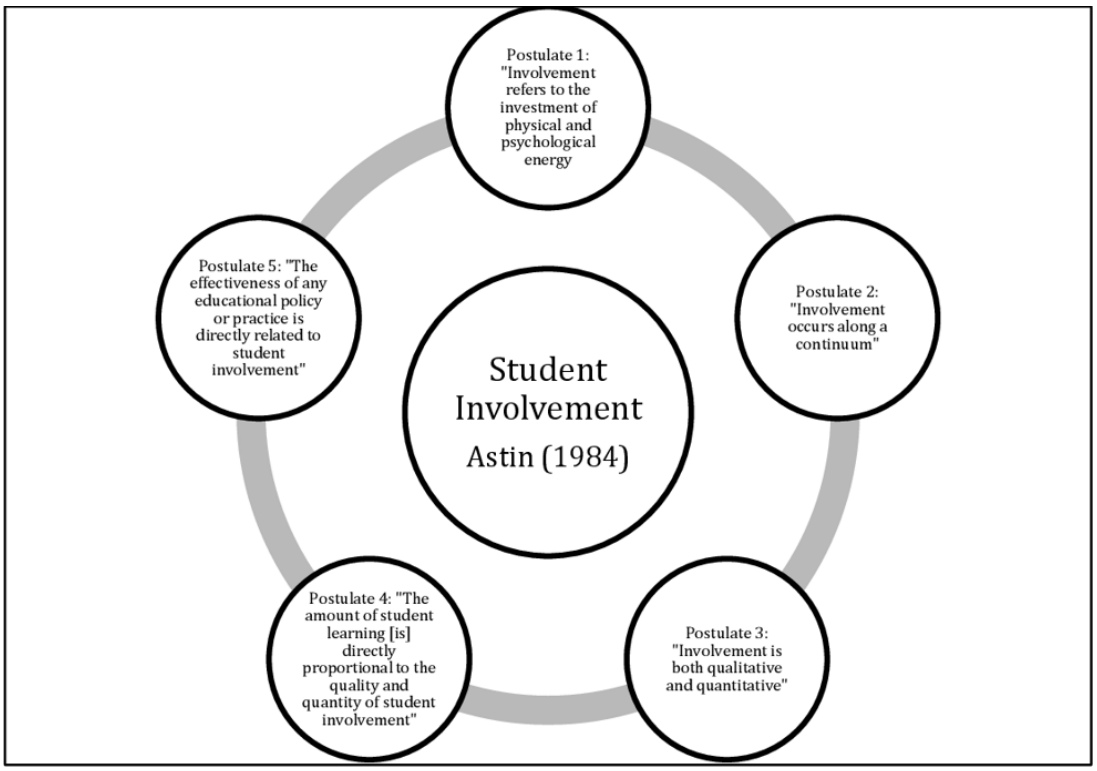


Figure 2.2: Astin's Student theory of involvement postulates [3, 76]

This model highlights the critical role of individual behaviours (inputs) surrounding conditions in moulding student experiences and their eventual outcomes. Interactions with staff, peers, and campus facilities (environment) are instrumental in shaping a student's initial year and their continued success. Recognising the impact of unmet expectations on first-year students is vital for universities aiming to facilitate their social and academic integration effectively. By implementing specific interventions and creating a nurturing environment, universities can alleviate stress and anxiety, enhance academic and social integration, and enable students to prosper in their university activities.

Astin's Student Involvement Theory explores how participation in co-curricular activities can lead to desired outcomes for higher education institutions. Engagement is positively associated with the ability to think critically, retain knowledge, and persist. A sense of belonging is related to self-worth, self-care, and reduced externalising problems

[3, 76]. Disruptive environments can negatively impact students' academic performance if they are distracted from their work. Consequently, first-year students often seek a sense of belonging, as they may experience doubts about their potential for success and completion of their studies [111].

Chickering's theory of identity development outlines seven vectors of development that students typically progress through during their tertiary years. These vectors include developing competence, managing emotions, and developing purpose. This theory complements Astin's Student Theory of Involvement by emphasising the developmental processes that underpin student engagement and participation in academic and extracurricular activities [26]. The integration of these theories highlights the multifaceted nature of student development, suggesting that both involvement in activities and progression through developmental stages are crucial for fostering a sense of belonging and academic success.

2.3.4 Lizzio's Five Senses of Success Framework

Although Tinto and Astin's theories remain valuable in describing students' experiences, they don't specifically target a particular moment, like the start of students' higher education journey. Lizzio's Five Senses of Success Framework (see figure 2.3), developed by Alf Lizzio at Griffith University in Australia presents a holistic model that addresses potential gaps in the early stages of the students' university experience [72]. This framework encompasses five key principles to support students' successful transition into university life. By strengthening all aspects of student transition, universities can overcome barriers, including the impact of disruptive environments on academic performance.

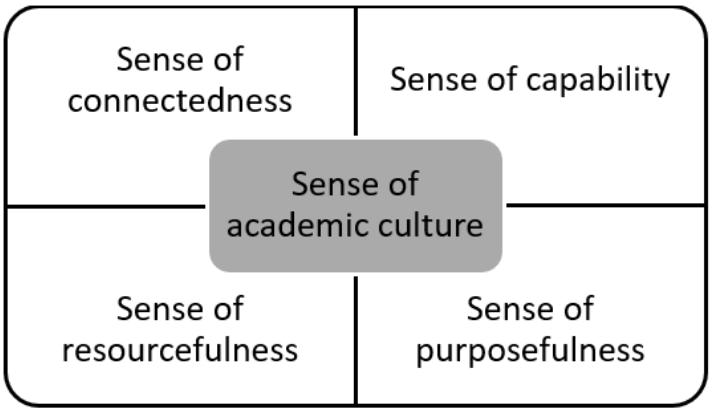


Figure 2.3: Lizzio's Five Sense of Success framework [72]

The Five Senses of Success Framework emphasises the importance of academic preparedness, self-awareness, connectedness, academic strategies, and a sense of purpose.

Academic preparedness encompasses the foundational knowledge and skills required for successful university studies. Self-awareness involves students understanding their strengths, weaknesses, and learning preferences, allowing them to adapt their approach to learning accordingly. Connectivity emphasises the importance of developing social connections with peers, faculty, and the broader university community to foster a sense of belonging and support. Using academic strategies involves equipping students with effective study skills and strategies to enhance their learning and academic performance. Lastly, a sense of purpose encourages students to cultivate personal goals and aspirations, providing direction and motivation throughout their university journey.

Bandura's social cognitive theory emphasises the role of self-efficacy and reciprocal determinism in shaping human behaviour [10]. Applying this theory to the context of first-year students highlights the importance of self-belief and the interactive influence of personal, behavioural, and environmental factors. This perspective aligns with Lizzio's Five Senses of Success Framework, particularly in understanding how students' beliefs and interactions with their environment contribute to their overall success [9, 10, 11]. Specifically, self-efficacy, or the belief in one's capabilities to achieve a goal, is crucial for academic preparedness and self-awareness. Reciprocal determinism, which posits that personal, behavioural, and environmental factors continuously interact, underscores the importance of connectivity and academic strategies. By integrating these elements, students are better positioned to develop a sense of purpose, thereby enhancing their academic and personal success.

By adopting the Five Senses of Success Framework, universities can proactively address the challenges when students enter higher education. Strengthening each aspect of the student's transition experience fosters an environment that supports their academic success and well-being. By removing barriers and ensuring a supportive and nurturing environment, universities can maximise students' potential for growth and achievement.

Sense of capability: A sense of capability is vital in preparing students for university life and equipping them with the necessary academic skills. It emphasises the importance of students achieving commendable grades from the beginning and mastering fundamental academic competencies. This is achieved by ensuring that students have essential prerequisites for higher education, such as basic knowledge and critical thinking skills. Additionally, fostering effective study habits and strategies, including time management and organisation, further nurtures students' sense of capability. Creating a supportive environment with accessible academic resources and personalised support also contributes to students' belief in their capabilities. Cultivating a sense of capability among students is crucial in preparing them for university, improving academic performance, and

fostering confidence and resilience in facing the challenges of higher education.

Sense of connectedness: Sense of connectedness in students refers to their ability to adapt to university life and establish meaningful relationships within the academic community. Students who move from schools that emphasise socialisation often perform better at university [108]. Hence, emphasising the significance of building relationships, accessing support, and developing peer networks is essential. Such a strategy promotes a feeling of community among students, enhancing their connection to the university and positively affecting their academic outcomes [3, 102]. Universities should promote student participation in curricular and co-curricular activities, as these facilitate student interaction and increase time spent on campus.

Sense of purpose: A sense of purpose in students refers to the clarity and determination with which they approach their academic and personal goals. Students who have a clear goal often find their academic efforts more fulfilling, as these pursuits align with their ambitions. This alignment improves their motivation and enjoyment of their studies. Furthermore, these students exhibit greater commitment and resilience when faced with academic challenges, understanding that these are part of their journey towards achieving their goals. This sense of purpose also allows for deeper engagement with their studies and the setting of personal goals, thereby driving them to excel in their academic journey.

Sense of resourcefulness: Sense of resourcefulness among students includes their propensity to independently gather information and use institutional support services, including seeking advice from lecturers and peers. It also involves forming and maintaining positive relationships within the university community. Additionally, resourceful students are proficient in identifying and effectively using resources contributing to their academic success. This proactive approach to education, which includes navigating challenges and seizing opportunities, plays a crucial role in a student's academic journey.

Sense of academic culture: Sense of academic culture refers to a student's understanding and appreciation of the values and norms within the educational environment and the wider university culture. This understanding extends beyond academic rules to include the ethos of intellectual curiosity and the pursuit of knowledge. A well-developed sense of academic culture can enhance academic aspirations, leading to deeper engagement with studies. It can also foster a sense of belonging within the university community, improving motivation and commitment. Studies, such as those by Freeman et al. [37] and Barefoot [13], have highlighted the positive correlation between a students' sense of academic culture and their engagement with academic work, underscoring its importance in promoting academic success.

When students' anticipations are often met, they tend to focus on their academic work

and hopefully perform well. It is important to focus on the broader student experience to support their all-inclusive development beyond just the academic content delivered in lectures. The first week of university is crucial as it sets a foundation for student's performance and satisfaction. Understanding the first-year experience is insufficient to grasp the full range of factors affecting students' underperformance or decision not to complete their qualification.

In selecting the theories for this study, the criteria focused on their relevance in understanding the complex nature of the transition of students to university life. The chosen theories—Astin's Student Involvement Theory, Gardner's Transition Theory, Tinto's Student Departure Theory, and Lizzio's Five Senses of Success Framework—were selected based on their empirical support, comprehensive frameworks, and applicability to various aspects of student engagement, retention, and success.

Astin's Student Involvement Theory was chosen for its emphasis on the importance of student participation in the educational process and how participation influences academic and personal development. Gardner's transition theory was included for its focus on the psychological and developmental processes that students undergo during significant life transitions, such as entering university. Tinto's theory of student departure was selected because of its well-established model explaining the factors that influence student persistence and attrition in higher education. Lastly, Lizzio's Five Senses of Success Framework was chosen for its holistic approach, capturing the diverse dimensions of student success and their interplay.

In addition to these primary theories, other applicable theories were considered to provide a broader perspective on student transitions and success. For example, Chickering's Seven Vectors of Identity Development offers insight into the stages of student development and identity formation during college years. Furthermore, Bandura's social cognitive theory highlights the role of self-efficacy and personal agency in student learning and achievement. Although these theories provide valuable information, the selected frameworks were deemed to be more directly aligned with the specific focus of this study on the transition to university life and the factors that contribute to student success and retention.

2.4 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is a statistical technique widely used in many research disciplines [35]. Its primary function is to contrast the latent relationships between the measured variables. This technique is particularly beneficial when researchers are con-

fronted with large data sets, such as those derived from Likert scale questionnaires with a substantial number of items [27]. EFA serves as a tool for simplifying data complexity by pinpointing a more concise group of factors responsible for explaining the relationships among variables [43].

In the domain of predictive modelling, the deployment of a 35-item Likert scale questionnaire can offer advantages and pose challenges. On the positive side, the richness and depth of the data obtained from such a comprehensive Likert scale questionnaire can provide a thorough understanding of the construct under investigation. It facilitates the capture of nuanced responses thereby increasing the model's predictive power.

However, dealing with data of high dimensionality presents notable difficulties. This situation can result in overfitting, a condition where a model, due to its excessive complexity, shows good performance on the training dataset but fails to generalise well to new, unseen datasets. The model may capture noise in the training data, mistaking it for a true underlying pattern. Furthermore, interpreting results can become complex with so many variables, making it difficult to draw clear and concise conclusions [98].

Here, the utility of EFA becomes evident. The EFA simplifies the data structure without significantly losing information by reducing the number of variables to a smaller number of more general factors. This reduction improves the interpretability of the model, making it easier to understand the relationships between the variables and the outcome of interest. Moreover, it mitigates the risk of overfitting by reducing the complexity of the model [43].

However, it is crucial to note that EFA has limitations. The process of factor extraction and deciding on the number of factors to retain can be subjective and relies heavily on the researcher's judgement. This introduces uncertainty and potential bias into the analysis [35]. Furthermore, EFA assumes that any observed correlation can be explained by underlying factors, which may not always be accurate. This assumption of common causality may not be valid and its violation can lead to misleading results [98].
The process of EFA can be broken down into several steps:

- **Data collection and preparation:** The first step in EFA is collecting and preparing data. This involves gathering responses from a sample of participants, typically through a questionnaire or survey. The data should be continuous or ordinal, with sufficient observations to perform a reliable analysis [98].

- **Computation of the correlation matrix:** The next step is to compute the correlation matrix of the variables. This matrix provides a summary of the pairwise relationships between the variables. It is the basis for the factor extraction process [43].

- **Factor extraction:** Extracting factors involves investigating the hidden variables or factors responsible for the observed correlations between variables. Different techniques, like Principal Component Analysis (PCA) or maximum likelihood, are available for this purpose. The selection of a specific method pivots on the data's characteristics and the aims of the research [35].

- **Factor rotation:** After extracting the factors, they are often rotated to achieve a simpler and more interpretable factor structure. Rotation can be either orthogonal (uncorrelated factors) or oblique (correlated factors). The choice between these two types of rotation depends on whether the factors are expected to be correlated [43].

- **Factor interpretation:** The final step in EFA is to interpret factors. This involves examining the factor loadings, which indicate the strength and direction of the relationship between each variable and the factor. Variables with high loadings on a factor are considered to be strongly associated with that factor [98].

- **Evaluating the results** of an EFA involves several considerations.

  - *Firstly*, the researcher should assess the suitability of the sample size. A common rule of thumb is to have at least five observations per variable. This guideline is based on the idea that each variable needs sufficient observations to estimate its properties and relationships with other variables [27].

    However, larger sample sizes can provide more reliable and stable factor solutions. This is because larger samples tend to provide more accurate estimates of population parameters and are more likely to meet the assumptions of factor analysis, such as multivariate normality [58]. Additionally, larger samples can better accommodate the complexity of the model, particularly when the number of factors and the number of items that load on each factor are large. Therefore, while a ratio of five observations per variable can serve as a minimum guideline, researchers should aim for larger sample sizes whenever possible to ensure the robustness and reliability of the factor analysis results [27].

  - *Secondly*, the researcher should examine the factorability of the correlation matrix. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and the Bartlett test can be used to assess whether the data are suitable for factor analysis [43].

    The KMO ranges between 0 and 1, where values close to 1 suggest that the correlation patterns among variables are fairly concentrated, making factor analysis likely to produce clear and dependable factors. Conversely, a KMO

value below 0.5 indicates that factor analysis might be unsuitable because of excessive unique variance or noise in the data [55].

Bartlett's test evaluates the correlation matrix to determine whether the variables do not relate and are thus not suitable for uncovering any underlying structure. A significant outcome ($p < 0.05$) in Bartlett's test indicates a notable deviation of the observed correlation matrix from an identity matrix, validating the use of factor analysis. It is critical to acknowledge that Bartlett's test reacts to the size of the sample, with large samples possibly producing significant outcomes even if the correlations among variables are minimal [14].

- *Thirdly*, The researcher must determine the number of factors to keep. Several guidelines, including the Kaiser criterion (eigenvalues over one), the scree plot, and parallel analysis, are available for this decision. The selection of these guidelines should be influenced by the goals of the research and how interpretable the factors are [35]. The Kaiser criterion, known as the eigenvalue-greater-than-one rule, is a widely used guideline in EFA for deciding how many factors to keep. Henry F. Kaiser introduced this criterion, which relies on the principle of eigenvalues within the context of factor analysis [55].

  Eigenvalues measure the variance explained by each factor. According to the Kaiser criterion, factors with an eigenvalue greater than one must be retained because they explain more variance than any single variable within the dataset. However, it is crucial to recognise that the Kaiser criterion, while a valuable initial guide, may overestimate the factor count, especially in larger datasets. Thus, it is advisable to complement it with additional techniques like the Scree Plot or parallel analysis for a more accurate determination of the ideal number of factors. [35].

  The Scree Plot shown in Figure 2.4 serves as a graphical tool used to identify the best number of factors to keep. This plot displays the eigenvalues associated with factors or components, arranged in descending order, through a line graph. The factors or components are organised from the first to the last along the X-axis, while the Y-axis shows their respective eigenvalues, reflecting the variance each factor contributes to the variables [35].

  In the Scree Plot, the factors are plotted in descending order of their eigenvalues. The plot typically starts with a steep slope, which then levels off to a point where the slope of the line becomes less steep, resembling a "scree". The point at which the slope of the line changes is often referred to as the "elbow" [23]. The "elbow", or the point of inflection on the plot, is used as a

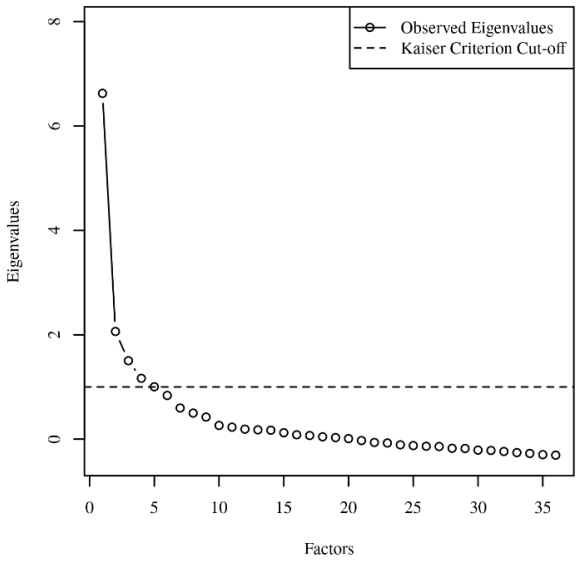


Figure 2.4: Scree Plot

criterion to decide the number of factors to retain. The factors to the left of the elbow (where the eigenvalues are relatively large) are retained. In contrast, factors to the right of the elbow (where the eigenvalues disappear) are generally ignored [35].

Parallel Analysis is a sophisticated and robust method used in Exploratory Factor Analysis (EFA) to determine the number of factors to retain. Introduced by Horn in 1965, it is considered one of the most accurate and reliable methods for deciding the number of factors, especially compared to other methods such as the Kaiser Criterion and the Scree Plot [50].

Parallel Analysis generates random data sets with the same number of variables and cases as the actual data. The eigenvalues are then calculated for these random data sets. The number of factors to retain is determined by comparing the eigenvalues of the actual data with those of the random data. Factors are retained as long as the actual eigenvalues exceed the corresponding percentile (usually the 95th) of the random-data eigenvalues. This method is based on the rationale that the factors extracted from the actual data should account for more variance than those extracted from the random data [48].

However, it is important to note that the interpretation of the Scree Plot can be somewhat subjective, as the 'elbow' is not always clear or easy to identify. Therefore, the Scree plot should be used with other criteria (such as the Kaiser criterion or parallel analysis) and substantive theory to decide the number of factors to retain [98].

- *Finally*, the researcher should evaluate the interpretability and reliability of the factors. This involves examining the factor loadings and the internal consistency of the items associated with each factor, often measured by Cronbach's alpha [98]. Cronbach's alpha was introduced by Lee Cronbach in 1951 and is a measure of internal consistency and reliability, widely used in psychometrics [28]. Figure 2.2 provides information on how well a set of items measures a single unidimensional latent construct.

Table 2.2: Interpretation of Cronbach's Alpha

Cronbach's Alpha (α)	**Interpretation**
0.00 to 0.20	Poor Internal Consistency / Poor Reliability
0.20 to 0.40	Fair Internal Consistency / Fair Reliability
0.40 to 0.60	Moderate Internal Consistency / Moderate Reliability
0.60 to 0.80	Good Internal Consistency / Good Reliability
0.80 to 1.00	Excellent Internal Consistency / Excellent Reliability

Cronbach's alpha is as a tool to assess the quality and reliability of identified factors. It assesses the internal consistency among the items associated with each factor. Typically, a value higher than 0.6 indicates that the items are closely interrelated and likely measure the same underlying construct [99].

Cronbach's alpha can also be used to evaluate the reliability of factor loadings, which are the correlations between the observed variables and the factors. High loadings and a high Cronbach alpha suggest that the factors are reliable and that the elements consistently measure the same underlying construct [43]. However, Cronbach's alpha has limitations, including the assumptions that all items are equally reliable and that they are all measuring a single underlying construct. Therefore, it should be used as one of several tools to comprehensively evaluate a scale's factor structure and reliability [99].

In conclusion, EFA is a complex process that requires careful consideration at each step; hence, it is important to evaluate the results and consider the suitability of the sample size, the factorability of the data, the number of factors to retain and the interpretability and reliability of the factors. Although using a 35-item Likert scale in predictive modelling can provide rich and detailed data, it also presents challenges due to high dimensionality; that is why EFA solves this problem by reducing the data into a manageable number of factors. This process is essential in enhancing the interpretability of the model and preventing overfitting [43]. However, using EFA should be carefully

considered, considering its assumptions and potential limitations.

2.5 Data Mining Techniques

Data mining is an analytical technique for exploring, characterising, and analysing data to reveal hidden patterns and insights [66, 107]. However, using incomplete data or inappropriate analytical methods can lead to errors, affecting the reliability of conclusions and the interpretation of models. Therefore, selecting the appropriate process, tools, and techniques is crucial for accurate data mining.

Several standard process models are widely recognised in data mining. These include SEMMA ("Sample, Explore, Modify, Model, Assess") developed by the Statistical Analysis System (SAS) Institute [88], the Cross Industry Standard Process for Data Mining (CRISP-DM) [24], and Knowledge Discovery in Databases (KDD) [56, 78]. KDD focuses on discovering new patterns through machine learning, statistics, and database technologies. While both SEMMA and KDD are comprehensive in their approach, they do not incorporate a stage for business understanding or extend to the deployment phase. They begin with the sampling process, bypassing the initial business understanding stage. In this regard, CRISP-DM distinguishes itself as a preferable process, as it includes the essential stage of business understanding and extends its reach to the deployment phase.

2.5.1 CRoss-Industry Standard Process for Data Mining (CRISP-DM)

This study employs the CRISP-DM methodology, following the lifecycle of a data mining project to develop predictive models. CRISP-DM is an adaptive and iterative approach, enabling the revisitation of each phase to improve the likelihood of developing a more precise predictive model. The process starts with a business understanding, progresses to data understanding, then modelling, and is followed by evaluation and deployment, details of which will be outlined subsequently.

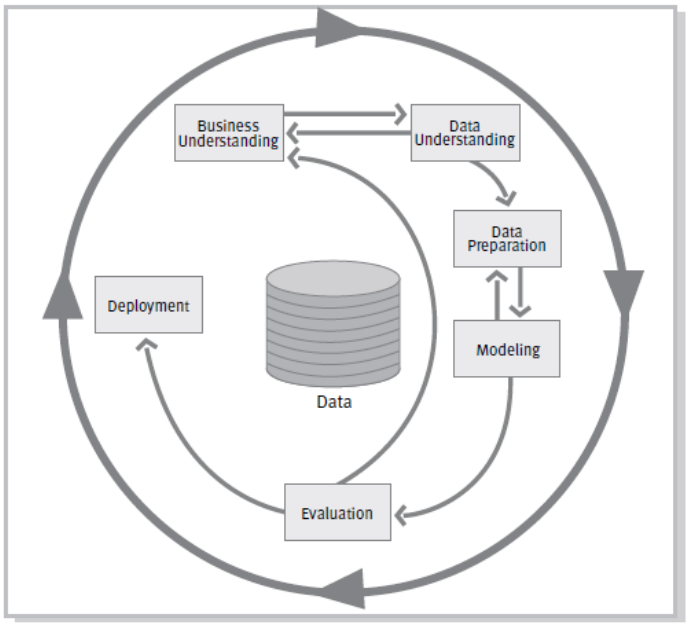


Figure 2.5: Phases of the CRISP-DM reference model [24]

There are six phases of CRISP-DM methodology which are shown in Figure 2.5:

- **Business Understanding:** This step involves clearly defining the research problem and setting specific objectives, ensuring that the data mining activities are directly targeted at addressing the core needs of the business. It is about translating the business requirements into actionable data-driven goals that guide the entire research process.

- **Data Understanding:** This stage is important for gaining familiarity with the data, which involves identifying patterns, anomalies, and trends within the dataset. The objective here is to develop a thorough understanding of the data regarding the research problem and objectives, laying the groundwork for effective data-driven decision-making.

- **Data Preparation:** This phase includes cleaning, selecting, describing, and transforming the data. The goal here is to refine the dataset so that it accurately addresses the research problem and objectives, ensuring that the data is in the best possible form for analysis and modelling.

- **Modelling:** This stage is focused on selecting and employing different types of predictive models that best align with the research problem and objectives. The aim is to leverage these models to uncover insights, ensuring that the chosen models are perfectly suited to the data and the goals of the research.

- **Evaluation:** This step is carried out in conjunction with the modelling phase, involving a thorough review and analysis of the models to ensure they meet the research problem and objectives. The purpose here is to validate the effectiveness and accuracy of the models before moving towards their deployment, making sure they are capable of providing the insights and predictions needed to solve the research problem.

- **Deployment:** This stage is about the insights generated from building the predictive model and organising them in a coherent, structured format. The aim is to ensure that the findings and the predictive model itself are accessible and usable, directly addressing the research problem and objectives and providing actionable recommendations. This ensures that the value derived from the predictive modelling process can be effectively applied to real-world decisions or further research.

The initial phase of the CRISP-DM process, focusing on business understanding, was discussed in previous section 1.1 (problem statement) and 1.2 (business objective). This foundational step is essential for guiding the subsequent data mining efforts, emphasising the necessity to identify data sources, the types of data needed to address the research questions, and determining whether the data is numerical or categorical. Understanding these aspects is crucial as it influences the selection of modelling techniques or methods to be employed.

In data mining, an excess of data can often weaken the effectiveness of the model by introducing too much noise, thereby diminishing its accuracy; hence, it is important to conduct a feature selection and extraction to refine the dataset. Feature selection involves choosing the most relevant variables, while feature extraction combines various attributes to form new ones, ultimately simplifying the dataset to a more manageable subset. Furthermore, the data must be divided into training and validation sets to effectively train and evaluate the model. For classification models, the target variable must be categorical, representing the outcome to be predicted, whereas input variables may be either numerical or categorical. This distinction helps in selecting an appropriate classification model, such as decision trees, logistic regression, naive Bayes, or support vector machines, which will be discussed in detail in the subsequent chapters and sections.

2.5.2 Data Mining Predictive Models

- **Decision trees:** are a popular data mining approach known for their ease of understanding and interpretability. They create a tree-like structure with nodes, each

representing a data split based on input variables. Decision trees are robust in handling missing data and can capture nonlinear relationships between input variables and one or more target variables. There are various algorithms used in decision tree models, including:

1. **Chi-square Automatic Interaction Detector (CHAID):** This algorithm is primarily used for categorical target variables and performs splits based on statistical significance.

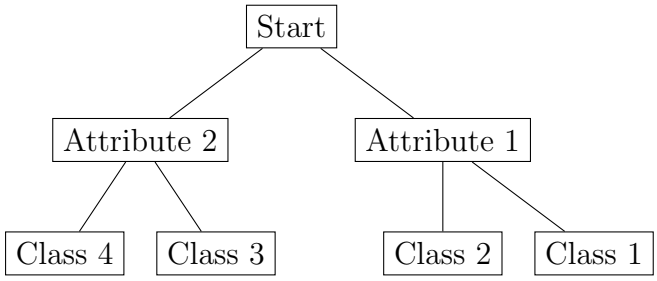


Figure 2.6: Simplified CHAID Diagram

2. **Classification And Regression Trees (CART):** CART is a versatile algorithm that can handle both categorical and continuous target variables. It employs measures like Gini impurity for classification and mean squared error for regression.

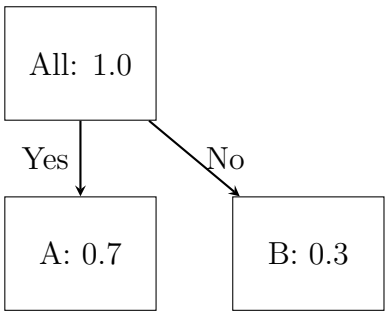


Figure 2.7: Classification and Regression Tree (CART) Example

3. **C4.5 Algorithm:** C4.5 is used for classification tasks and employs information gain to make splits. It can handle both categorical and continuous input variables.

One of the advantages of decision trees is their ability to handle missing values in the data without requiring imputation. When constructing a decision tree, it considers missing values as a separate category and makes decisions accordingly.

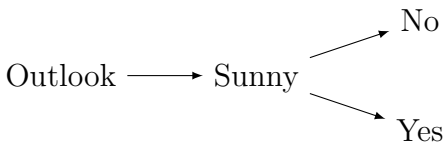


Figure 2.8: C4.5 Algorithm Example

Each node represents a test on a specific attribute or feature. The tree's structure is determined by recursively selecting the attribute that best splits the data based on certain criteria (e.g., Gini impurity, information gain). The data is divided into subsets at each node based on the attribute's values. A leaf node in a decision tree denotes a class or a prediction for the target variable. When a data point reaches a leaf node, it is assigned the class or value associated with that leaf.

While constructing a decision tree, the algorithm automatically ranks input variables based on their contribution to the tree's structure. Variables that best discriminate between classes are favoured for splitting. However, decision trees may become overly complex, leading to overfitting. Pruning can be applied to simplify the tree by removing nodes that do not significantly improve predictive performance. Pruning helps prevent overfitting and enhances the tree's generalisability.

Decision Trees are versatile and interpretable machine learning models that can handle missing data, accommodate nonlinear relationships, and are easy to understand. They provide valuable insights into feature importance and can be controlled through pruning to balance complexity and predictive accuracy.

- **Neural networks:** are widely used in data mining for various tasks, including pattern classification and nonlinear regression. Neural networks leverage the concept of artificial neurons to perform complex computations and discover patterns and relationships within data. They consist of interconnected layers of artificial neurons, nodes or units. These neurons process and transform input data, passing it through multiple layers to produce an output. Each connection between neurons has a weight associated with it, which determines the strength of the connection.

  The input layer receives the initial data features ($\mathbf{X}$), where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Here, $(X_1, X_2, \ldots, X_n)$ represent individual attributes or variables related to the dataset. For example, (X_1) might be age, (X_2) might be income, and so forth. There can be one or more hidden layers between the input and output layers. Each neuron in a hidden layer computes a weighted sum of the inputs and applies an activation function to produce an output. The output of a neuron in the i-th hidden

layer can be represented as:

$$h_i = f\left(\sum_{j=1}^{n} w_{ij} X_j + b_i\right) \tag{2.1}$$

Where: - h_i is the output of the i-th neuron in the hidden layer. - f is the activation function (e.g., sigmoid). - w_{ij} represents the weight associated with the connection between the i-th neuron and the j-th input feature. - b_i is the bias term for the i-th neuron.

The output layer produces the final predictions or classifications depending on the task. It produces a continuous output for regression tasks, whereas for classification tasks, it generates class probabilities.

In pattern classification, neural networks can classify data points into predefined categories or classes. By adjusting the weights of connections during training, neural networks can adapt and improve their ability to classify data accurately. For a binary classification task, the output y is determined using a threshold function:

$$y = \begin{cases} 1 & \text{if } h_i \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Neural networks are also valuable for nonlinear regression tasks. They can model complex, nonlinear relationships between input variables and output variables. This makes them useful for tasks where traditional linear regression models may be insufficient. Neural networks can approximate functions not easily captured by linear models, enabling accurate predictions and forecasting.

Neural networks are powerful tools in data mining, offering the capability to classify, predict, and forecast dependent variables. Their ability to capture complex patterns and relationships in data makes them indispensable in modern machine learning and data analysis. As data mining continues to evolve, neural networks remain at the forefront of innovation and research.

- **Logistic Regression:** is used to predict the likelihood of an outcome based on one or more predictor or independent variables. These independent variables can either be continuous or categorical, making logistic regression a versatile tool for binary classification problems [2]. The appropriateness of a model, particularly in logistic regression analysis, relies on its fit. Including predictor variables in a logistic

regression model can enhance the amount of variance accounted for in the log odds, making the selection of the model a critical step.[2].

For logistics with a binary dependent variable: 1 = Yes, 0 = No, then the probability will be

$$Pr\{Y = 1|\mathbf{X} = \mathbf{x}\} = \pi \tag{2.3}$$

All models will use a generalised linear model (GLM) model with β_0 as an intercept and β_k which will increase in log odds of $Y = 1$ when x_k increase by one unit when all other independent variables are held constant.

The logistic regression model is formulated as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2.4}$$

where π is the probability of the observation belonging to the target category, $\log(\cdot)$ is the natural logarithm, β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables, and $x_1, x_2, \ldots, x_n$ are the independent variables.

The left-hand side of the equation represents the log odds or the logit of the probability. By taking the natural logarithm of the odds, the log-odds are transformed into a linear combination of the independent variables on the right-hand side. The selection of input variables in a logistic regression model is crucial for its performance. Including relevant input variables can increase the variance explained in the log odds, leading to a better model fit.

On the contrary, including irrelevant or redundant variables can introduce noise into the model. The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are estimated from the data using techniques like Maximum Likelihood Estimation (MLE). The logistic regression model fits the data by finding the values of these coefficients that maximize the likelihood of the observed outcomes.

Logistic regression is a specific case of the more general Generalized Linear Model (GLM) framework. GLM expresses the relationship between the dependent and independent variables through a link function. For logistic regression, the link function is the logistic function (also known as the sigmoid function):

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} \tag{2.5}$$

where $P(Y = 1)$ is the probability of the dependent variable being in category 1, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables,

$X_1, X_2, \ldots, X_n$ are the independent variables, and e is the base of the natural logarithm.

The choice of input variables and the quality of the model fit are critical in logistic regression. The goal is to select relevant independent variables that contribute significantly to predicting the binary outcome. Model fit can be assessed through various statistical measures, such as the likelihood-ratio test, Wald test, and deviance statistic. Including appropriate independent variables in the logistic regression model can increase the variance explained in the log odds, leading to a more accurate prediction of the binary outcome.

- **Support Vector Machine (SVM):** It is used to perform classification and regression analysis, being particularly effective in binary classification tasks where the goal is to separate data points into two distinct classes while maximising the margin between them. SVM is known for its ability to handle non-probabilistic classification challenges [39]. In binary classification, SVM aims to find a hyperplane that best separates the data points of two classes. The hyperplane is chosen to maximise the margin between the two classes. The decision function for SVM can be represented as:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i \langle x, x_i \rangle + b\right) \tag{2.6}$$

Where: $f(x)$ is the decision function; x is the input data point; N is the number of support vectors; α_i are the Lagrange multipliers; y_i is the class label of the i-th support vector; x_i is the i-th support vector; and b is the bias term.

The Lagrange multipliers (α_i) are determined through the optimisation process, and support vectors are the data points that lie closest to the hyperplane. The optimal hyperplane is defined as:

$$w \cdot x + b = 0 \tag{2.7}$$

where w is the weight vector perpendicular to the hyperplane, and b is the bias term.

The key idea behind SVM is to maximise the margin between the two classes. The margin is the perpendicular distance between the hyperplane and the nearest support vectors from each class. Mathematically, the margin can be calculated as:

$$\text{Margin} = \frac{2}{\|w\|} \tag{2.8}$$

These support vectors are crucial because they are the only data points used to determine the optimal hyperplane. They essentially fix learning support the hyperplane, meaning that the position of the hyperplane is entirely dependent on these points. Data points that are not support vectors do not influence the placement of the decision boundary, and thus, can be considered less critical in the context of SVM. The importance of support vectors lies in their ability to define the margin and ensure that the hyperplane is maximally separated from the nearest points of each class, which helps in achieving better generalization to unseen data.

Moreover, in the case of non-linearly separable data, SVM employs a kernel trick to transform the input space into a higher-dimensional space where a hyperplane can be used to separate the classes. Even in this transformed space, the concept of support vectors remains pivotal. The support vectors are identified in the transformed space and play the same role in defining the decision boundary.

The objective function of SVM is to maximise this margin while minimising the classification error. It can be formulated as a constrained optimisation problem, often solved using Lagrange multipliers and quadratic programming methods. Support vector machines are a powerful tool for binary classification tasks, focusing on maximising the margin between classes. They are widely used in machine learning for their ability to handle non-probabilistic classification challenges effectively.

- **k-Nearest Neighbours (kNN):** is a non-parametric method used for classification and regression. The principle behind kNN is to find a predefined number of training samples closest in distance to the new point and predict the label from these.

  A common distance metric used in kNN is the Euclidean distance. For two points $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ in an n-dimensional space, the Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j)$ is given by:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{2.9}$$

  The prediction equation for the k-Nearest Neighbours (k-NN) algorithm is given by:

$$\hat{y} = \text{mode}\left(\{y_i : (x_i, y_i) \in N_k(x)\}\right) \tag{2.10}$$

In this equation, $\hat{y}$ represents the predicted output or class for a new input x. The set $\{(x_i, y_i)\}$ denotes the training data, where x_i are the feature vectors and y_i are the corresponding labels or responses. The notation $N_k(x)$ refers to the set of the k-nearest Neighbours to the input x based on a chosen distance metric, such as Euclidean distance. The prediction $\hat{y}$ is determined by taking the mode of the labels y_i of the k-nearest Neighbours. Mathematically, this means we count the frequency of each label among the k-nearest Neighbours and select the label that appears most frequently. If there is a tie, a tie-breaking rule or additional criteria may be used to decide the predicted class.

This approach is particularly effective in classification problems, as it bases the prediction on the majority vote of the closest Neighbours, thus leveraging local patterns in the data. The choice of k is crucial: a smaller k can capture more local variations, while a larger k provides a more general view, reducing the impact of noise but possibly missing finer details.

- **Naive Bayes:** is a probabilistic classifier based on Bayes' theorem with the assumption of independence among features.

  Bayes' theorem is stated as:

  $$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \tag{2.11}$$

  where y is the class variable and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are the feature variables.

  The naive Bayes classifier assumes that the features are conditionally independent given the class. Thus,

  $$P(\mathbf{x}|y) = \prod_{i=1}^{n} P(x_i|y) \tag{2.12}$$

  The classification rule is to predict the class y that maximizes $P(y|\mathbf{x})$:

  $$\hat{y} = \arg\max_y P(y|\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.13}$$

  In this context, $\arg\max_y$ means finding the class y that yields the highest value for the product $P(y) \prod_{i=1}^{n} P(x_i|y)$. This approach is commonly used in the Naive Bayes classifier, which relies on the assumption of conditional independence among features to simplify the calculation of the likelihood $P(\mathbf{x}|y)$.

- **Random Forests:** is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes of the individual trees.

Each decision tree is trained on a bootstrap sample of the training data. Let $\mathcal{D}$ be the training dataset. A bootstrap sample $\mathcal{D}_b$ is created by sampling $\mathcal{D}$ with replacement.

Each tree is grown by selecting the best split from a random subset of features at each node. Let $\mathcal{F}$ be the set of all features and $\mathcal{F}_b \subseteq \mathcal{F}$ be a random subset of features. The best split is chosen by maximizing a criterion such as the Gini impurity or information gain.

- **Gradient Boosted Machines:** are a powerful class of machine learning algorithms designed to enhance predictive performance through an iterative, stage-wise approach. In classification tasks, GBMs combine multiple weak learners, typically decision trees, into a single strong learner by sequentially optimising a chosen loss function. This method builds the model incrementally, where each new learner focuses on correcting the errors made by the ensemble of previous learners.

  The process of the GBM model can be described by the following equation:

$$\hat{F}_{m+1}(\mathbf{x}) = \hat{F}_m(\mathbf{x}) + \gamma_m h_m(\mathbf{x}). \tag{2.14}$$

  In this equation, $\hat{F}_{m+1}(\mathbf{x})$ denotes the updated ensemble model at iteration $m+1$, while $\hat{F}_m(\mathbf{x})$ represents the ensemble model from the previous iteration m. The term $h_m(\mathbf{x})$ is the new weak learner added at the m-th iteration, and γ_m is the learning rate, which moderates the contribution of the new learner to the updated model.

  Each iteration in the GBM process involves adding a scaled version of the new weak learner, $\gamma_m h_m(\mathbf{x})$, to the current model $\hat{F}_m(\mathbf{x})$. This iterative addition allows the model to adapt and improve by focusing on the residual errors from prior iterations. The choice of the learning rate γ_m and the weak learner $h_m(\mathbf{x})$ is critical, as they directly influence the model's ability to minimise prediction errors effectively.

  In practical applications, $h_m(\mathbf{x})$ is often selected to be the weak learner that best addresses the specific classification errors in the existing ensemble. This selection process ensures that each new learner incrementally enhances the overall model's performance, making GBMs a robust choice for complex classification tasks.

2.5.3 Data Mining Evaluation Measures

After implementing the modelling techniques, the effectiveness of these models is assessed using the statistical outcomes they produce. The confusion matrix emerges as a crucial instrument for generating metrics that help determine the most effective predictive model.

Figure 2.9 outlines the structure of the confusion matrix used for validating the models' accuracy and evaluating the performance of the classification model by comparing actual outcomes with predicted results.

		Actual	
		No	**Yes**
Predicted	**No**	True Negative (TN)	False Negative (FN)
	Yes	False Positive (FP)	True Positive (TP)

Figure 2.9: Confusion Matrix

- **Accuracy:** From the confusion matrix, the accuracy of the model is calculated by adding the true positive and the true negative and dividing by the grand total. The equation is

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.15}$$

- **Precision:** from the confusion matrix, the precision is calculated by dividing the true positive by the sum of the true positive and false positive (which can be referred to as total predicted positive), and the equation is

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.16}$$

  It demonstrates how accurate the model is in predicting those who are positive from the actual positive.

- **Sensitivity (Recall):** is calculated by dividing the true positive by the sum of the true positive and false negative (which can be referred to as total actual positive), and the equation is

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.17}$$

  It is the best measure to use if there is a high cost relating to the false negative. This is to prevent a loss or classifying incorrectly.

- **Specificity:** is calculated by dividing the true negative with the sum of the true negative and false positive (which can be referred to total actual negative), and the equation is

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.18}$$

  It is the best measure to classify an individual as negative correctly. It is the probability (as a percentage) that an individual is negative, given that the individual is negative.

- **F1-score:** provides a measure of the incorrectly predicted values, the weighted average of the recall and precision. It is useful when the data has imbalanced classes and is used to evaluate the models.

$$\text{F1-score} = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{2.19}$$

2.6 Summary

This chapter explained the key concepts and theoretical frameworks related to first-year students' academic performance. Furthermore, it discussed Tinto's model of student integration, Astin's student theory of involvement, and Lizzio's five senses of success framework, providing insights into the factors contributing to student success. It covered methodological considerations, including using exploratory factor analysis for variable reduction and applying data mining classification techniques and Models. These techniques help identify patterns and relationships within the data, offering valuable insights into student performance. This chapter laid a solid foundation for understanding the complexities of first-year student academic performance, highlighting the importance of academic and social integration, student involvement, and a holistic view of success. These concepts and frameworks will guide the subsequent first-year student academic performance analysis.

Chapter 3

Literature Review

This chapter provides a discussion on understanding the factors influencing student performance (3.1), the features of the first-year experience (3.2), and the application of various data mining techniques (3.3). The chapter reviews the findings from previous studies, identifies gaps in the existing body of knowledge, and establishes a solid foundation for the current research.

3.1 Understanding Features Affecting Student Performance

Research shows that predicting student academic performance is not just about grades and personal information. There are other important factors. Instead of focusing only on demographics, this study looks at different aspects that influence how well students do in their studies. These aspects include social factors, such as how well they fit into the university community, academic factors, like their motivation and integration into the courses, and wellness aspects [3, 63, 82, 101, 102, 103].

In this section, the study explores the factors researchers have found important to student success. The goal is to identify gaps and highlight the key factors to consider in the study to assess how these factors impact the predictive models. A better understanding of these factors will assist in creating more reasonable models that consider the complex mix of factors that impact students' academic performance. A comprehensive overview of these features used to predict student academic performance is provided in Table 3.1 below.

Table 3.1: Factors and Features Used for Understanding Student Academic Performance

Type	Features	Previous Studies
Background Information	Gender	[1, 15, 42, 56, 61, 62, 97, 113]
	Race	[18, 30, 42, 56, 60, 61, 62, 75, 97, 113]
	Financial aid	[7, 16, 62, 97, 103, 106]
	Socio-economic status	[18, 106]
Pre-university Information	Type of school	[42, 56, 60, 61]
	High school mathematics	[42]
	High school admission score	[1, 7, 15, 18, 42, 56, 60, 62, 97]
University Information	Course level	[15, 60, 61, 113]
	Grade point average (GPA)	[1, 7, 15, 42, 51, 113, 75, 94]
	Course attendance	[62, 63, 97, 94]
	Technological support	[60, 62, 97]
	Social integration	[4, 18, 60, 62, 63, 81, 82, 85, 97, 103]
	Academic integration	[4, 18, 60, 62, 81, 82, 85, 97, 103]
	Resourcefulness	[81, 82]
	Motivation	[60]
	Institutional commitment	[18, 103]
	Leadership and self-efficacy	[85, 100]

Previous studies examined various features to predict student performance. They considered students' backgrounds, like their socio-economic status, cultural background, and past educational experiences [30, 60]. Motivation has also been studied as a significant factor in academic success [60]. Additionally, how well students integrate into the academic community and social environment [4, 85] can play a large role [60, 15, 82, 81]. With technology becoming more important in education, researchers have also investigated how technological support affects student performance [60].

Demographic information, such as gender, race, and socio-economic status, impacted students' academic performance [60, 94]. Moreover, academic factors like the level of courses taken, credit ratings, previous education, and class attendance also affect how well students do.

Some studies have shown that non-academic factors, like leadership skills and self-efficacy [85], can significantly affect students' grades, but the impact may vary for different groups of students [100]. Other research suggested that some non-academic factors may

have less influence on the academic success of particular student groups, like first-year female engineering students [18]. This suggests that the influence of these non-academic factors may differ depending on the specific group of students.

The traditional approach of only looking at academic and demographic data might not be sufficient. Many non-academic factors can affect students' performance, such as their well-being, mental health [4], family relationships, and friendships [110]. Unfortunately, these non-academic factors are often self-reported and can change over time, making it challenging to study. To create a more acceptable predictive model, one must consider academic and non-academic factors and understand how they interact and change over time.

This review identifies significant gaps and limitations in the current understanding of factors influencing student performance. While the focus on transition features, including social, academic, and wellness aspects is essential, it may overshadow other non-academic factors that play important roles in academic success, such as personal motivation, emotional intelligence, and socio-economic background. To gain a more comprehensive understanding, it is imperative to consider a wider array of academic and non-academic variables.

The predominance of quantitative research methods, particularly those relying on surveys and self-reported data, offer valuable insights but may be susceptible to response biases. To enrich the analysis, incorporating qualitative methods, such as in-depth interviews and case studies, it is necessary to explore students' experiences and perspectives.

Moreover, the reviewed studies often focus on specific student populations and academic disciplines, potentially limiting the generalisability of findings [84]. A broader investigation across diverse contexts would yield a more comprehensive understanding of factors affecting academic success. Considering the temporal dimension of student experiences is vital. The academic journey is dynamic, with factors evolving over time and impacting students differently at various stages. A longitudinal approach or analysis of data at multiple points in a student's academic trajectory would provide valuable insights into these changes.

Furthermore, the possibility of publication bias in the reviewed literature should be acknowledged. To counter this, conducting a systematic review that includes unpublished studies or grey literature would provide a more balanced overview.

This literature review reveals important gaps and limitations in the current understanding of factors influencing student performance. Addressing these issues through an inclusive approach, incorporating qualitative methods, and considering diverse contexts and temporal dimensions will enhance understanding and support the development of

effective educational interventions to promote academic achievement. This emphasises the importance of considering various factors, including social, academic, and wellness, as well as demographic and non-academic variables when predicting student academic performance. By considering all these factors, this study aims to develop a more acceptable model that better reflects the complex nature of student success.

3.2 Understanding First-Year Experience Features

The first-year experience (FYE) at universities is a critical phase in students' academic journey. It includes various programs, initiatives, and support systems designed to help students transition to university life. This section explores the essential features of the FYE, focusing on their role in academic [62, 81, 82, 85, 94, 97], social [4, 62, 81, 82, 85, 97, 110], and personal development, and their impact on student engagement [81, 82, 110], retention rates, and overall success.

Orientation programs provide incoming students with an overview of campus resources, academic expectations, and campus culture. These initiatives are crucial in helping new students acclimatise to the university environment, understand its policies, and use its facilities effectively [74]. Orientation programs offer new students a holistic view of campus life. They typically include tours of key facilities like libraries, lecture halls, student centres, and recreational areas. This physical orientation is crucial for students to feel comfortable and confident in navigating the campus.

These programs also focus on setting clear academic expectations [85]. This might include workshops or seminars on the university's academic standards [110], the rigours of coursework, and the expectations for classroom participation and independent study. Such sessions are invaluable in bridging the gap between high school and university academic cultures.

Orientation often includes introductions to various campus resources students can use throughout their university journey. This includes academic support services like tutoring centres, writing labs, career counselling, and health services. Knowing about these resources can significantly enhance a student's ability to cope with academic and personal challenges.

An essential aspect of orientation is helping students integrate socially [4, 85]. Activities are designed to foster connections among new students [110] and with faculty and staff [110]. Ice-breaking sessions, group activities, and social events facilitate the formation of new friendships and networks, which are critical for a well-rounded university experience.

Academic advising is integral to the university experience, serving as a foundation for

student success [105]. Advisors are vital in guiding students through the course selection, ensuring that their academic choices align with their personal interests and career aspirations, and fulfilling the necessary graduation requirements. Personalised guidance is crucial, particularly for first-year students often unfamiliar with university course planning [105]. By providing insights into different academic paths and opportunities, advisors enable students to make informed decisions that shape their educational and professional trajectories. They also play a critical role in identifying and addressing academic difficulties early on. They connect students with additional resources such as tutoring or study skills workshops, thereby fostering a proactive approach to academic challenges.

Peer mentoring programs complement the role of academic advising by offering a more relatable and accessible form of support. These programs pair incoming students with experienced senior students, creating a supportive peer-to-peer network [110] that eases the transition into university life. Having navigated the early stages of university, senior students are well-positioned to offer practical advice on managing coursework [85], balancing extracurricular activities, and making the most of campus resources. This form of mentoring provides a unique blend of academic guidance, emotional support [85], and social integration [4, 110], which is particularly valuable for new students who may feel overwhelmed by the sudden shift to a more autonomous learning environment. Furthermore, peer mentoring fosters a sense of belonging and community, which is fundamental to student retention and satisfaction.

Co-curricular activities, encompassing a diverse array of clubs, organisations, and community service projects [97], serve as a dynamic and integral component of the university experience [81, 82, 110]. Participation in these activities offers students a platform to immerse themselves in the university community, fostering a deep sense of belonging and connection. Engaging in such activities allows students to explore and develop personal interests [30], cultivate new skills, and form meaningful relationships outside the confines of the classroom. Involvement in clubs and organisations encourages teamwork, leadership, and organisational skills, all of which are invaluable in personal and professional development. It provides a balanced environment where academic pursuits are complemented by personal growth and community involvement, which is crucial for holistic student development.

Recognising the diverse academic needs of students, particularly those in their first years, universities offer various academic support services, including tutoring programmes, writing assistance, and study groups [62, 81, 82, 85, 94, 97]. These services are designed to address the specific challenges that first-year students often encounter as they acclimate to the demands of higher education. Tutoring services, often provided by students

or faculty members, offer group assistance in various modules, enabling students to grasp complex concepts and enhance their understanding of the coursework. Writing centres play an important role in enhancing students' writing skills, which is critical for success across all disciplines. These centres guide structuring essays, conducting research, and developing arguments, skills that are essential for academic writing but not always covered in standard curricula [110]. Study groups, facilitated by the university, create a collaborative learning environment where students can share ideas, clarify doubts, and learn from each other. Such group settings not only aid in academic learning but also foster a sense of community and mutual support among students, which is vital for maintaining motivation and engagement in the academic journey.

Wellness and counselling services provided by universities are essential components in supporting the holistic well-being of students [4, 85], particularly during their critical first year. These services address various personal and mental health challenges that students may encounter as they transition into a new and often demanding academic and social environment. Counselling centres on campus offer confidential sessions where students can discuss various issues, including stress, anxiety, depression, and relationship problems [33]. These services are not only about mitigating mental health crises but also about providing a supportive space for students to navigate their emotions and challenges[85]. Wellness programs complement counselling services by focusing on broader aspects of health, such as physical fitness, nutrition, and mindfulness. Workshops and activities centred around these areas aim to equip students with the skills and knowledge to maintain a healthy lifestyle intrinsically linked to their academic success and overall quality of life.

The transition to university life can be particularly challenging for underprepared students who enter the academic setting. This lack of preparedness, which can be academic, social, or emotional, may lead to heightened stress and anxiety as students struggle to meet the expectations and demands of higher education [85]. The pressure to catch up academically, coupled with the need to adapt to a new social environment, can be overwhelming, making these students more susceptible to mental health issues. Universities must recognise and proactively address these challenges by providing targeted support to underprepared students. By identifying and supporting underprepared students from the onset, universities can help mitigate the potential escalation of stress and anxiety that might otherwise hinder their academic journey and personal development.

Given the increasing complexity of challenges students face in higher education, there is a growing need for universities to adopt a comprehensive approach to wellness and mental health [45]. This approach should integrate various services and resources to create a supportive and nurturing environment. Strengthening partnerships between academic

departments, wellness centres, and student organisations can facilitate a more cohesive and effective support system. Additionally, universities should endeavour to destigmatise mental health issues and encourage open discussions about well-being [45]. This can be achieved through awareness campaigns, peer-led initiatives, and integrating well-being topics into the curriculum. Ultimately, the goal is to create a university culture that responds to mental health needs when they arise and actively promotes and prioritises the overall well-being of every student.

The timing of interventions is a critical factor in the successful transition and retention of students in higher education. Studies have shown that early interventions, particularly during the first few weeks of the academic term, are important for helping students adapt to university life and can significantly impact their retention and success rates [63, 104]. For instance, Tinto (2006) emphasizes that the first year is a critical period for student retention and that interventions during this time can help mitigate challenges students face as they transition to university [104]. Similarly, Kuh et al. (2005) highlight that timely academic support and engagement opportunities are essential in promoting student success and preventing early departure [63].

Moreover, interventions must be tailored to align with key milestones and potential stress points in the student lifecycle. Research by Yorke and Longden (2008) suggests that timely support during high-stress periods, such as midterms and finals, can help students manage their workload and reduce the likelihood of dropout [112]. This highlights the importance of not only the content but also the timing of interventions in addressing student needs effectively.

The effectiveness of the FYE is contingent upon a comprehensive approach that addresses the academic, social, and personal needs of students. By integrating orientation programs, academic advising, specialised courses, peer mentoring, residential learning communities, engagement opportunities, academic support, and wellness services, universities can create a supportive and enriching environment [4, 85, 110]. This holistic approach is fundamental in facilitating a successful transition into university life, thereby promoting retention and academic success.

Academic institutions have implemented interventions such as academic development and supportive initiatives for students [30]. While these interventions have assisted many students, the progress observed through these measures has been limited and produces only moderate results [31]. This highlights the need for more targeted and personalised strategies considering each student's unique circumstances and needs. By doing so, universities can better support students on their academic journey and enhance their chances of success.

This research is important as it not only builds upon the existing body of knowledge but also seeks to provide actionable insights that can be used to improve student outcomes. By developing a predictive model, universities can proactively identify potential challenges and implement appropriate interventions, enhancing overall student experience and academic performance.

This study represents a significant step forward in understanding student success and dropout prevention. By integrating the findings of previous descriptive analyses into a predictive model, universities can more effectively address gaps in student expectations and experiences, ultimately contributing to improved academic performance.

3.3 Data Mining Techniques

Various techniques can be used to predict student academic performance; selecting an appropriate technique mainly depends on whether the variable in question is numerical or categorical. Among the commonly used methods are classification, regression, and clustering. Many researchers have preferred classification techniques in predicting student academic performance, encompassing a range of algorithms including decision trees, naive Bayes, logistic regression, k-nearest neighbour, neural networks, and support vector machines.

Each of these algorithms has distinct characteristics and offers specific advantages. Decision trees are highly regarded for their interpretability, which allows researchers to gain insight into the decision-making process. It has the flexibility to handle both numerical and categorical data, making them suitable for various types of predictors [15, 42, 73, 113]. On the other hand, Naive Bayes is valued for its simplicity and efficiency. It assumes independence among predictors, allowing quick computation and making it particularly effective in high-dimensional data sets [29, 42, 93, 113]. Logistic regression, a widely used algorithm, is well suited for binary classification tasks, providing probabilistic outputs and the ability to model the relationship between predictors and the likelihood of an event occurring [15, 30, 61, 73].

The K-nearest neighbour is a nonparametric method that determines the class membership of an instance based on its proximity to the nearest neighbours in the feature space. This algorithm is advantageous when the decision boundary is irregular or the data exhibits complex relationships. Neural networks, known for their ability to learn intricate patterns and relationships in data, have gained popularity in recent years [20, 56, 106]. They consist of interconnected layers of nodes (neurons) that mimic the structure and functioning of the human brain. Neural networks excel at capturing nonlinear relation-

ships and can handle high-dimensional data effectively. On the other hand, support vector machines are robust algorithms that can handle linear and non-linear data. They constructed an optimal hyperplane that maximally separated the different classes, allowing for accurate classification.

In the context of this study, a comprehensive exploration of each of these algorithms will be conducted. These techniques' theoretical foundations and practical applications will be examined, providing a comprehensive understanding of their underlying principles and potential use cases. Additionally, related research that used these techniques in predicting student academic performance will be reviewed. This comprehensive overview will contribute to the existing knowledge base, shedding light on the current landscape of predictive modelling in the context of student academic performance.

When predicting student academic performance, researchers can choose from various techniques based on the nature of the variable predicted. Classification techniques, such as decision trees, naive Bayes, logistic regression, k-nearest neighbour, neural networks, and support vector machines, gained prominence due to their versatility and effectiveness. Each algorithm offers unique advantages and characteristics, making them suitable for different scenarios. This study explores these algorithms, examining their theoretical foundations, practical applications, and previous use in predicting student academic performance. By comprehensively exploring these techniques, this study seeks to advance knowledge in the field and inform the development of an effective predictive model for student academic performance. Table 3.2 below shows the techniques used to predict an outcome.

3.4 Summary

This chapter provided a comprehensive overview of prior research focused on applying data mining techniques to predict student academic performance, particularly in the context of the first-year experience. The review synthesises the findings of previous studies, identifies gaps in existing knowledge, and lays the groundwork for the present research. The chapter covered three main sections: understanding the characteristics of student performance, the characteristics of the first-year experience, and exploring data mining techniques.

The first section demonstrated and highlighted the limitations of previous research because it relied mainly on demographic data and individual grades to predict student success. This study seeks to shift the focus to transitioning characteristics, specifically social, academic, and wellness, as they might be important determinants of student performance. By considering a broader range of factors, this study seeks to develop a more

Table 3.2: List of Data Mining Techniques Used to Predict an Outcome

Techniques	Prediction objective	Previous Studies
Logistic Regression	Academic performance	[30, 61]
	Attrition	[15, 73]
Neural Network	Academic performance	[20, 56, 106]
	Attrition	[15, 73]
Decision Tree	Student retention	[30]
	Attrition	[15, 42, 73, 113]
	Academic performance	[1, 29, 56, 60, 61, 106]
Naive Bayes	Attrition	[42, 113]
	Academic performance	[29, 93]
Support Vector Machine	Academic performance	[29]
	Attrition	[42, 113]

comprehensive predictive model that accurately reflects the multifaceted nature of student experiences.

The second section moved into the features of the first-year experience that significantly impact student success. These characteristics include a sense of capability, connectedness, resourcefulness, purposefulness, and the institution's academic culture. Drawing on information from previous studies on the first-year experience, this research aims to identify targeted interventions that bridge the gap between student expectations and actual experiences, ultimately improving academic results and retention rates by using predictive modelling capabilities.

Lastly, the chapter discussed various data mining techniques, such as classification and regression that can be used to predict student academic performance. Each technique offers unique advantages, and their theoretical foundations and practical applications were explored. By examining related work that used these techniques, this study seeks to improve the understanding of predictive modelling in the context of student performance and inform the development of an effective predictive model.

Chapter 4

Methodology

The chapter outlines the methodological approach used to achieve the aims and objectives of the study. The study uses the CRISP-DM methodology, which provides a structured plan for data analysis and modelling [24]. This methodology consists of six stages 4.1 business understanding, 4.2 data understanding, 4.3 data preparation, 4.4 model development, 4.5 model evaluation, and 4.6 model deployment and recommendations. The final section will discuss the ethical considerations 4.7 relevant to the study.

4.1 Business Understanding

Figure 4.1: Business Understanding

The data mining steps will be highlighted as the study progress through the sections such as in figure 4.1. According to Latief [67], there has been a notable transformation in the student population at South African universities, with an increase in enrollment of individuals from diverse socio-economic backgrounds. This expansion of access to higher education institutions in South Africa poses a complex challenge: accurately assessing incoming students' academic readiness and unique requirements, especially during their critical first year of study. Additionally, students' level of engagement and active involvement in their university experience can be significantly influenced by their expectations and perceptions [83]. These anticipations may even be crucial in determining whether students persist or discontinue their studies [103].

Although the importance of the first-year experience is widely acknowledged, our understanding of its impact on academic performance is limited. This necessitates further research on how diverse student bodies can measurably improve this experience. Notably, understanding a diverse student population can enrich their academic and social experience but concurrently introduces outlooks and anticipations that can create target holistic support to ensure successful outcomes for all students involved [63].

Research on student experience has uncovered that many first-year students start their academic year with unrealistic expectations of their university journey [83]. These assumptions often include social and academic integration [81, 82, 83], and the development of generic academic skills [8]. These can lead to a disconnect between expectations and reality, potentially contributing to student dissatisfaction and attrition.

This study aims to examine the influence of expectations and experiences of first-year students on academic performance. The primary objective is to explore the profiles of students, their pre-entry information, academic and social expectations, and overall university experience. The dataset was drawn from the university's administrative systems, including information about student academic performance and the online survey results of student expectations and experiences. Further details about the data set will be elaborated upon in the Data Understanding Section.

4.2 Data Understanding

Figure 4.2: Data Understanding

The data used in this study consisted of the first-year student population registered at the University of the Western Cape, South Africa. Specifically, the study population consisted of 4 500 students starting their first degree and registering for the first time at the university. The sample for this study was selected using purposeful sampling methods. The sample consisted of students who participated in the First Year Experience and Expectation survey conducted in 2019.

Purposeful sampling is a non-random sampling technique where participants are selected based on specific characteristics or criteria that align with the research objectives [80]. In this case, the study seeks to examine the experiences and expectations of

first-year students. Thus, only those students who completed the survey were purposively sampled.

Selecting students who had already completed the First Year Experience and Expectation survey ensures that the sample would be representative of the population of interest - first-year students willing to participate in such surveys. However, it is important to note that purposeful sampling may introduce potential bias into the sample, as it relies on researchers' judgment and subjective decision-making [80]. Overall, purposeful sampling allowed for a targeted sample for the study while ensuring relevance to research questions.

Table 4.1: Population and Sample

	Count or Percentage
Population	4 500
Sample	2 054
Proportion	41%

The data for this study were obtained from multiple database sources within the university's business intelligence data warehouse. This includes academic records, administrative records, and responses to student surveys. The survey records are based on a pilot study conducted in 2018 by Pather and Booi [82], which was subsequently adopted and formalised as part of the university registration process. The formalisation of this survey was required by the limitations of administrative systems, which do not capture comprehensive information about students' pre-university attributes, particularly their expectations and experiences. The datasets were anonymised with a unique identifier to facilitate the integration or joining of the multiple datasets.

Table 4.2: Description of Student Administration Dataset

Field	**Attributes**	**Measurement Level**
ID	Unique identifier	Interval
Gender	Gender	Nominal
FirstGen	First Generation	Nominal
Bursary	Bursary	Nominal
Residence	Residence	Nominal
PopGroup	Population Group	Interval
Programme	Qualifications	Nominal
APS	Admission point score	Interval
GPA	First Year Grade Point Average	Interval

Table 4.2 shows a detailed list of attributes of the Student Administration Dataset.

The student academic and administrative data files included pre-university entry information, such as school quintile and admission point score, and university academic information, such as average academic performance during the first year and details about the type of programme and credit load.

The dataset on student expectations and experiences consisted of 35 statements to capture various aspects of students' perceptions. These statements were responded to using a five-point Likert scale, a commonly used tool for collecting subjective data such as personal attitudes and opinions [6]. The Likert scale allows respondents to indicate their level of agreement or disagreement with each statement on a scale of 1 to 5. A score of 5 indicates strong agreement with the statement, while a score of 1 signifies strong disagreement. This range of options provides a nuanced understanding of the students' views by capturing the strength of their agreement or disagreement with each statement. A comprehensive list of all student expectations and experience statements is provided in Table A.1 in Appendix A [82].

4.3 Data Preparation

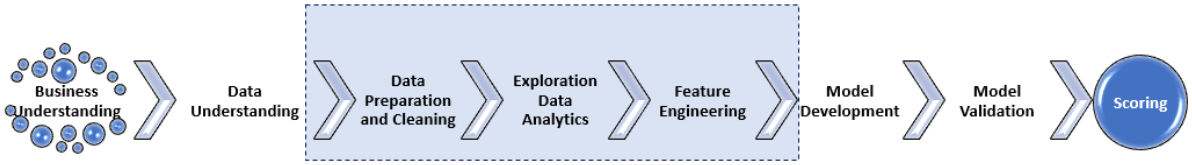


Figure 4.3: Data Preparation

Merging data sets to create a single input file for analysis is essential in the data understanding phase. Combining multiple data sets allows for a more comprehensive and holistic view of available information. According to Kimball [57], data integration or merging is critical to the data process. This consolidation enables uncovering patterns, relationships, and insights that may not have been apparent when examining individual datasets separately. Merging data sets allows for more robust analysis by incorporating a broader range of variables and observations.

After combining the data sets, an in-depth analysis or profiling of the data was undertaken. This process involved reviewing each variable to ensure consistency and completeness in confirming missing values. Data missing values could skew the results and lead to inaccurate conclusions during the analysis. Data profiling is crucial to maintaining data

integrity and ensuring its readiness for analysis [109].

When problems were identified during the data profiling stage, appropriate transformation methods were applied to each variable to rectify these problems. Data transformation involves various techniques, from simple procedures like filling in missing values to more complex operations like normalisation or standardisation. These transformations aimed to correct inconsistencies or errors in the data, ensuring they are in the best possible form for analysis [79].

Data transformation helped to manage outliers, handle skewness, and improve the performance of data mining models. For example, applying the normalisation of the factors, a method that adjusts the values in the dataset to a common scale, was applied particularly to benefit the models because of convergence issues since the datasets had large variations in values [109]. It is important to emphasise that the unique characteristics of the data dictated the selection of a data transformation method.

The data analysis was performed using R Studio [87], a powerful and widely used open-source software application for statistical analysis and data visualisation. R Studio offers a range of tools and functionalities that make it an ideal choice for handling complex datasets and performing sophisticated analyses.

4.3.1 Data Cleaning

The dataset underwent a meticulous cleaning process to ensure the accuracy and relevance of the data for analysis. This process was guided by a set of predefined criteria, which are explained below.

The Admission Points Score (APS) is a numerical value that reflects a student's high school aggregated score of their school results. Entries with an APS of less than 20 were deemed inappropriate. The rationale behind this decision was to focus on data representing students with a higher level of academic school achievement, as well as the fact that the minimum entrance APS score at the university is 27. The other reason was that there were no APS score records between 20 and 27. Therefore, all entries with an APS of less than 20 were excluded from the dataset.

The programme was characterised by an extensive range of categories, including programme specialisation. However, this level of detail was deemed excessive for the analysis. Therefore, the programme categories were grouped into more manageable broader categories, such as Mainstream Programmes and extended Programmes. This recording process facilitated a more streamlined and efficient analysis.

The final criterion for the data cleaning process was related to the completeness of the average first-year marks of the students. These marks provide valuable information about

a student's academic performance during their first year at the university. However, some students in the data set did not have a complete record of these marks. Incomplete data could potentially skew the analysis results and compromise its validity. Therefore, to maintain the integrity and consistency of the analysis, students with incomplete first-year average marks were excluded from the study. After processing and cleaning the data, the sample used for the analysis consisted of 1 764 records. Strict criteria are necessary to ensure that the data collected are relevant and reliable. This provided a solid foundation for the following stages of the investigation.

4.3.2 Data Imputation

Addressing missing data is a common challenge in data analysis. This study used the Multiple Imputation by Chained Equations (MICE) methods to handle missing values in the expectation and experience Likert statements [5]. MICE generates multiple predictions for each missing value by modelling each feature with missing data as a function of other features, rendering reasonable values that maintain the statistical properties of the data.

For each Likert statement in the expectations dataset, the former assumes that the observed data adheres to a multivariate normal distribution, and the algorithm used by R packages, such as Amelia, to impute missing values draws values from this assumed distribution [49]. However, the imputed values may be incorrect if the data does not follow a multivariate normal distribution.

On the other hand, when the variable does not follow the abovementioned criteria, a Conditional Multiple Imputation was used. It employs an iterative procedure, modelling the conditional distribution of a specific variable given the other variables [5]. This technique allows for greater flexibility as a distribution is assumed for each variable rather than the entire dataset. This approach is particularly useful when dealing with Likert scale data that often have a non-normal distribution.

Figure 4.4 illustrates that most of the 'experiences' variables and a few 'expectations' variables exhibited missing data. These missing data points tend to occur in blocks of a few observations. Contrarily, missing data points were predominantly found towards the end of each cross-section. Identifying these patterns was facilitated by using a missingness map, a crucial tool for understanding the patterns of missingness in the data. Missing values are in white and observed values are in blue.

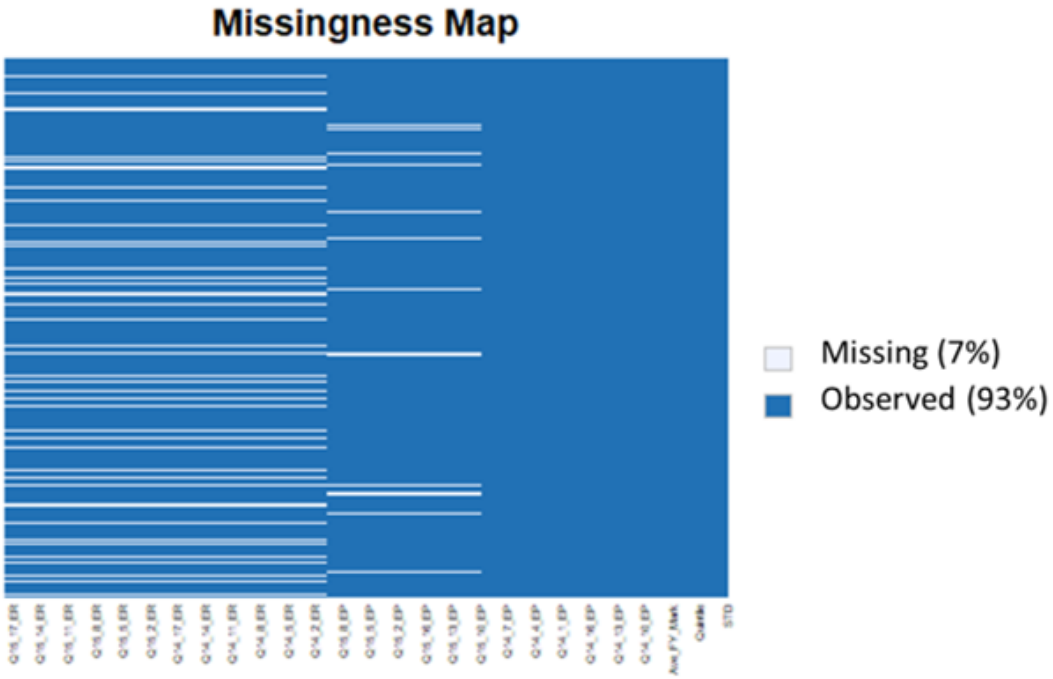


Figure 4.4: Missingness map of the Dataset

Given the non-symmetrical and self-reported nature of the data, median imputation was initially employed for the individual Likert statements within the 'experiences' dataset. This approach was chosen because it is less sensitive to outliers and skewed distributions, offering a robust estimate of the central tendency [70]. However, the use of median imputation has limitations, such as underestimating the standard error, which needs to be considered in the context of the analysis [49].

For more complex missing data patterns, particularly those involving entire responses missing for the "ER" variables, conditional multiple imputation using the MICE package was applied. This method allows for a more nuanced imputation by considering the conditional distributions of other variables, making it suitable for handling missing data in Likert scales [21]. However, it is crucial to assess whether the assumptions of this method are met and to fully understand its limitations. Imputing entire responses for the "ER" variables is a complex decision. On the one hand, imputing these responses can preserve the completeness of the data set, which is beneficial for maintaining statistical power and enabling comprehensive analyses. However, there are significant considerations:

Imputing entire responses can introduce bias if the missingness is not at random (MNAR) and is related to unobserved variables. This could distort the relationships within the data and lead to inaccurate conclusions [70]. The assumptions underlying the chosen imputation method, such as MAR (Missing at Random) or MCAR (Missing Completely at Random), need careful evaluation to ensure that they align with the dataset's

characteristics.

If the missing data for the "ER" variables follows a specific pattern, such as being predominantly at the end of the cross sections, it might indicate a systematic issue rather than random occurrences. In such cases, simply imputing the data may not be appropriate, and other strategies, such as investigating the cause of missingness, might be warranted [70].

If imputation is deemed inappropriate, other methods such as analysing the missingness pattern itself or employing model-based approaches that can handle missing data implicitly, such as mixed models or Bayesian methods, could be considered [22].

Given the use of median imputation to handle missing 'ER' variables, this approach has effectively preserved the integrity of the data set for missing isolated responses. However, the limitation of the current study is the handling of missing data for the "ER" variables. Although median imputation was applied effectively for individual missing responses, this approach may not fully address the issue of entire missing responses for certain individuals. This could introduce bias and affect the reliability of the findings. Future research should consider conducting a sensitivity analysis to understand the impact of these imputations on the results. Additionally, exploring advanced imputation techniques such as multiple imputation or model-based approaches could provide a more robust solution to deal with extensive missing data. Addressing these limitations will enhance the accuracy and credibility of future studies in this area.

4.3.3 Data Selection

The selection of appropriate data is crucial for ensuring the reliability and meaningfulness of results obtained from the data mining process [56]. This section provides an in-depth understanding of data selection and emphasises its significance in maintaining the integrity and applicability of findings. Data selection is a fundamental step within the pre-processing phase of data mining. It involves identifying and extracting relevant subsets of data from a larger pool for further analysis. The objective is to filter out noise, irrelevant variables, or redundant information that may hinder accurate predictions or increase computational costs.

By eliminating unnecessary variables or instances that do not contribute significantly to the desired outcomes, the algorithm can focus on processing a reduced dataset with higher relevancy. This targeted approach saves computational resources and time required for analysing large datasets. Moreover, data selection plays a pivotal role in improving prediction accuracy. By carefully selecting relevant variables or features with strong predictive relationships with the target variable, models can be trained more effectively [56].

Removing noisy or irrelevant attributes minimises the chances of overfitting and reduces model complexity, thus leading to better generalisation and more reliable predictions.

4.3.3.1 Cronbach alpha coefficient

The reliability of the Student Expectation and Experience scale, which encompasses Question 1 through Question 35, was assessed using the Cronbach alpha coefficient. This statistical measure was evaluated based on established guidelines, categorising a coefficient greater than 0.80 as excellent, above 0.60 as good, above 0.40 as moderate, above 0.20 as fair, below 0.20 as poor [40].

Table 4.3: Reliability Analysis for Student Expectation and Experience

Description	**Number of Items**	**Cronbach's Alpha**
Student Expectations	35	0.86
Student Experiences	35	0.87
Expectations & Experiences Gap	35	0.87

The items for Student Expectation had a Cronbach's alpha coefficient of 0.86, indicating excellent reliability. The following variables were negatively correlated with the overall composite score: Q12 and Q26. These variables were automatically reverse-coded to improve reliability.

The reliability of the Student Experience scale was 0.87, which indicates excellent reliability, suggesting that the items consistently measure the intended construct. However, a specific item, Question 18, negatively correlated with the overall composite score. To enhance the reliability of the scale, this item was automatically reverse-coded.

4.3.3.2 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) was an essential component of the data selection process for 35 statements on expectation and 35 on experience to identify underlying factors or dimensions within the dataset. The primary goal was to explore whether expectation and experience statements could be grouped into distinct factors or constructs. This approach enabled the study to uncover any underlying themes or commonalities within these statements, which would help better understand the phenomenon under investigation [79].

Furthermore, EFA allowed for the identification of redundant or overlapping items within each set of statements. Through this process it could eliminate items that did not contribute significantly to capturing the variability in expectations and experiences, thereby refining our measurement instrument. By employing EFA on the dataset com-

prising 35 expectation statements and 35 experience statements, the study can gain a deeper understanding of the underlying dimensions present within these sets.

The Kaiser criterion, a widely accepted statistical method, was used to determine the number of factors to be retained. This criterion is based on the eigenvalues of the factors, where factors with eigenvalues greater than one are retained. To enhance the interpretability of the factors, a varimax rotation was applied. This orthogonal rotation method simplifies factor loadings, making understanding the relationship between variables and factors easier [55].

The factor extraction method used in this study was factoring the principal axes. This method is a common choice in EFA as it does not assume multivariate normality, unlike other extraction methods. Therefore, the assumption of multivariate normality, which assumes that all variables in the analysis are normally distributed, does not apply in this context. The normality scatterplot in Figure 4.5 shows that the data deviate from normality. While most points are relatively close to the identity line, indicating some alignment with a normal distribution, there are noticeable deviations, particularly with outliers at the higher end, which suggests that the data do not strictly follow a normal distribution.

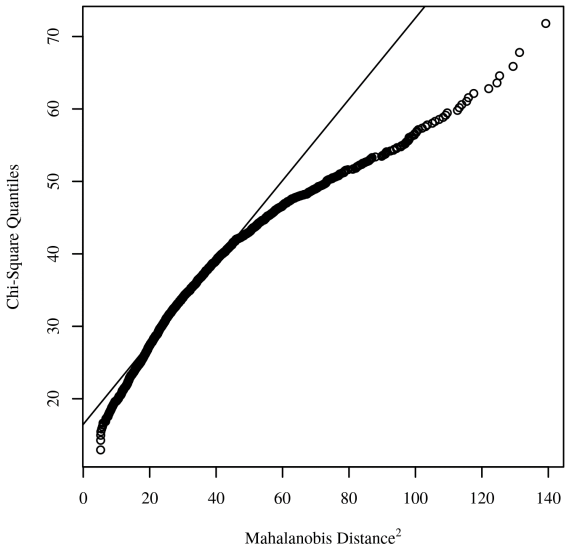


Figure 4.5: Scatterplot testing multivariate normality

The suitability of the data for factor analysis, also known as factorability, was assessed by calculating Pearson's correlation coefficients. These coefficients provide a measure of the linear relationship between pairs of variables. According to Tabachnick and Fidell [98], a correlation coefficient should exceed 0.30 to justify the factorability of the data.

All variables met this criterion in this analysis, with at least one correlation coefficient greater than 0.30, suggesting that the data were suitable for factor analysis.

Although variables must be intercorrelated to some degree for factor analysis, excessive correlation or multicollinearity can cause problems. Multicollinearity can inflate the variance of factor loadings, making them unstable and difficult to interpret. The determinant of the correlation matrix was calculated to assess the presence of multicollinearity. Field [36] suggests that a determinant value ≤ 0.00001 indicates multicollinearity. However, in this analysis, the determinant value was 0.00002, suggesting that multicollinearity was not a concern in this dataset.

In determining the optimal number of components, a criterion where components with eigenvalues over 1 were selected and applied to the first four components.

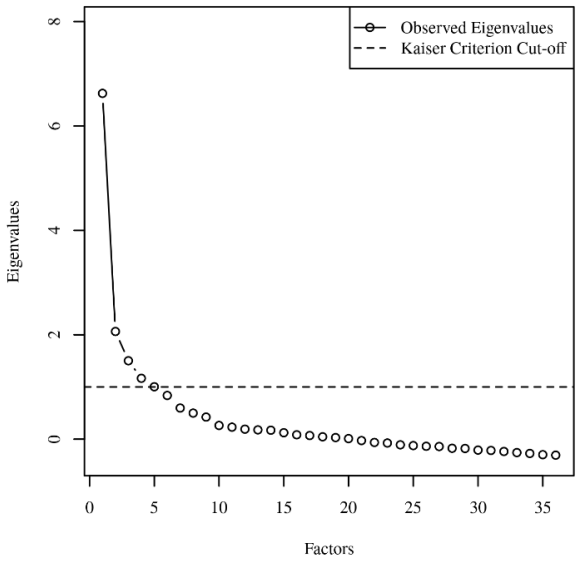


Figure 4.6: Scree Plot

The Scree Plot figure 4.6 affirmed this selection, displaying an "elbow" at the fourth component, suggesting no further extraction is necessary. The identified factors are Factor 1 - Effective Learning, Factor 2 - Social Wellbeing; Factor 3 - Academic Support; and Factor 4 - Access to Information. Refer to table 4.4 and Factor 1 - Effective Learning is created from questions focusing on the direct facilitation of learning at a university. It includes the availability of lecturers for extra assistance, regular academic feedback, access to learning resources like the internet and computers, and clear campus navigation. These elements are important for a conducive academic environment, directly impacting a student's capacity to learn effectively. Questions 1, 2, 20, and 35 were omitted following extraction due to their insufficient correlation with the central factors. Questions 1 and 2 had low factor loadings, indicating a negligible shared variance with other vari-

ables. Questions 20 and 35 were removed due to their negative questioning statement, making them outliers or significant cross-loadings, which can obscure the clarity of factor interpretation.

Table 4.4: Factor Selection

Factor	Description	Questions
FA1	Effective Learning	Q22, Q23, Q24, Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q33
FA2	Social Wellbeing	Q4, Q5, Q6, Q7, Q32
FA3	Academic Support	Q8, Q9, Q11, Q12, Q13, Q14, Q15
FA4	Access to Information	Q3, Q10, Q16, Q17, Q18, Q19, Q21, Q34

Cronbach Alpha. A Cronbach alpha coefficient was calculated for the student Expectation and Experience scale by factors, consisting of Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18, Q19, Q21, Q22, Q23, Q24, Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q32, Q33, and Q34. Table 4.5 shows the reliability analysis of four factors, with Cronbach's alpha (α) ranging from 0.60 to 0.79, indicating good reliability for Effective Learning, Social Wellbeing, and Academic Support, and moderate reliability for Access to Information.

Table 4.5: Reliability Analysis of Factors

Description	Items	α	Results
All Expectation Factors			
FA1 - Effective Learning	11	0.78	Good reliability
FA2 - Social Wellbeing	5	0.72	Good reliability
FA3 - Academic Support	7	0.67	Good reliability
FA4 - Access to Information	8	0.70	Good reliability
All Experiences Factors			
FA1 - Effective Learning	11	0.79	Good reliability
FA2 - Social Wellbeing	5	0.75	Good reliability
FA3 - Academic Support	7	0.72	Good reliability
FA4 - Access to Information	8	0.60	Moderate reliability

4.3.4 Data Coding

The dataset variables were coded using binary values of 0 and 1 for categorical variables, while numerical variables (APS and all factors) were left unchanged. The importance

of accurately coding dataset variables cannot be overstated, as this process improves data interpretability and compatibility with several statistical algorithms and models [47]. Table 4.6 summarises the binary variables where '1' denotes the presence of a condition such as 'MA' for Performance, 'PASS' for Outcome, being Female, having a Residence or Bursary, being First Generation, identifying as African or Coloured, and being in a Mainstream program, while '0' represents their absence or alternative options.

Table 4.6: Variables Coding

Variables	**Value - Binary Value**
Performance	MA - 1 ; MB - 0
Outcome	PASS - 1 ; FAIL - 0
Female	FEMALE - 1 ; MALE - 0
Residence	YES - 1 ; NO - 0
Bursary	YES - 1 ; NO - 0
FirstGen	YES - 1 ; NO - 0
African	AFRICAN or COLOURED - 1 ; OTHERS - 0
Mainstream	MAINSTREAM - 1 ; EXTENDED - 0

In the context of the study, the variable 'African' is a combination of African and Coloured students to account for the shared socio-cultural and educational experiences prevalent in the South African context, which are often similar for these groups, thus providing a more cohesive and contextually relevant analysis of the data [19].

Assigning binary values to code categorical variables yields multiple advantages, such as facilitating a clear distinction between various categories, an attribute that greatly eases the interpretation of research results [47]. Using 0 and 1 values for disparate categories simplifies understanding each category's influence on the outcome variable within the model development process.

4.3.5 Data Profiling

To summarise, present, and organise the data, this study used the method of descriptive statistics. This approach involves the application of numerical calculations, supplemented by graphs and tables, to thoroughly comprehend the data to be presented [2].

For numerical data, measures of a central location, including mean and median, were used to summarise. Furthermore, measures of variation, such as standard deviation and skewness, were used to describe the distribution and variability of the data [2]. These measures are useful in clearly presenting the data's overall characteristics and facilitating

a more accurate analysis.

In contrast, summarisation was achieved by creating tables and charts for categorical data. These visual representations offer an intuitive and easily understandable method of data presentation, highlighting the differences and similarities across various categories and assisting in identifying patterns [2]. Through descriptive statistics, data profiling enabled a broad understanding of the dataset, ensuring the process was well-informed about the data's characteristics before any advanced analysis could be undertaken.

4.3.6 Data Partition

Data partitioning is a crucial step typically undertaken before constructing a predictive model. This process involves dividing the dataset into training and validation subsets, serving two primary purposes: preventing overfitting of the model and facilitating model comparison. By partitioning the data, biases are mitigated, and the accuracy and reliability of the model's outcomes can be assessed, enabling better decision-making and increased confidence in the model's precision [71].

Various compositions can be employed to determine the appropriate partition ratio, such as a 60:40, 80:20, or 70:30 split [41]. The data was partitioned into 70% training data and 30% test data; the training data set is used to train the model [71]. The test data set was used to measure the accuracy and prevent overfitting to enable the selection of the best-performing model [41].

A comprehensive approach was implemented that integrates the 70:30 split with 10-fold cross-validation to address these potential drawbacks and improve the robustness of the evaluation. Initially, the data were divided into 70% for training and 30% for testing. This approach provided a basic, yet direct, performance assessment of the test data. To further refine the model evaluation and optimise hyperparameters, 10-fold cross-validation was incorporated during the training phase. This process involved partitioning the training data into 10 equal folds. Each fold was then used as a validation set once, while the remaining nine folds were used for training, cycling through all folds [86]. Such an iterative procedure is instrumental in tuning the hyperparameters and selecting the most effective model configuration. It offers a robust mechanism to evaluate the performance of the model across different subsets of data, thus reducing overfitting and enhancing generalisation capabilities [54].

After the optimal model was determined through cross-validation, its performance was ultimately assessed using the reserved 30% test set. This two-pronged strategy, combining the train split with cross-validation, leverages the strengths of both methodologies. Specifically, the train split offers a straightforward and clear performance measure on new

data, while cross-validation provides a thorough and reliable tuning process, ensuring that the model is finely tuned and not overfitted to the training data. By integrating these methods, the study ensured that the models were not only well optimised, but also capable of making accurate predictions on previously unseen data, thus improving the rigour and robustness of the research findings [47].

Furthermore, this comprehensive validation approach effectively prevented overfitting, allowed the comparison of different models, and facilitated the selection of the most suitable model to make accurate predictions [41]. Data partitioning is indeed a critical step in predictive modelling, as it enables the optimisation of model performance and significantly enhances the reliability of the model's outcomes. Recognising the importance of balanced data in preventing bias towards any specific class during model training and evaluation, the study employed techniques to achieve a balanced dataset. Balanced data refers to scenarios where each class or category within the dataset has an equal number of instances or observations [44]. To achieve this, the study used random oversampling [25] or undersampling [44], depending on the initial distribution of classes in our dataset. These methods helped create a representative subset with equal proportions for each class, ensuring that the models learnt from a balanced dataset.

On the contrary, unbalanced data occurs when there are significant disparities between the number of instances in different classes [44]. This imbalance often poses challenges in classification tasks due to the model's tendency to be biased towards the majority classes. Therefore, special attention was given to the handling and evaluation of unbalanced data sets in this study. Ensuring appropriate data handling was crucial for maintaining the integrity and reliability of the model's predictions, particularly in contexts where class distributions were initially skewed. By addressing these aspects comprehensively, the study not only optimised model performance, but also protected against biases and inaccuracies that could arise from imbalanced data, ultimately contributing to more robust and reliable research findings.

4.4 Model Development

Figure 4.7: Model Development

Multiple models were employed to predict students' academic performance, aiming to identify and select the most effective model. This study used classification models because the variable of interest, student academic performance or outcome, were treated as a binary outcome variable.

The `caret` package (*short for Classification And REgression Training*) in R provides a comprehensive suite of tools that streamline the creation and evaluation of predictive models [64]. This package offers functionality for data splitting, pre-processing, feature selection, model tuning using resampling, and variable importance estimation, among others. Its design addresses the challenge of diverse syntax and interfaces across different modelling functions in R by offering a standardised and uniform approach. By consolidating these functionalities, `caret` facilitates efficient and reproducible workflows, especially in complex data science tasks [64]. This capability makes it an ideal choice for this study, which requires robust model training and evaluation. `Caret`'s extensive support for a wide array of algorithms and its ability to manage the entire modelling process, from initial data preparation to final model validation, significantly enhances its utility. Consequently, `caret` was chosen as the best package for this research due to its versatility, ease of use, and comprehensive nature.

In this study, cross-validation was used as a robust method to evaluate the performance of the classification models and mitigate overfitting. Specifically, 10-fold cross-validation was utilised for all models. This method involves partitioning the data into ten equal subsets or "folds." In each iteration, nine folds are used to train the model, while the remaining fold is reserved for testing. This process is repeated ten times, and each fold serves as the test set exactly once. The model's performance is then averaged over the ten iterations to provide a comprehensive evaluation metric. This approach ensures that every data point is used for both training and validation, enhancing the reliability of the model's performance assessment [47]. Cross-validation, especially 10-fold, is widely recognised for its balance between computational efficiency and robust model evaluation [59].

The `caret` package in R was integral to this study, providing a streamlined interface for model training and evaluation. One of the key features of `caret` is its built-in support for hyperparameter tuning. For each classification model, `caret` automates the process of selecting the optimal hyperparameters, which are crucial for model performance. The package performs a grid search over a specified range of hyperparameter values and evaluates each combination using cross-validation. By default, `caret` sets a grid size of three for each hyperparameter, providing a balance between exhaustive search and computational feasibility [64]. The use of default values facilitates rapid prototyping and provides

a reasonable starting point for model tuning. This systematic tuning process is important for optimising model performance and generalisability, as it prevents overfitting and enhances predictive accuracy.

The following models were implemented:

- A Logistic Regression (LR) using the Generalized Linear Model (GLM) was used to model the classification task using 10-fold cross-validation to ensure robust evaluation and prevent overfitting. Hyperparameter tuning was conducted using `caret`'s default settings. This approach ensures balanced model performance and computational efficiency.

- the K-Nearest Neighbours (KNN) algorithm was employed for classification tasks. We used the `caret` package in R, applying 10-fold cross-validation to assess model performance and ensure robustness. This method provided a balanced evaluation of the model, ensuring each data point contributed to both training and validation processes.

- A Gradient Boosting Model (GBM) was implemented to sequentially build an ensemble of weak decision trees. Hyperparameters, including learning rate, maximum depth, and number of iterations, were optimised through cross-validation.

- A Random Forest (RF) with an ensemble method combining multiple decision trees was utilised, where each tree is constructed on a random subset of features and observations. The optimal number of trees and other hyperparameters were determined through cross-validation.

- A Decision Tree (DT) algorithm was developed using either an entropy-based approach or a classification and regression tree (e.g., CART). The optimal hyperparameters for tree pruning were determined via cross-validation on the training dataset.

- Support Vector Machine (SVM) with non-linear kernels such as radial basis function (RBF) was employed to develop classification models. Tuning parameters, including kernel type and regularisation parameters, were optimised through grid search and cross-validation.

- The Naive Bayes (NB) algorithm was applied, assuming independence between variables given the target variable. Different variations, such as Gaussian Naive Bayes or Multinomial Naive Bayes, were explored based on the nature of input data.

- Regularised Discriminant Analysis (RDA) models were developed assuming Gaussian distributions for both classes. Regularisation parameters were chosen via cross-validation to improve generalisability.

Various classification models are employed to accurately predict student academic performance to determine the most effective model. The chosen models were specifically suited to address the binary aspect of the target variable. By training and evaluating these models, the study was able to gauge their effectiveness and identify the model that most accurately predicts student academic performance. Furthermore, by thoroughly documenting the optimisation and cross-validation processes, provides a clear and detailed account of the methods used to select the most appropriate models for the analysis.

4.5 Model Evaluation

Figure 4.8: Model Evaluation

During the model evaluation phase, various measures were used to determine the most suitable model for predicting student academic performance. These measures, including the F1-score, classification matrix, sensitivity, specificity, precision, and accuracy, offer critical insights into each model's effectiveness. They play a pivotal role in guiding the selection of the optimal model. To thoroughly evaluate each model's performance, various relevant metrics such as accuracy, precision, recall, and F1-score were used, chosen based on the specific requirements of the problem at hand. Additionally, to prevent overfitting and accurately estimate the models' ability to generalise, cross-validation methods like k-fold cross-validation were implemented.

The F1-score was the best measure because it combines precision and recall, providing a balanced measure of the model's accuracy. It takes into account both the model's ability to correctly identify positive instances (precision) and its ability to identify all positive instances (recall). The F1-score ranges from 0 to 1, with a higher score indicating better performance.

The classification matrix, also known as the confusion matrix, was another measure used as it provided information on the percentage of correctly predicted instances (accuracy), as well as the number of true positives, true negatives, false positives, and false

negatives. The matrix assisted in assessing the model's ability to classify instances correctly and identify any imbalances or biases in the predictions.

Other measures used were sensitivity (true positive rate), specificity (true negative rate), precision (positive predictive value) and accuracy to evaluate the model's performance. A model with a high F1-score, high accuracy, and a balanced classification matrix was typically preferred as the best model.

4.6 Deployment and Recommendation

After evaluating the results, recommendations are made based on the results. The information guides the institutions on areas of concern or interventions they can create. Since this study is based on first-year students and looks at the early days of their studies, it means the institutions can put programs in place to address the gaps in student expectations and experience as they transition through the university. The process includes the first-year life cycle within the university or its value chain process.

The recommendation will include implementing the results for target-specific interventions that will enhance the student's academic performance while at the university and promote student success. Effective interventions can prevent student dropout by tailoring student experiences to expectations and needs, enhancing access to resources for academic planning and goal setting, and providing pre-emptive academic and psycho-social support to all students.

4.7 Ethical Consideration

The study used data from the University of the Western Cape (UWC), South Africa, adhering to ethical standards in collection and analysis. It used individual students' demographic and academic data, which was strictly used for analysis. Before sharing, this data was anonymised to maintain confidentiality. A key focus of the study was to ensure the integrity of the data, emphasising quality and accuracy in recording, which facilitated anonymity and secure storage with password protection; such measures were crucial in protecting the data.

Central to the study's ethos was safeguarding the student's rights, dignity, safety, and privacy. The guiding principle of 'do no harm' was integral, aiming to ensure that the analytical outcomes benefited the students and the institution. Rigorous steps were taken to rectify inaccuracies, address missing data effectively, and prevent misleading correlations from upholding data validity. Both the collection and interpretation of data

were approached with sensitivity and precision.

Ethical approval was sought and received from the University of South Africa (UNISA) - Appendix B and the University of the Western Cape (UWC) - Appendix C from which the data was sourced.

4.8 Summary

This chapter systematically described the methodological framework of this study. Adopting the CRISP-DM methodology established a well-defined and efficient data analysis and modelling pathway. The CRISP-DM framework covered six pivotal stages, each crucial in the research process. Firstly, the business understanding stage involves comprehending the study's overarching goals and needs, setting a solid foundation for subsequent stages. Secondly, the data understanding phase was critical in gaining insights into the data structure. Thirdly, data preparation, a vital step, elaborated on cleaning and preparing the data for analysis.

The fourth stage, model development, was where predictive models were crafted and refined. Following this, the fifth stage, model evaluation, entailed assessing the models' performance using various metrics. The sixth stage focused on model deployment and the formulation of recommendations, translating the study's findings into actionable insights. Lastly, the chapter concluded with a discussion on ethical considerations, ensuring the study's adherence to ethical standards, practices, and limitations. This comprehensive methodological approach ensured the study's rigour, relevance and applicability in addressing the research objectives.

Chapter 5

Results

This chapter presents the findings and results. As a reminder, this study aimed at investigating and analysing the impact of first-year expectations and experiences on student academic performance. The primary objective of this study was to examine and analyse how the expectations and experiences of first-year students influence their academic performance. Specifically, the study focused on two dependent variables: Academic Performance, measured by a student's First Year Median Grade Point Average (GPA), and Academic Outcome, determined by whether a student passed or failed. The chapter is organised into several sections. The first section provides a data profiling analysis of student academic performance, expectations, and experiences. This analysis aimed to understand the patterns in these variables among first-year students.

Next, predictive models were developed to examine the relationship between student expectations, experience, and academic performance or outcome. These models sought to determine if there was any predictive value in the expectations and experience that first-year students held regarding their academic success. This chapter highlights the importance of understanding the factors to develop effective strategies for supporting students during their transition to higher education.

5.1 Data Profiling

To effectively develop a predictive model for student academic performance, it is crucial to thoroughly explore and profile the available data. This section focuses on data profiling various student attributes related to academic performance, expectations, and experiences. Subsection 5.1.1 Student Demographic and Categorical Academic Attributes explores key factors such as gender, first-generation status, population group, bursary recipients, residence type, and programme types. Additionally, the subsection 5.1.2 Numer-

ical Academic Attributes examines important numerical measures, including Admission Points Score (APS) and Grade Point Average (GPA). The section includes these subsections: student expectation profile, student experience profile, expectation and experience gap, and exploratory factor analysis. By analysing these attributes in detail, the study will gain valuable insights that will aid in developing an accurate predictive model for student academic outcomes.

5.1.1 Student Demographic and Categorical Academic Attributes

The data presents a descriptive statistical analysis of the *nominal data* and academic performance and Outcome. Performance is categorised into two groups: those who scored below the median (MB) and those who scored above the median (MA). Students' academic outcomes are categorised as whether they passed or failed. The total number of students (n) and the percentage (%) of each category are also provided.

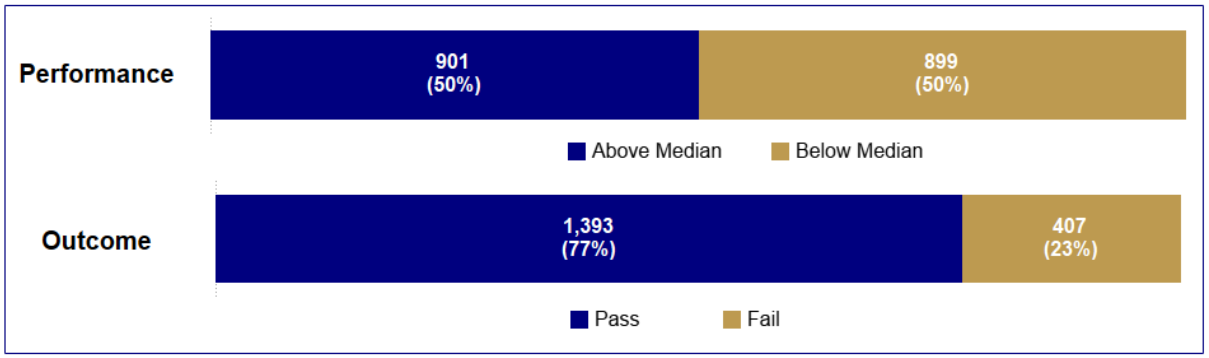


Figure 5.1: Academic Performance and Outcome

Figure 5.1 performance data shows 50% scored above the median, while 50% scored below the median. The outcome data shows that 77% scored passed and 23% failed.

5.1.1.1 Gender

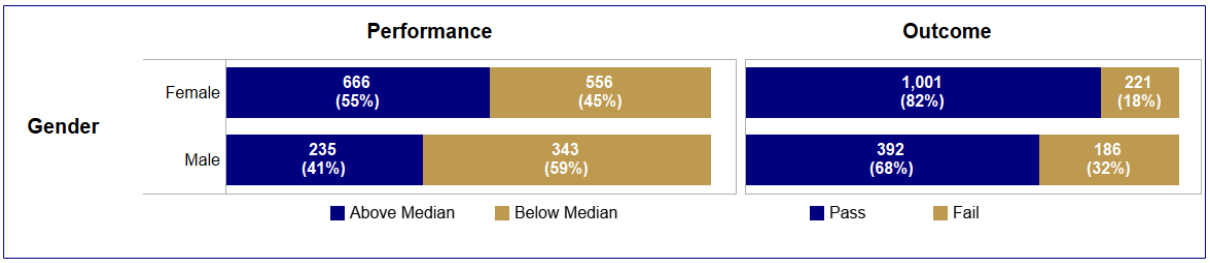


Figure 5.2: Gender by Academic Performance and Outcome

On the other hand, of the 2 222 female students, 45% scored below the median, and 55% scored above the median. The total number of male students was 578, of which 392 (68%) passed and 186 (32%) failed. The total number of female students was higher, at 1 222, with 1 001 (82%) passing and 221 (18%) failing. Figure 5.2 shows that female students have a higher pass rate than male students.

5.1.1.2 First Generation

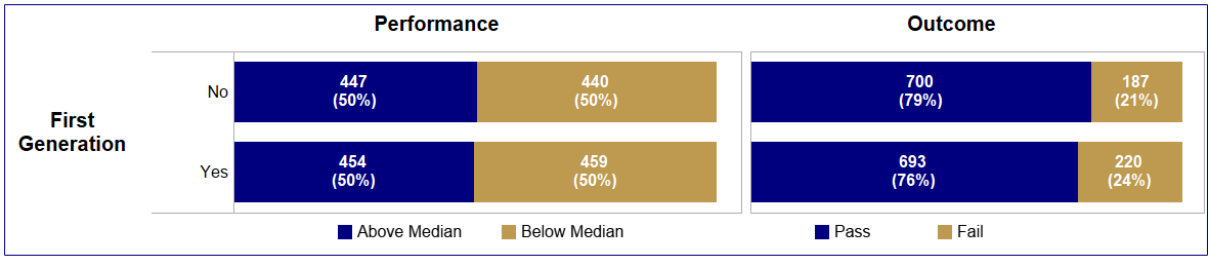


Figure 5.3: First Generation by Academic Performance and Outcome

Figure 5.3 provides information on whether the students are first generation. Of 887 students who are not first-generation, 50% scored below the median, and 50% scored above the median. Among the 913 first-generation students, 50% scored below the median, while 50% scored above the median. 887 students were not first-generation students, with 700 (79%) passing and 187 (21%) failing. Of the 913 first-generation students, 693 (76%) passed and 220 (24%) failed.

5.1.1.3 Population Group

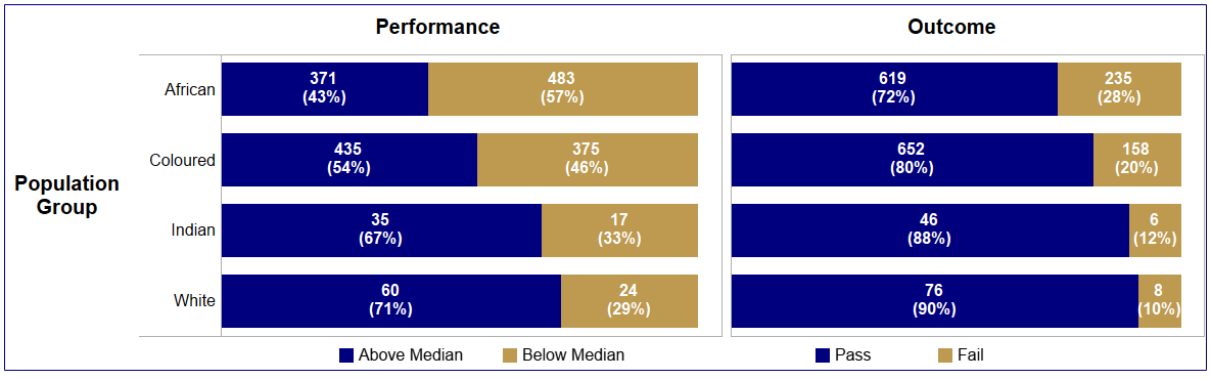


Figure 5.4: Population Group by Academic Performance and Outcome

Figure 5.4 shows the breakdown of performance by population group. Among the 810 Coloured students, 46% scored below the median, and 54% scored above the median.

Of 84 white students, 29% scored below the median, and 71% scored above the median. Among the 854 African students, 57% scored below the median, and 43% scored above the median. Of 52 Indian students, 33% scored below the median, and 67% scored above the median. Data were also analysed by ethnicity. Of the 810 Coloured students, 652 (80%) passed and 158 (20%) failed. Among the 84 White students, 76 (90%) passed and 8 (10%) failed. The African student group was larger, with 854 students, of which 619 (72%) passed and 235 (28%) failed. The smallest group was Indian students, with 52 students, 46 (88%) of whom passed and 6 (12%) failed.

5.1.1.4 Bursary

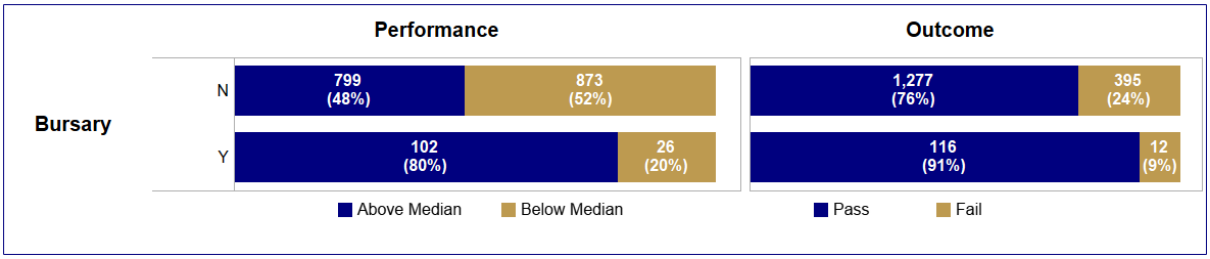


Figure 5.5: Bursary by Academic Performance and Outcome

Figure 5.5 shows the bursary status of the students. Of 1 672 students who did not receive a bursary, 52% scored below the median, and 48% scored above the median. Among the 128 students who received a bursary, 20% scored below the median, and 80% scored above the median. The same pass and fail rates were observed when the data were analysed by population group. The bursary status of the students was also considered. Of the 1 672 students without a bursary (N), 1277 (76%) passed and 395 (24%) failed. Of the 128 students with a bursary (Y), 116 (91%) passed and 12 (9%) failed.

5.1.1.5 Residence

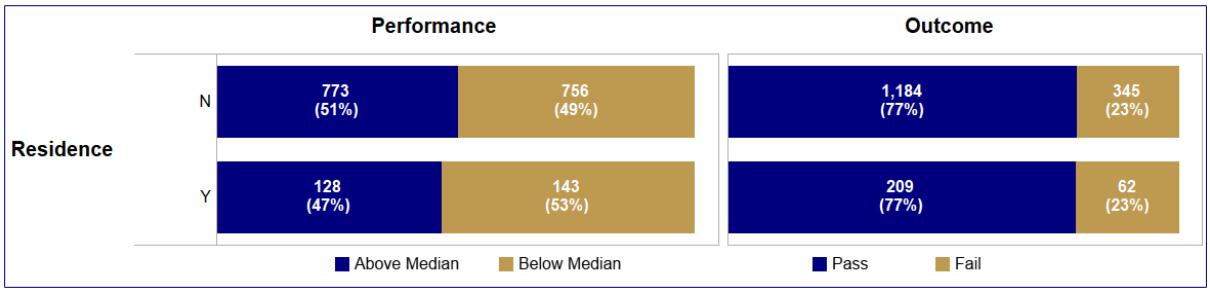


Figure 5.6: Residence by Academic Performance and Outcome

Figure 5.6 shows residence status is another factor in the data. Among the 1 529 students who do not reside on campus, 49% scored below the median, and 51% scored above the median. Of 271 students who reside on campus, 53% scored below the median, and 47% scored above the median. Data were also analysed by whether the students lived in residence. Of the 1 529 students not living in residence (N), 1 184 (77%) passed and 345 (23%) failed. Of the 271 students who lived in residence (Y), 209 (77%) passed and 62 (23%) failed.

5.1.1.6 Programme

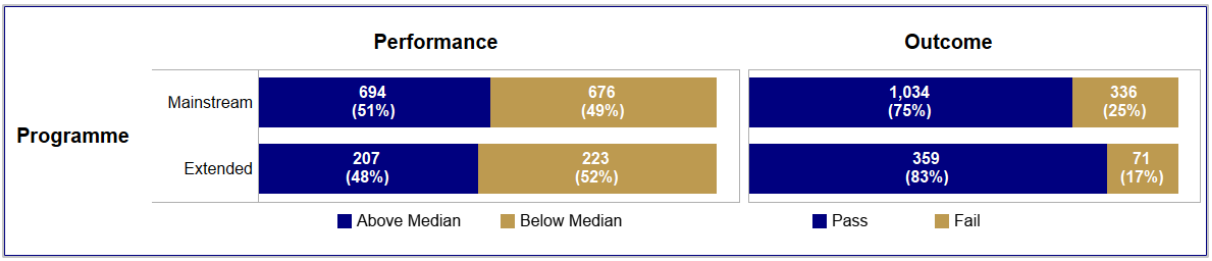


Figure 5.7: Programme by Academic Performance and Outcome

Figure 5.7 shows a breakdown of two academic programs — the mainstream programme and the extended programme — based on the first-year average marks of the students about the median GPA. For the mainstream programme, there is a near-equal distribution between students below and above the median GPA, with 676 students (49%) falling below the median and 694 students (51%) exceeding it. In the extended programme, however, there is a slight skew towards students performing below the median, with 223 students (52%) falling below the median and 207 students (48%) above it. These figures suggest that the mainstream programme has a more balanced performance among students regarding the median GPA. In contrast, the extended programme has a higher percentage of students scoring below the median.

The mainstream and extended programme are categorised by pass or fail rates. In the mainstream programme, 1 370 students are accounted for, with 1 034 students (75%) having passed and 336 students (25%) having failed. This indicates that most students in the mainstream programme successfully met the criteria to pass their first year. In comparison, the extended programme shows a higher success rate, with 359 of its 430 students passing, which translates to 83% of its population. Failures in the extended programme are notably lower, with only 71 students (17%) not meeting the passing criteria. Overall, both programs demonstrate most students passing, with the Extended Programme exhibiting a higher pass rate than the mainstream programme.

The mainstream programme shows nearly equal distribution around the median GPA, while the extended programme shows a slight bias towards lower performance. Although more students in the Extended Programme fall below the median GPA, it has a higher pass rate of 83% compared to the mainstream's 75%. This indicates an association between programme type and student performance, with the extended programme seemingly better supporting students to pass despite a lower median GPA.

5.1.2 Numerical Academic Attributes

The data presents a descriptive statistical analysis of the *numerical data*: APS (Admission Point Score) and GPA (Grade Point Average) for the academic performance (Below median and Above Median) and outcome (Pass and Fail).

5.1.2.1 Academic Performance

The statistics summary in Table 5.1 include the frequency, mean, median, standard deviation, and skewness for each category (Below median and Above Median) within each variable:

Table 5.1: Summary Statistics Table for Interval and Ratio Variables by Performance

Variables	Performance	Count	Mean	Median	Std. Dev.	Skewness
APS	Below Median	899	39.1	39	6.26	-2.21
	Above Median	901	40.9	41	8.53	-1.98
GPA	Below Median	899	47.9	52.8	14.2	-1.98
	Above Median	901	67.8	67.1	5.08	0.73

For the APS, there were 901 instances in the 'Above Median' category. The mean APS for this group was 40.9, with a median of 41 and a mode of 34. The standard deviation was 8.53, indicating a relatively wide spread of scores around the mean. The minimum APS was 0, while the maximum was 61. The skewness of -1.98 suggests a negatively skewed distribution with a heavy tail.

The 'Below Median' category for APS comprised 899 instances. The mean APS was slightly lower at 39.1, with a median of 39 and a mode of 37. The standard deviation was 6.26, suggesting a tighter distribution of scores around the mean compared to the 'Above Median' group. The minimum APS was 0, and the maximum was 58. The skewness of -2.21 indicates a more negatively skewed distribution with a heavier tail than the 'Above Median' group.

For the GPA, the 'Above Median' category contained 901 instances. The mean GPA was 67.77, with a median of 67.1 and a mode of 62. The standard deviation was 5.08, indicating a relatively narrow spread of scores around the mean. The minimum GPA was 60.63, and the maximum was 85.13. The skewness of 0.73 suggests a slightly positively skewed distribution with a light tail.

The 'Below Median' category for GPA consisted of 899 instances. The mean GPA was significantly lower at 47.96, with a median of 52.75 and a mode of 0. The standard deviation was 14.2, suggesting a wider distribution of scores around the mean compared to the 'Above Median' group. The minimum GPA was 0, and the maximum was 60.6. The skewness of -1.98 indicates a negatively skewed distribution with a moderately heavy tail.

The APS and GPA data suggest different distributions for students above and below the median. The 'Above Median' category for APS shows a wider spread of scores and a negative skew, indicating that while most students scored high, there is a significant number with much lower scores. The 'Below Median' APS group has a tighter score distribution but a more negative skew, indicating a clustering of scores closer to the median, although lower overall. For GPA, the 'Above Median' group shows a narrower spread and a slight positive skew, suggesting most students are achieving close to the average with a few high achievers. Meanwhile, the 'Below Median' GPA group shows a significantly lower mean with a wide spread of marks and a negative skew, showing a considerable number of students with marks much lower than the median. This indicates variability in student performance, with a general trend of those below the median struggling more significantly across both measures.

5.1.2.2 Academic Outcome

The statistics include the frequency, mean, median, standard deviation, and skewness for each category (Pass and Fail) within each variable:

Table 5.2: Summary Statistics Table for Interval and Ratio Variables by Outcome

Variables	Performance	Count	Mean	median	Std. Dev.	Skewness
APS	Pass	1 393	40.3	40	7.73	-1.93
	Fail	407	39	39	6.72	-2.31
GPA	Pass	1 393	63.31	63.25	7.89	-0.53
	Fail	407	39.3	44.5	16.66	-1.12

For the APS variable, it is observed that among the students who passed, the frequency is 1 393, with a mean value of 40.3 and a median of 40. The standard deviation is 7.73,

indicating moderate dispersion in the data. The skewness value of -1.93 suggests that the distribution is negatively skewed, with a tail extending towards lower values. On the other hand, among the students who failed, the frequency is 407, with a mean and median value of 39. The standard deviation is 6.72, and the skewness is -2.31, indicating a similar negatively skewed distribution.

Turning to the GPA variable, for the students who passed, the frequency remains at 1 393. The mean GPA is 63.31, and the median is 63.25. The standard deviation is 7.89, suggesting a moderate dispersion of data points. The skewness value of -0.53 indicates a slight negative skewness, but the distribution is relatively symmetrical compared to the APS variable. For the students who failed, the frequency is 407. The mean GPA is 39.3, and the median is 44.5. The standard deviation is 16.66, reflecting more variability in the data. The skewness value of -1.12 indicates a negatively skewed distribution.

The APS and GPA variables shows differences in the academic outcomes of students who passed versus those who failed. Students who passed have higher mean and median values in both APS and GPA, indicating better overall academic performance. These suggest that failing students not only have lower average scores but also a wider dispersion of outcomes, particularly in GPA, highlighting a noticeable variation in academic achievement within this group.

5.1.3 Student Expectation Profile

The summary statistics analysis in Tables B.1 and B.2 (**note:** refers to tables presented in the Appendices) on student expectation responses indicates a range of student views concerning their expectations and intentions as they enter university life. The mean scores for the questions, rated on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree), provide an average indication of the students' positions. Here are some key takeaways from the data:

Students generally plan to join social organisations and expect to make many new friends, including those from different racial groups. The mean scores for these expectations are high (Q1: 3.98, Q4: 3.99, Q5: 3.71). Most students anticipate engaging in academic discussions outside formal lectures and believe that making new friends will contribute to their academic success (Q6: 4.00, Q7: 4.49). Students strongly intend to use librarians and peer tutors to assist with their assignments and studies (Q8: 4.34, Q9: 4.22).

There is a high expectation that lecturers will be a source of academic support and that they will provide feedback on assignments and tests. Students also expect to manage their learning to some extent (Q14: 4.16, Q16: 4.15, Q17: 4.11, Q18: 4.12, Q23: 3.64). Students

have high expectations regarding the affordability and healthiness of food provided by the university cafeteria and generally feel safe on campus (Q24: 4.07, Q25: 3.81, Q29: 4.07). Access to the Internet and other learning resources is expected to be highly available (Q27: 4.39).

Students have positive expectations about the supportiveness of classmates and the university's care for their welfare (Q30: 3.95, Q31: 4.05, Q32: 4.01). Students are generally confident they can balance their studies with other responsibilities (Q34: 3.96). Financial issues are expected to concern a significant portion of the student body, potentially distracting them from their studies (Q35: 4.09). The lower mean scores are associated with the clarity of the workload expectation (Q10: 3.02), the possibility of spending much time in the library (Q11: 3.16), and the awareness of academic integrity and plagiarism requirements (Q33: 3.03). These areas may indicate where students feel less confident or have lower expectations.

Overall, students enter university with high expectations for social engagement, academic support, and personal welfare but also recognise potential challenges related to workload management, financial issues, and academic integrity.

5.1.4 Student Experience Profile

The summary tables B.3 and B.4 (**note:** refers to tables presented in the Appendices) provide an analysis of the student experience based on responses to a questionnaire with 35 items. Each question was rated on a 5-point Likert scale, with options ranging from "1: Strongly Disagree to "5: Strongly Agree. The questions have been rephrased from future expectations ("I will...) to present experiences ("I have...) to reflect the current state of the respondents better. The mean scores and the distribution of responses were analysed to provide insights into the students' experiences. Here is a summary of the insights:

Most students (52%) did not join social organisations, and an equal percentage (49%) believed that joining such organisations did not distract from their academic work, with a mean score of 1.96 for both Q1 and Q3, indicating a tendency towards disagreement. A sizable number of students (40%) also did not attend many social functions, with a mean score of 2.24, suggesting a slight inclination towards disagreement. A significant number of students (37%) made new friends. Similarly, 36% of students made new friends from different racial groups, with high mean scores of 3.88 and 3.80, respectively, indicating agreement that these experiences were common. Making new friends was seen as supportive of academic success (mean score of 3.67). Financial issues were a concern for many students, with 41% agreeing that it distracted them from studies, reflected in a mean score of 3.70.

Most students (69%) were involved in academic discussions outside formal lectures, which was perceived to enhance learning (mean score of 3.85). Usage of peer tutors was common (55% agreed or strongly agreed), with a mean score of 3.40, suggesting a positive experience. Students were generally clear on the expected workload (mean score of 3.58) and spent significant time in the library (mean score of 3.10). There was a strong expectation for academic essay writing and proper referencing, with high mean scores (above 4.0) indicating that students felt these were clear expectations.

Many students (85%) agreed they could self-manage their learning and were resourceful in finding university information (mean score of 4.20). However, 50% of the students disagreed or were neutral about finding university procedures and support independently (mean score of 2.48). The majority felt safe on campus (57%), with a high mean score of 4.37, indicating strong agreement. The university's infrastructure was generally rated positively for signage and accessibility (mean score of 3.95 for Q28).

In summary, social activities did not play a central role for most students, and friendships were considered supportive for academic success. There was a strong engagement in academic discussions outside of classes, and students used academic support services. Students felt the expectations of academic skills were clear, and they could self-manage their learning effectively. Infrastructure and safety on campus were generally viewed positively, whereas financial issues presented a notable distraction for students. These insights can inform strategies to enhance student experience and academic success.

5.1.5 Expectation and Experience Gap

The data presented in Tables B.5 and B.6 (**note:** refers to tables presented in the Appendices) shows insights into the gap between student expectations and their actual university experiences, focusing on various academic and campus life aspects. The analysis measures include mean, standard deviations, and Z-scores, which help understand the data distribution between students' expectations and actual experiences. The interpretation is based on a rule where a negative mean indicates high expectations that exceed the experience, while a positive mean suggests that the experience exceeds expectations. A mean of zero indicates that expectations and experiences are met.

The standard deviation measures the degree of variation or spread in the responses for each question. A higher standard deviation indicates a wider range of responses, signifying greater variability in student perceptions regarding that aspect. The Z-scores provide a standardised measure of how each question's mean deviates from the overall mean. A negative Z-score indicates that the expectation for that aspect was significantly higher (in terms of standard deviations) than the overall mean. At the same time, a positive Z-score

suggests that the experience exceeded expectations significantly. A Z-score close to zero implies that the expectation and experience were relatively in line with the overall mean.

Students generally have high expectations in several areas. For example, in questions like "I will join social organisations/clubs on campus this year" and "Joining social clubs/organisations at university will distract me from my academic work", students express high expectations, as indicated by the negative mean of -2.016 and -1.847, respectively. However, their experiences tend to fall short of these high expectations, with negative Z-scores indicating a substantial gap between what they anticipate and what they encounter. Several other questions, such as Q1, Q3, Q8, Q30, and Q32, have negative mean values, indicating students' high expectations in these areas.

Some aspects of university life show a trend where experiences exceed expectations. For example, in questions like "My lecturers will expect me to attend all my lectures" and "I will be able to balance my first-year university study with other responsibilities", the positive mean values of 0.749 and 0.399 indicate that students have relatively lower expectations. However, their experiences surpass these expectations, as evidenced by positive Z-scores. This suggests that in certain areas, students may underestimate the support and opportunities available to them at the university. Questions like Q10, Q19, and Q28 have positive mean values, suggesting that students had relatively low expectations in these areas.

Some questions, such as Q4, Q5, Q11, and Q24, have mean values close to zero, indicating that students' expectations were reasonably aligned with their experiences. In these cases, students neither had exceptionally high nor low expectations, and their experiences were relatively consistent with their expectations.

This statistical analysis highlights that there is a significant gap between student expectations and their actual experiences at the university. These disparities are reflected in the mean values, standard deviations, and Z-scores, demonstrating that students often have high expectations in some areas but encounter lower-than-expected experiences. In other aspects, students appear to have modest expectations that their experiences exceed. Understanding these gaps can help the university tailor their support and services to meet student needs better and align expectations with reality, ultimately enhancing students' overall satisfaction and success during their academic journey.

5.1.6 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is a powerful statistical technique to uncover underlying patterns and structures within a dataset by identifying latent factors that explain the observed correlations among variables. This section focuses on the results and inter-

pretation, providing insights into the process of factor extraction, determination of the number of factors to retain, evaluation of sample size adequacy, and the interpretation of factor loadings.

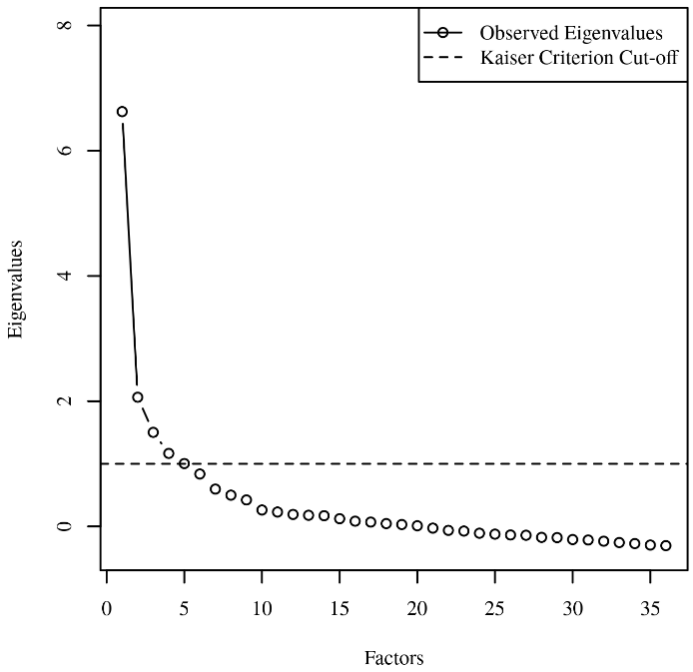


Figure 5.8: Scree plot incorporating the Kaiser criterion

The Kaiser criterion was used to decide the number of factors to retain. This rule stipulates that all factors with an eigenvalue greater than one should be retained for interpretation. The eigenvalues were extracted from the correlation matrix, with the diagonal of the matrix replaced by the squared multiple correlations of each variable [68] to estimate the communality of each variable [32]. Applying Kaiser's rule of greater eigenvalue than one is a common research practice [68]. Figure 1 presents the scree plot alongside the Kaiser criterion to determine the number of significant factors. On examination of Figure 5.8, five factors with an eigenvalue greater than one were identified. Consequently, four factors were retained for the factor.

Factor 1 accounted for 12.82% of the variance with an eigenvalue of 4.61. Factor 2 explained 7.56% of the variance and had an eigenvalue of 2.72. Factor 3 accounted for 5.30% of the variance with an eigenvalue of 1.91. Factor 4 explained 5.07% of the variance and had an eigenvalue of 1.83. Factor 5 accounted for 3.81% of the variance with an eigenvalue of 1.37. The five-factor model collectively represented 34.55% of the total variance in the data.

The Scree Plot affirmed this selection, displaying an "elbow" at the fourth component, suggesting no further extraction is necessary. The identified factors are Factor 1 - Effective

Table 5.3: Eigenvalues, Percentages of Variance, and Cumulative Percentages for Factors

Factor	Eigenvalue	% of Variance	Cumulative %
1	4.61	12.82	12.82
2	2.72	7.56	20.38
3	1.91	5.30	25.68
4	1.83	5.07	30.75
5	1.37	3.87	34.55

Learning, Factor 2 - Social Wellbeing, Factor 3 - Academic Support, and Factor 4 - Access to Information.

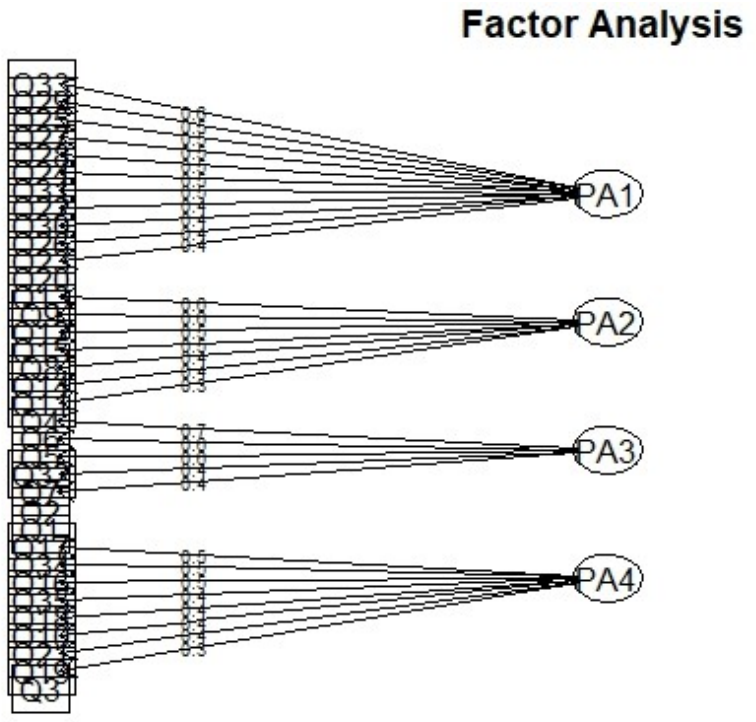


Figure 5.9: Exploratory Factor Analysis Plot

This suggests that the factor structure may be suitable for the data [27] for only four factors. Costello and Osborne [27] also suggest dropping variables with low communality, cross-loadings, and any variable that is the only significant load on a factor, which can prevent a weak factor structure and alleviate these problems hence the following variables were excluded in the factor: Q1, Q2, Q20 and Q35.

In Summary, the exploratory factor analysis conducted on the 35-item variable set revealed the presence of four factors, each contributing to the explanation of variance in the data. Factor 1 showed the highest eigenvalue, followed by Factors 2, 3, and 4.

Cumulatively, these factors accounted for a substantial portion of the variance.

5.2 Student Expectation - Predictive Models

This section explores the results of the Student Expectation - predictive models. It covers the measures that underpin the usefulness of various prediction models, each tailored to anticipate academic success with increasing precision. The models include Generalized Linear Model (GLM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Regularized Discriminant Analysis (RDA). Subsequent subsections will explain the results derived from confusion matrices, offering a transparent view of each model's performance in classifying outcomes correctly. Finally, a discussion is presented on the variables of importance as they form the cornerstone of the predictions, shaping interventions that strengthen student success.

5.2.1 Student Performance - Models

The student expectation-performance predictive model evaluates the capability of various statistical learning models to predict students' academic performance, specifically distinguishing between those who score above and below the median first-year average mark. The key performance metrics include Accuracy, Sensitivity (Recall), Specificity, Precision, and the F1 Score, which is the harmonic mean of Precision and Recall.

Table 5.4: Balanced Model Performance Measures for Student Expectation and Performance

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.60	0.54	0.65	0.61	**0.54**	**0.57**
KNN	0.58	**0.55**	0.61	0.58	0.55	0.56
GBM	0.60	0.52	0.66	0.61	0.52	**0.57**
DT	0.58	0.45	0.72	0.61	0.45	0.52
RF	0.59	0.54	0.64	0.60	**0.54**	**0.57**
SVM	0.60	0.53	0.65	0.60	0.53	0.56
NB	0.58	0.23	**0.90**	**0.72**	0.23	0.38
RDA	**0.61**	0.48	0.73	0.64	0.48	0.55

The RDA model shows the highest Accuracy (0.61) and Specificity (0.73), suggesting a robust ability to correctly identify students who will not score above the median (True

Negatives). However, its Sensitivity is moderate (0.48), indicating room for improvement in correctly predicting students who will score above the median (True Positives).

On the other hand, the NB model, despite having a high Specificity (0.90) and Precision (0.72), suffers from the lowest Sensitivity (0.23) and F1 Score (0.38), which is not ideal for a balanced predictive performance where the 'Positive' class is of primary interest. Models like the GLM, GBM, and SVM present a balance between the metrics, with F1 Scores ranging from 0.56 to 0.57, indicating a relatively moderate capability to predict above-median performance.

The RDA model could be considered the most accurate overall. Yet, the choice of the model may depend on the specific cost-benefit analysis of false positives versus false negatives in the context of the educational institution's objectives.

5.2.1.1 Confusion Matrix

The confusion matrices for the Student Expectation Performance Predictive Model (fig. 5.10) provide insights into the accuracy of each predictive model in classifying students above (Positive Class: 1) or below (Negative Class: 0) the median first-year average mark.

GLM

		Reference	
		0	1
Prediction	0	173	121
	1	92	142

KNN

		Reference	
		0	1
Prediction	0	160	119
	1	105	144

GBM

		Reference	
		0	1
Prediction	0	175	124
	1	90	139

DT

		Reference	
		0	1
Prediction	0	191	146
	1	74	117

RF

		Reference	
		0	1
Prediction	0	169	120
	1	96	143

SVM

		Reference	
		0	1
Prediction	0	173	125
	1	92	138

NB

		Reference	
		0	1
Prediction	0	238	195
	1	27	68

RDA

		Reference	
		0	1
Prediction	0	194	137
	1	71	126

Figure 5.10: Expectation Balanced Dataset Confusion Matrix

GLM shows a tendency to identify better students who are above the median mark (142 true positives) compared to the KNN model, which shows a slightly lower number of true positives (144) but more false negatives (119). The GBM model shows a balanced classification with many true negatives (175) but a lower count of true positives (139) than GLM and KNN. The DT model tends to predict more false negatives (146) than other models, suggesting a conservative bias towards predicting students as below the median.

RF and SVM models have similar true positive rates (143 and 138, respectively). However, RF has slightly fewer false negatives (120). Notably, the NB model is the most conservative, with the highest number of true negatives (238) and the highest number of false positives (195), indicating a high level of misclassification for students who are

actually above the median. RDA offers a middle ground, with fewer false positives (137) than NB and a modest number of true positives (126).

Overall, the models demonstrate varied abilities to classify students accurately. Models like GLM and KNN are more balanced in predicting positive and negative classes. In contrast, models like NB may require further calibration to reduce the high number of false positives. The choice of model may thus depend on the specific educational context and whether the cost of false positives or false negatives is more critical to address.

5.2.1.2 Variable of Importance

This subsection focuses exclusively on the GLM and the KNN algorithm, identifying them as the most effective models. However, it is important to note that while assessing feature importance is straightforward for the GLM, the KNN Classification algorithm does not inherently provide this information. Figure 5.11 shows the GLM summary output of the Expectation Performance Predictive Models that predict whether a student's first-year average mark is above (1) or below (0) the median; the results highlight significant factors that correlate with academic performance outcomes.

The logistic regression model's summary indicates several variables significantly influencing a student's likelihood of achieving an above-median first-year average mark. The variable Bursary1 (β = 0.9833, p = 0.00130) suggests that students with a bursary are more likely to achieve above-median marks. This statistically significant finding could reflect the positive impact of financial support on academic performance.

The APS coefficient ($\beta = 0.0692$, $p < 0.0001$) is positive and highly significant, showing that higher Admission Points Scores are strongly associated with being above the median in terms of marks. This could indicate that the APS is a robust predictor of academic success.

Variables FA3 (Academic Support) and Female1 show positive relationships with the outcome (β = 0.4512, p = 0.00308 for FA3 and β = 0.4787, p = 0.00019 for Female1), implying that academic support and being female are linked to higher academic achievement.

On the other hand, FA4 (Access to Information) has a negative coefficient (β = -0.3808, p = 0.02186), African1 (β = -0.6050, p = 0.02454), and Mainstream1 (β = -0.3451, p = 0.02104]) also have negative coefficients, suggesting that access to information, being an African student, and being in the mainstream program are associated with lower likelihoods of achieving above-median marks. These could be areas of concern that may require further investigation and intervention.

The Akaike Information Criterion (AIC) serves as a statistical measure designed to

```
> summary(ESP_perf_glm_mod)

Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.385  -1.097  -0.641   1.142   1.837

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2912     0.8730   -3.77  0.00016 ***
Bursary1      0.9833     0.3058    3.22  0.00130 **
APS           0.0692     0.0130    5.34  9.2e-08 ***
Residence1   -0.1313     0.1721   -0.76  0.44560
FA1           0.0506     0.1698    0.30  0.76582
FA2           0.1063     0.1463    0.73  0.46734
FA3           0.4512     0.1524    2.96  0.00308 **
FA4          -0.3808     0.1661   -2.29  0.02186 *
Female1       0.4787     0.1281    3.74  0.00019 ***
African1     -0.6050     0.2690   -2.25  0.02454 *
Mainstream1  -0.3451     0.1496   -2.31  0.02104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1713.4  on 1235  degrees of freedom
Residual deviance: 1602.7  on 1225  degrees of freedom
AIC: 1625

Number of Fisher Scoring iterations: 4
```

Figure 5.11: Expectation Balanced Dataset: Logistic Regression Model

assess the relative quality of statistical models for a specific dataset. It indicates that the model in question exhibits a superior fit compared to the null model, as evidenced by its lower AIC value. The model fits better than the null model (as indicated by the lower AIC). However, with residual deviance of 1602.7 on 1225 degrees of freedom, there is still unexplained variability in the model, and further improvements could be made.

In summary, variables such as having a bursary, higher APS, increased academic support, and being female are associated with an increased likelihood of passing. Meanwhile, being in a mainstream program and being African are associated with a decreased likelihood of passing. The variable Access to Information shows a surprising negative impact on the likelihood of passing, which suggests that the relationship between information

access and academic success might be more complex than initially presumed. These insights can be used to tailor student interventions and support services to improve student outcomes.

5.2.2 Academic Outcome - Models

This section explores the development and evaluation of predictive models for academic outcomes, explicitly focusing on binary classification: Pass (1) and Fail (0). Addressing the challenge posed by the unbalanced nature of these categories, the analysis includes four distinct modelling approaches: the Unbalanced Model, which serves as a baseline by using the original dataset without adjustments for imbalance; the ROSE (Random OverSampling Examples) Model, which generates synthetic samples to achieve class balance; the Oversampling Model, which replicates instances of the minority class to equalise representation; and the Undersampling Model, which reduces instances of the majority class to match the minority class count. These methods are crucial in providing a holistic understanding of how different data balancing techniques impact the predictive accuracy of academic outcomes.

Evaluating the performance of these models requires a nuanced approach due to the dataset's unbalanced nature. The study uses various measures, including Accuracy, Precision, Recall (or Sensitivity), and F1 Score. Each measure offers unique insights: while accuracy measures overall correctness, Precision and Recall provide a more detailed view of positive class prediction quality, and the F1 Score balances them. A confusion matrix further enriches the analysis by clearly illustrating the types and numbers of correct and incorrect predictions, thereby providing a deeper understanding of model strengths and weaknesses.

A critical aspect of our investigation is identifying and understanding the variables that significantly influence academic outcomes. This involves analysing feature importance across different models to determine which factors most predict academic success or failure. Such insights are pivotal for developing targeted interventions and informed policy decisions to enhance educational achievement. Additionally, our analysis pays special attention to how each model addresses the challenges posed by the unbalanced data, evaluating the effectiveness of techniques like ROSE, oversampling, and undersampling in mitigating this imbalance and their subsequent impact on model interpretation and performance. This comprehensive approach ensures a robust understanding of predictive modelling in the academic context, focusing on addressing data imbalance challenges.

5.2.2.1 Unbalanced Models

Table 5.5 shows the performance measures for various predictive models applied to the unbalanced academic outcome dataset, where the positive class (Pass) is denoted as '1' and the negative class (Fail) as '0'. The models assessed include GLM, KNN, GBM, DT, RF, SVM, NB, and RDA. The key performance measures for these models are Accuracy, Sensitivity, Specificity, Precision, Recall, and F1 Score.

Table 5.5: Unbalanced Model Performance Measures for Student Expectation and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	**0.78**	**1.00**	0.02	**0.78**	**1.00**	**0.87**
KNN	0.77	0.98	0.07	**0.78**	0.98	**0.87**
GBM	0.77	0.98	0.07	**0.78**	0.98	**0.87**
DT	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**
RF	0.77	0.99	0.01	0.77	0.99	**0.87**
SVM	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**
NB	0.77	0.96	0.09	**0.78**	0.96	0.86
RDA	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**

A consistent observation across almost all models is the high Sensitivity (also equal to Recall in this context), which indicates a strong ability to correctly identify the positive class (Pass). However, this is contrasted by notably low Specificity scores, particularly in GLM, DT, SVM, and RDA models. Specificity is at 0.00 or 0.02, indicating a poor performance in correctly identifying the negative class (Fail). This imbalance suggests that the models are biased towards predicting the more prevalent class (Pass), a common issue in unbalanced datasets.

The Accuracy scores hover around 0.77 to 0.78 for all models, which might seem adequate at first glance. However, given the unbalanced nature of the dataset, this metric might be misleading. The Precision, Recall, and F1 Scores offer a more nuanced insight into model performance. Precision remains relatively stable across models, around 0.77 to 0.78, indicating a decent but not outstanding ability of the models to predict the positive outcomes out of all correctly predicted positives. The F1 Scores, which balance Precision and Recall, are consistently high, hovering around 0.87 for most models. This high F1 Score is primarily driven by the high Recall values, again highlighting the models' tendency to favour the positive class.

In summary, while the models demonstrate high Sensitivity/Recall and F1 Scores, the extremely low Specificity scores across most models reveal a significant bias towards

predicting the positive class. This highlights the challenges and limitations of using these models in an unbalanced dataset context, especially when the goal is to accurately predict both positive (Pass) and negative (Fail) classes. The analysis underscores the importance of considering a range of performance metrics and not relying solely on Accuracy, especially in scenarios with unbalanced class distributions.

Confusion Matrix: The confusion matrix for the Student Expectation - Unbalanced Outcome Model reveals the performance of various machine learning models in predicting academic outcomes where 'Pass' is labelled as '1' and 'Fail' as '0'.

GLM

		Reference	
		0	1
Prediction	0	2	0
	1	120	417

KNN

		Reference	
		0	1
Prediction	0	8	8
	1	114	409

GBM

		Reference	
		0	1
Prediction	0	8	10
	1	114	407

DT

		Reference	
		0	1
Prediction	0	0	0
	1	122	413

RF

		Reference	
		0	1
Prediction	0	1	4
	1	121	413

SVM

		Reference	
		0	1
Prediction	0	0	0
	1	122	417

NB

		Reference	
		0	1
Prediction	0	11	15
	1	111	402

RDA

		Reference	
		0	1
Prediction	0	0	0
	1	122	417

Figure 5.12: Expectation Unbalanced Dataset Confusion Matrix

The GLM has almost no true negatives or false positives, with all predictions favouring a pass. This indicates a high bias towards predicting a pass regardless of the actual outcome, evident from the 120 false negatives. The KNN slightly improves with eight true negatives and eight false positives. However, it still leans heavily towards predicting a pass, as the 114 false negatives indicate. The GBM and DT models are similar in their predictions, with the DT having slightly more false negatives (122) than GBM's 114 and no true negatives or false positives, indicating a bias towards predicting passes.

The RF model has a single true negative and four false positives, with 121 false negatives. This model slightly improves specificity over the GLM but still shows a strong bias towards predicting passes. The SVM and RDA predict that all students will pass, as evidenced by the absence of true negatives and false positives. This results in a model that does not effectively differentiate between the outcomes. The NB model has the highest number of true negatives (11) and false positives (15), suggesting a better balance in prediction but still with a substantial number of false negatives (111).

In summary, all models show a strong propensity to predict that students will pass, which could be due to the unbalanced nature of the dataset. The KNN and NB models can only predict true negatives, indicating some ability to identify students who will fail. However, the high number of false negatives across all models suggests that they are not

effectively capturing the characteristics of students who fail. The high false negative rate implies that these models may not be reliable for intervention strategies to assist students at risk of failing. The lack of true negatives and the high number of false negatives in models like GLM, DT, RF, SVM, and RDA indicate a need for further model tuning or consideration of alternative models or resampling techniques to handle the unbalanced dataset better.

5.2.2.2 Random Over Sampling Examples (ROSE) Model

In the Student Expectation - ROSE (Random Over-Sampling Examples) Outcome Model summary, several predictive models were assessed to determine their effectiveness in predicting student academic outcomes where 'Pass' is denoted as '1' and 'Fail' as '0'. The models were evaluated based on their accuracy, sensitivity (also known as recall for the positive class), specificity, precision, recall, and F1 score.

Table 5.6: ROSE Model Performance Measures for Student Expectation and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.60	0.56	0.71	**0.87**	0.56	0.68
KNN	0.61	0.66	0.43	0.80	0.66	0.72
GBM	0.64	0.65	0.57	0.84	0.65	0.74
DT	**0.65**	**0.69**	0.52	0.83	**0.69**	**0.76**
RF	**0.65**	**0.69**	0.53	0.83	**0.69**	**0.76**
SVM	**0.65**	0.68	0.53	0.83	0.68	0.75
NB	0.35	0.19	**0.89**	0.86	0.19	0.31
RDA	0.59	0.57	0.69	0.86	0.57	0.68

The GLM model demonstrated a moderate accuracy of 0.60 and a sensitivity of 0.56, indicating that it correctly predicted 56% of the students who would pass. The GLM's precision was quite high at 0.87, suggesting that when it predicted a student would pass, it was correct 87% of the time. However, its F1 score was 0.68, reflecting a need to improve its balance between precision and recall. KNN model slightly improved accuracy to 0.61 and sensitivity to 0.66, meaning it correctly identified 66% of the students who would pass. However, KNN's specificity was lower at 0.43, indicating a less effective performance in correctly identifying students who would fail.

GBM and DT models offered better accuracy at 0.64 and 0.65, respectively. Both models had an F1 score higher than GLM and KNN, with GBM at 0.74 and DT at 0.76, suggesting a more robust predictive performance. The RF model matched the

DT's accuracy at 0.65 and demonstrated the highest sensitivity at 0.69, indicating that it correctly predicted 69% of the students who would pass. The RF's F1 score was also 0.76, tied for the highest among all models, indicating a strong performance balance.

The SVM shared the highest accuracy with DT and RF at 0.65 and had an F1 score of 0.75, just slightly below RF and DT, making it a competitive option. On the contrary, the NB model significantly underperformed in accuracy (0.35) and sensitivity (0.19), with the lowest F1 score at 0.31, indicating a poor overall predictive performance. RDA had an accuracy of 0.59 and an F1 score of 0.68, placing it in the mid-range of model performance.

Considering the balance of all metrics and the importance of correctly predicting students who will pass and minimising false positives, the RF model stands out as the best model. Its high sensitivity, combined with an accuracy of 0.65 and the top F1 score of 0.76, reflects its capacity to provide reliable predictions, making it the most suitable model for educational institutions seeking to allocate resources effectively to support student success.

Confusion Matrix: The confusion matrix summary for the Student Expectation - ROSE (Random Over-Sampling Examples) Outcome Model with Academic Outcome demonstrates the performance of various predictive models in determining whether students pass (1) or fail (0).

GLM		Reference	
		0	1
Prediction	0	86	182
	1	36	235

KNN		Reference	
		0	1
Prediction	0	53	143
	1	69	274

GBM		Reference	
		0	1
Prediction	0	69	144
	1	53	273

DT		Reference	
		0	1
Prediction	0	63	128
	1	59	289

RF		Reference	
		0	1
Prediction	0	65	130
	1	57	287

SVM		Reference	
		0	1
Prediction	0	65	133
	1	57	284

NB		Reference	
		0	1
Prediction	0	109	339
	1	13	78

RDA		Reference	
		0	1
Prediction	0	84	181
	1	38	236

Figure 5.13: Expectation ROSE Dataset Confusion Matrix

The GLM model correctly predicted a pass for 235 students and a fail for 86 students but also incorrectly predicted 182 students as failing and 36 as passing. The KNN algorithm showed more true positives, correctly predicting 274 students as passing. However, it also had many false negatives, with 143 students who passed being predicted as failing. The GBM presented a balanced performance, with 273 true positives and a similar number of false negatives to the KNN model. On the other hand, the DT classifier predicted pass outcomes with higher accuracy, correctly identifying 289 students as passing while having fewer false negatives (128) compared to KNN and GBM.

The RF model showed strong results with the highest number of true positives (287) among the models, indicating its strength in correctly identifying students who will pass. However, it also incorrectly predicted that 130 students would fail. The SVM model had a similar pattern to the RF model, with more true positives (284) but slightly more false negatives (133). The NB classifier model had the poorest performance with the highest number of false negatives (339), indicating a tendency to predict a failed outcome when students actually passed. Lastly, the RDA showed a better balance, with 236 true positives and fewer false negatives (181) than other models.

Overall, the Random Forest, Decision Tree and Support Vector Machine models exhibited the highest number of correct predictions for students passing the course, making them potentially the most effective models for predicting academic success in this oversampled dataset. However, the final selection of the best model would depend on the specific costs associated with false positives and false negatives in the educational context.

Variable of Importance: The Random Forest model is best for predicting student outcomes; it provides a strong balance between identifying students who will pass and those who will fail, making it the most suitable choice for determining which variables are most important for predicting academic success. The paragraphs below analyse and discuss the importance of the variable specifically for the Random Forest model to understand which factors most strongly predict academic outcomes.

Table 5.7: Expectation ROSE Variable Importance in the Random Forest Model

Importance Variable	**Mean Decrease Gini**
FA4 (Access to Information)	23.8
APS (Admission Points Score)	20.0
FA3 (Academic Support)	18.1
FA2 (Social Well-being)	17.1
FA1 (Effective Learning)	16.6
African	9.6
Female	9.4
Bursary	7.6
Mainstream	5.4
Residence	2.2

The Random Forest model's analysis for the Student Expectation - ROSE Outcome Model highlights the significance of various factors in predicting academic success. The

Access to Information (FA4) stands out with a Mean Decrease Gini of 23.8, suggesting that students with better access to resources and data tend to perform better academically. This highlights the necessity for institutions to ensure that all students have equal access to the information they need to succeed.

Following closely, the Admission Points Score (APS) is the second most influential factor, with a Mean Decrease Gini of 20.0. This indicates that the APS strongly predict their likelihood to pass, reinforcing the traditional view that past academic performance can be a reliable indicator of future success. Academic Support (FA3) is also a significant determinant of student outcomes, with a Mean Decrease Gini of 18.1. This reflects the importance of the support provided by the institution, including tutoring, advising, and other academic services that can help students achieve their academic goals.

The model also places considerable importance on Social Well-being (FA2) and Effective Learning (FA1), with scores of 17.1 and 16.6, respectively. These factors highlight the role of a supportive social environment and effective learning strategies in contributing to a student's academic performance. Moreover, 'Ethnicity' (African) and 'Gender' (Female) have scores of 9.6 and 9.4, indicating potential demographic influences on educational outcomes that may necessitate further research into how these attributes intersect with academic success.

Financial factors and program types also play a role, though to a lesser extent. A bursary and enrollment in a mainstream program are associated with Mean Decrease Gini scores of 7.6 and 5.4, respectively, suggesting that while financial aid and curriculum structure have an impact, they are less predictive than the previously mentioned variables. Residence, with the lowest score of 2.2, appears to have a minimal impact on academic success, which could indicate that other factors outweigh the importance of living arrangements.

Overall, the Random Forest model points to Access to Information, APS, and Academic Support as the three most critical areas for institutions to focus on to boost student achievement. There is a need to channel resources and interventions towards these key areas to make a substantial difference in fostering student success.

5.2.2.3 Oversampling Model

Table 5.8 Student Expectation - Oversampling Outcome Model with Academic Outcome provides a comparative overview of several statistical models, evaluating their ability to predict whether a student will pass (1) or fail (0).

The RF model stands out with the highest accuracy of 0.75, indicating that it correctly predicts 75% of the outcomes and F1 scores of 0.84. It also has the highest sensitivity,

Table 5.8: Oversampling Model Performance Measures for Student Expectation and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.69	0.75	0.48	0.83	0.75	0.79
KNN	0.66	0.79	0.21	0.77	0.79	0.78
GBM	0.67	0.79	0.26	0.79	0.79	0.79
DT	0.60	0.62	0.53	0.82	0.62	0.71
RF	**0.75**	**0.86**	0.37	0.82	**0.86**	**0.84**
SVM	0.69	0.77	0.45	0.83	0.77	0.80
NB	0.38	0.24	**0.84**	**0.84**	0.24	0.38
RDA	0.69	0.76	0.47	0.83	0.76	0.79

at 0.86, suggesting it is most effective at correctly identifying students who will pass. Furthermore, the RF model achieves the best F1 score of 0.84, which indicates a strong balance between precision and recall, which is important for models where false positives and false negatives have significant implications.

The SVM and GLM models also perform well, each with an accuracy of 0.69 and F1 scores of 0.80 and 0.79, respectively. These models show a good balance between sensitivity and specificity, although they do not outperform the RF model. The NB model, despite having high specificity and precision, suffers from low overall accuracy (0.38) and sensitivity (0.24), reflected in its low F1 score (0.38). This indicates that while it is good at predicting failures, it fails to identify many students who will pass, which could be a major drawback in an educational context.

Given the high importance of correctly identifying as many true positive outcomes as possible without incurring many false positives, the Random Forest model is the most suitable choice. It provides the highest accuracy and the best balance of sensitivity and precision, making it the best model among those evaluated for the oversampling outcome model.

Confusion Matrix: The provided confusion matrices represent the outcomes of various models applied to an oversampled dataset concerning academic performance, where '1' indicates students who have passed and '0' indicates those who have failed.

The GLM model correctly predicted 314 students as passing and 58 as failing. However, it also incorrectly predicted 103 students as failing and 64 as passing. The KNN model had many true positives, correctly predicting 328 students as passing and 25 as failing, with 87 false negatives and 97 false positives. The GBM model demonstrated a similar pattern with 330 true positives and 32 true negatives, accompanied by many false

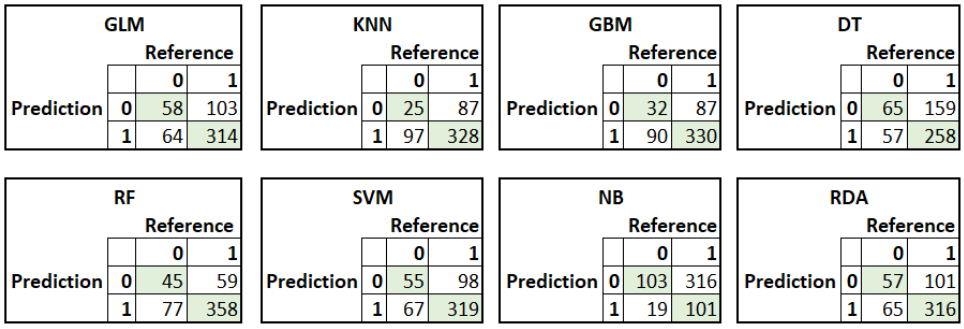


Figure 5.14: Expectation Oversampling Dataset Confusion Matrix

negatives (87) and false positives (90). The DT classifier showed a more balanced distribution with 258 true positives and 65 true negatives but with a relatively high number of false positives (57) and false negatives (159).

Moving on to the RF model, it correctly identified 45 students as failing and 302 as passing, showing fewer false negatives (59) compared to other models. The SVM model had a modest number of true positives (98) and true negatives (55), but also 98 false negatives and 55 false positives. The NB classifier had the highest number of false negatives (316), indicating a conservative tendency to predict failures, with 103 true negatives and 316 false negatives, accompanied by a relatively low number of false positives (18). Lastly, the RDA model correctly predicted 101 students as passing and 57 as failing, with a balanced number of false predictions in both classes (101 false negatives and 57 false positives).

Overall, these models show varying degrees of sensitivity and specificity. While some, like KNN and GBM, prioritise identifying as many true positives as possible, others, like NB, demonstrate a conservative approach with higher false negatives. There is a need to consider the trade-offs between these models regarding misclassification costs when choosing the most appropriate one for their context.

Variable of Importance: The best model would depend on what is most important for the application. If you prioritise accuracy, precision (minimising false positives) and F1-score, then RF is the best model. If recall (minimising false negatives) is more critical, the NB model might be considered despite its lower accuracy, precision and F1 score. However, considering a balance of all metrics, the RF model offers the best overall performance. The model assigns the greatest importance to the Admission Points Score (APS), indicating that this score is a key predictor of academic success.

Following APS, the most influential factors are those related to the accessibility of information (FA4) and effective learning strategies (FA1), which suggest that both the re-

Table 5.9: Expectation Oversampling Variable Importance in the Random Forest Model

Importance Variable	**Mean Decrease Gini**
APS (Admission Points Score)	25.6
FA4 (Access to Information)	23.1
FA1 (Effective Learning)	19.9
FA2 (Social Well-being)	19.4
FA3 (Academic Support)	16.3
Female	15.5
Mainstream	11.5
Bursary	8.8
African	6.1
Residence	3.3

sources available to students and how they use them are critical for academic performance. Additionally, a student's social well-being (FA2) and the level of academic support they receive (FA3) are also important, pointing towards the need for a supportive environment and infrastructure.

Gender (female) also shows significant importance, reflecting potential differences in performance outcomes between genders in this academic environment. The type of academic program (mainstream vs. extended), financial support (bursary), racial background (African), and whether a student stays in residence (Residence) have a noticeable but lesser impact on the prediction of academic outcomes according to this model.

The Random Forest model suggests that while individual abilities and conditions are key determinants of academic success, environmental factors and support systems also play non-negligible roles.

Figure 5.15 shows the statistical summary of the logistic regression output for the Student Expectation - Oversampling Outcome Model explains the factors contributing to academic success, defined here as passing (1) or failing (0).

The intercept's p-value is above the conventional alpha level of 0.05, suggesting that the model does not significantly predict the likelihood of passing versus failing when all other variables are held at zero. The negative coefficient for students with a bursary is significant and indicates a lower likelihood of passing. This could reflect that bursaries may be awarded to students who, despite the financial support, still face significant challenges that negatively impact their academic performance. This variable's importance should be interpreted with caution, and it calls for a deeper investigation into the profiles of bursary recipients and the challenges they encounter.

```
> summary(ESP_outc_glm_over)

Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -1.96   -1.06   -0.61    1.12    2.38

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.86208    0.66066    1.30   0.1919
Bursary1    -2.00333    0.43818   -4.57  4.8e-06 ***
APS         -0.02496    0.00774   -3.23   0.0013 **
Residence1  -0.22211    0.14257   -1.56   0.1193
FA1         -0.35090    0.14090   -2.49   0.0128 *
FA2         -0.54092    0.12778   -4.23  2.3e-05 ***
FA3         -0.14768    0.12162   -1.21   0.2246
FA4          0.78365    0.13612    5.76  8.6e-09 ***
Female1     -0.65674    0.10581   -6.21  5.4e-10 ***
African1     0.69219    0.24974    2.77   0.0056 **
Mainstream1  0.84166    0.13271    6.34  2.3e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2425.3  on 1763  degrees of freedom
Residual deviance: 2213.5  on 1753  degrees of freedom
AIC: 2235

Number of Fisher Scoring iterations: 5
```

Figure 5.15: Expectation Oversampling Dataset: Logistic Regression Model

The negative and significant coefficient for APS is counterintuitive, as higher academic scores are generally expected to correlate with better academic outcomes. This may suggest an upturn in the expected relationship within this academic context, or other confounding factors may be at play, such as the possibility that students with higher APS scores face more challenging curricula or greater performance pressure.

The lack of statistical significance for the residence variable suggests that living in residence does not impact the likelihood of passing in this model. FA1 (Effective Learning) and FA2 (Social Well-being): These factors show a significant negative relationship with the likelihood of passing. This is surprising, as effective learning practices and social well-being are typically seen as supportive of academic success. The negative coefficients

indicate the need for a deeper examination of how these factors are measured and the potential stressors or unmeasured variables that may affect students negatively.

The coefficient for academic support is not statistically significant, which suggests that the variable, as modeled, does not have a noticeable impact on passing rates. It could be that the type, quality, or relevance of academic support needs further refinement to understand its true effect. The positive and highly significant coefficient of FA4 (Access to Information) aligns with expectations. Access to information is crucial in academic settings, and this result highlights its importance as a predictor of academic success.

The significant negative coefficient of being female suggests gender-based differences in academic outcomes. This finding highlights the need to investigate potential systemic issues or biases contributing to this disparity. The positive coefficient for African (black and coloured) students suggests they are more likely to pass than their counterparts. This result could reflect effective support mechanisms within the institution for these students or other positive factors influencing their academic outcomes. The strong positive coefficient of mainstream indicates that students in mainstream programs are more likely to pass, which could reflect the structured support and resources typically available in such programs.

GLM model reveals a complex interplay of variables affecting student outcomes. Variables like APS, access to information, and program type show expected relationships with academic success, while factors like bursary, effective learning, and social well-being show unexpected negative associations. Gender and ethnicity also emerge as important factors, highlighting the need for educational institutions to consider these demographic variables in their support strategies. The results advocate for a closer look at the institutional context and potentially re-evaluating the support structures in place to ensure they effectively aid in student success.

In conclusion, both models highlight the complex relationship of various variables in influencing student outcomes. While individual abilities and conditions remain crucial, environmental factors and support systems cannot be ignored. Therefore, there is a need to closely examine the institutional context and re-evaluate existing support structures to ensure they effectively contribute to student success by considering demographic variables such as gender and ethnicity alongside relevant factors like access to resources and social well-being to develop more inclusive and tailored approaches to supporting students throughout their academic journey.

5.2.2.4 Undersampling Model

The table 5.10 provides a summary of the Student Expectation - Undersampling Outcome Model, with the outcome of academic success being Pass (1) or Fail (0), reveals various performance measures across different models. These measures include Accuracy, Sensitivity (Recall for the positive class), Specificity, Precision, Recall, and the F1 Score.

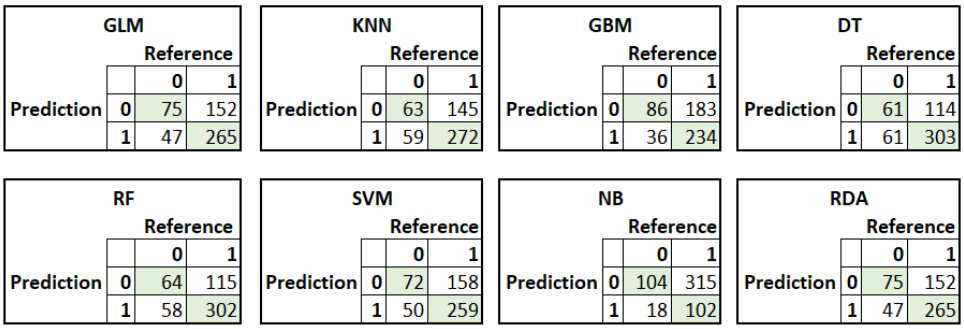
Table 5.10: Undersampling Model Performance Measures for Student Expectation and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.63	0.64	0.62	0.85	0.64	0.73
KNN	0.62	0.65	0.52	0.82	0.65	0.73
GBM	0.59	0.56	**0.71**	**0.87**	0.56	0.68
DT	**0.68**	**0.73**	0.50	0.83	**0.73**	**0.78**
RF	**0.68**	0.72	0.52	0.84	0.72	**0.78**
SVM	0.61	0.65	0.59	0.84	0.65	0.71
NB	0.38	0.25	0.25	0.85	0.25	0.38
RDA	0.63	0.64	0.64	0.85	0.64	0.73

The DT and RF models stand out with the highest accuracy from the performance measures, each scoring 0.68. This implies that approximately 68% of their predictions are correct. The DT model exhibits the highest sensitivity at 0.73, indicating that it correctly identifies 73% of students who will pass. The RF model matches the DT in precision with a score of 0.84, meaning that when it predicts a student will pass, it is correct 84% of the time. Both models also share the highest F1 score of 0.78, suggesting a balanced trade-off between precision and recall. On the other hand, models like the NB show significantly lower performance across most measures, with an accuracy of only 0.38 and a sensitivity of 0.25, indicating a high misclassification rate. Given these findings, the Decision Tree and Random Forest models are strong contenders. Considering the balance of all measures, the RF model is identified as the best performer. It provides high accuracy and a good balance between sensitivity and precision, making it a reliable choice for predicting student outcomes in an undersampled dataset. The ability to correctly identify students who will pass is invaluable for institutions to allocate resources effectively and to support students in achieving academic success.

Confusion Matrix: The confusion matrices for the student expectation models present the performance of eight different machine learning algorithms on an unbalanced dataset

related to academic outcomes, where the positive class '1' represents a pass and the class '0' represents a fail.



GLM		Reference	
		0	1
Prediction	0	75	152
	1	47	265

KNN		Reference	
		0	1
Prediction	0	63	145
	1	59	272

GBM		Reference	
		0	1
Prediction	0	86	183
	1	36	234

DT		Reference	
		0	1
Prediction	0	61	114
	1	61	303

RF		Reference	
		0	1
Prediction	0	64	115
	1	58	302

SVM		Reference	
		0	1
Prediction	0	72	158
	1	50	259

NB		Reference	
		0	1
Prediction	0	104	315
	1	18	102

RDA		Reference	
		0	1
Prediction	0	75	152
	1	47	265

Figure 5.16: Expectation Undersampling Dataset Confusion Matrix

For the GLM model, there were 75 true negatives (students correctly identified as failing) and 265 true positives (students correctly identified as passing). However, there were also 47 false negatives (students who passed but were predicted to fail) and 152 false positives (students who failed but were predicted to pass). The KNN model showed a slight increase in true positives (272) compared to the GLM and a slight decrease in true negatives (63). It had 59 false negatives and 145 false positives.

The GBM model had the highest number of true negatives (86) among the models, indicating it was more conservative in predicting passes. However, it also had the lowest true positives (234), suggesting a tendency to underestimate students' likelihood of passing. The DT model classifier presented a balanced prediction of true positives (303) and true negatives (61), but it also produced a relatively high number of false negatives (61) and false positives (114). The RF model had similar results to the DT, with a high number of true positives (302) but also a significant number of false positives (115) and false negatives (58).

The SVM model produced fewer true positives (259) and true negatives (72) compared to the RF and DT, with 50 false negatives and 158 false positives. The NB classifier had a markedly different distribution, with the highest number of false positives (315) and the lowest number of true positives (102), indicating a bias towards predicting failure. Lastly, the RDA was moderate in its true positive predictions (265), with a comparable number of false positives (152) and false negatives (47) to the GLM.

Each model presents trade-offs between false positives and false negatives, and the choice between them may depend on the cost of misclassification in predicting student outcomes. For example, a model with fewer false negatives might be preferred if the cost of failing to identify a student who needs help is high.

Variable of Importance: The Random Forest model's performance is the most robust across various evaluation metrics, making it the preferred choice for predicting student academic outcomes in this scenario. Its balance of sensitivity, specificity, and precision makes it particularly well-suited for applications where it is essential to correctly identify as many true cases of a 'pass' as possible without a substantial cost for false positives. Therefore, in the context of academic outcomes, where the goal is often to ensure that students who are likely to succeed are correctly identified and supported, the Random Forest model would be the most appropriate choice among the ones considered.

Based on the Random Forest algorithm for the Student Expectation - expectation-undersampling outcome Model with Academic Outcome, the best model presents a hierarchy of variables regarding their importance as measured by the Mean Decrease in Gini impurity. The Admission Points Score (APS) emerges as the most crucial variable, with a score of 10.6, underscoring its strong predictive power for student success (passing, denoted as '1').

Table 5.11: Expectation Undersampling Variable Importance in the Random Forest Model

Importance Variable	**Mean Decrease Gini**
APS (Admission Points Score)	10.6
FA4 (Access to Information)	9.4
FA2 (Social Well-being)	9.4
FA1 (Effective Learning)	9.3
FA3 (Academic Support)	7.3
Female	6.0
Bursary	3.3
Mainstream	3.0
Residence	2.0
African	1.9

Variables related to Access to Information (FA4) and Social Well-being (FA2) follow closely, each with a 9.4 score indicating their significant impact on academic outcomes. Effective Learning (FA1) and Academic Support (FA3) are prominent factors, with scores of 9.3 and 7.3, respectively, suggesting that the learning environment and support structures are key determinants of student performance.

Gender (female) has a notable importance score of 6, reflecting potential gender-based differences in academic outcomes. Financial support (bursary), the type of academic program (Mainstream), living arrangements (residence), and racial background (African)

also contribute to the model's predictions, albeit with less influence than the other factors.

5.3 Student Experience - Predictive Models

This section explores the results of the Student Experience - predictive models. It covers the measures that underpin the efficacy of various prediction models, each tailored to anticipate academic success with increasing precision. Subsequent subsections will explain the results derived from confusion matrices, offering a transparent view of each model's performance in classifying outcomes correctly. Finally, discuss the variables of importance as they form the cornerstone of the predictions, shaping interventions that strengthen student success.

5.3.1 Student Performance - Models

The Student Experience Performance Predictive Model Table 5.12 provides a comparative overview of various data mining models' capabilities to predict student performance, specifically their likelihood of achieving an above-median first-year average mark (denoted as 'Positive' Class: 1).

Table 5.12: Balanced Model Performance Measures for Student Experience and Performance

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.57	0.54	0.60	0.58	0.54	0.56
KNN	0.56	0.55	0.57	0.56	0.55	0.55
GBM	0.57	0.53	0.61	0.57	0.53	0.55
DT	0.58	0.45	0.72	0.61	0.45	0.52
RF	0.54	0.52	0.56	0.54	0.52	0.53
SVM	**0.60**	**0.59**	0.61	0.60	**0.59**	**0.60**
NB	0.57	0.26	**0.88**	**0.69**	0.26	0.38
RDA	0.55	0.52	0.59	0.56	0.52	0.54

The SVM model displays the highest overall accuracy at 0.60, corresponding with its sensitivity (or recall) and precision scores at 0.59, leading to the highest F1 score among the models at 0.60. These measures suggest that SVM is the most balanced model in identifying true positives and minimizing false positives, making it potentially the most reliable model within this specific dataset for predicting student performance.

In contrast, despite having an accuracy of 0.57, the NB model shows a significant disparity between its sensitivity at 0.26 and specificity at 0.88. This suggests that while it is

quite good at predicting students who will not achieve an above-median mark (true negatives), it falls short in accurately identifying those who will (true positives), as evidenced by the lowest F1 score of 0.38.

The NB model offers the highest specificity at 0.88 but has one of the lowest sensitivity scores at 0.26. This indicates a tendency of the NB model to predict that a student will score below the median more often than above it, which could lead to many false negatives.

Other models like GLM, KNN, GBM, RF and RDA present moderate performance across the board, with F1 scores ranging from 0.52 to 0.56. These models demonstrate neither a significant bias towards false positives nor false negatives but do not excel in any particular metric.

While some models show potential, the generally modest performance across all models suggests room for improvement. Future work might involve exploring additional features, fine-tuning model parameters, or employing more advanced modelling techniques to improve predictive accuracy. It is also important to consider that while SVM shows the best performance in this analysis, the choice of the model should be guided by the specific context and the consequences of false predictions.

5.3.1.1 Confusion Matrix

The confusion matrices 5.17 for the Student Expectation Performance Predictive Model provide a detailed comparison of the predictive capabilities of eight different models.

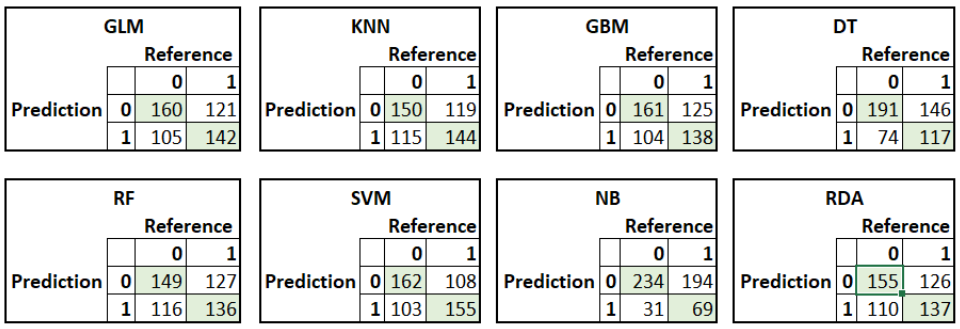


Figure 5.17: Experience Balanced Dataset Confusion Matrix

Starting with the GLM, we see that it predicted 160 students correctly scoring below the median (true negatives) and 142 students correctly scoring above the median (true positives), with 121 false positives and 105 false negatives. The KNN model had a similar performance with slightly more false negatives (115) but fewer false positives (119).

The GBM showed an improved ability to predict true negatives (161) but had a comparable number of false negatives (104) to GLM. The DT model had the highest number

of true negative predictions (191), suggesting a conservative bias towards predicting students below the median and the highest number of false negatives (74). This could lead to missing out on identifying students who perform above the median.

RF and SVM models presented a similar number of true negatives. However, the SVM stood out with a higher number of true positives (155) and fewer false negatives (103), indicating a better balance in prediction accuracy for the 'Positive' class.

The NB model had the highest number of true negatives (234) and false positives (194), indicating a tendency to over-predict the number of students performing below the median. RDA fell in the middle ground, with false positives and negatives close to the average of the other models.

While each model has strengths, the SVM appears to provide the best balance between sensitivity and specificity, making it potentially the most reliable for identifying students likely to score above the median. However, all models show significant misclassifications, indicating room for improvement in the predictive performance.

5.3.1.2 Variable of Importance

Figure 5.18 shows a logistic regression model output for the Student Experience Performance Predictive Model provides insights into factors influencing a student's likelihood of achieving an above-median first-year average mark.

Significant predictors include having a bursary, the APS, and females, all show positive associations with the probability of being in the 'Positive' class. A bursary and a higher APS score increase the likelihood of above-median performance. At the same time, being female is also a strong positive predictor.

Conversely, negative coefficients for FA1 - Effective Learning and African Student suggest that challenges associated with these factors negatively impact the likelihood of achieving above-median marks. The negative relationship with the Mainstream programme indicates that students in the mainstream program are less likely to achieve above-median marks than their counterparts.

The model's accuracy in predicting student performance is indicated by a residual deviance of 1603.0 on 1225 degrees of freedom, suggesting a reasonable fit to the data. However, the presence of significant predictors with both positive and negative associations emphasizes the need for targeted interventions to support student's academic success, particularly for those identified as being at a disadvantage.

```
> summary(EXP_perf_glm_mod)

Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.373  -1.089  -0.696   1.146   1.706

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2485     0.7873   -4.13  3.7e-05 ***
Bursary1      1.0020     0.3056    3.28  0.00104 **
APS           0.0692     0.0129    5.36  8.2e-08 ***
Residence1   -0.1422     0.1702   -0.84  0.40324
FA1          -0.2657     0.1211   -2.19  0.02822 *
FA2           0.1829     0.0920    1.99  0.04690 *
FA3           0.0371     0.0829    0.45  0.65466
FA4           0.2820     0.1309    2.15  0.03117 *
Female1       0.4794     0.1275    3.76  0.00017 ***
African1     -0.6589     0.2662   -2.48  0.01329 *
Mainstream1  -0.3564     0.1492   -2.39  0.01688 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1713.4  on 1235  degrees of freedom
Residual deviance: 1603.0  on 1225  degrees of freedom
AIC: 1625

Number of Fisher Scoring iterations: 4
```

Figure 5.18: Experience Balanced Dataset: Logistic Regression Model

5.3.2 Academic Outcome - Models

This section examines the development and evaluation of predictive models for academic outcomes in Student Experience, specifically focusing on binary classification: Pass (1) and Fail (0). The analysis considers the challenge posed by the imbalanced nature of these categories. Four distinct models are employed: the Unbalanced Model, the ROSE (Random OverSampling Examples) Model, the Oversampling Model, and the Undersampling Model. Various measures are used to assess these models' performance, including Accuracy, Precision, Recall (or Sensitivity), and F1 Score. These measures evaluate how well the models classify student outcomes.

In order to gain deeper insights into model strengths and weaknesses, a confusion matrix is incorporated into the analysis. This matrix visually shows the types and quan-

tities of correct and incorrect predictions made by each model. This approach provides an understanding of how each model addresses challenges associated with unbalanced data. Furthermore, it allows for assessing the effectiveness of ROSE, oversampling, and undersampling in mitigating data imbalance. By adopting this holistic approach, the analysis ensures a comprehension of predictive modelling. It emphasises addressing challenges related to data imbalance while evaluating their impact on model interpretation and performance.

5.3.2.1 Unbalanced Models

The results of the Unbalanced - Outcome Model for the Experience Gap with a 'Positive' class, denoting a pass. The imbalance in the dataset suggests a disproportionate number of instances in one class over another, which is a common challenge in classification problems as it can lead to models that are biased towards the majority class.

Table 5.13: Unbalanced Model Performance Measures for Student Experience and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	**0.78**	**1.00**	0.03	**0.78**	**1.00**	**0.87**
KNN	0.74	0.95	0.03	0.77	0.95	0.85
GBM	**0.78**	**1.00**	0.03	**0.78**	**1.00**	**0.87**
DT	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**
RF	0.77	0.99	0.02	0.77	0.99	**0.87**
SVM	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**
NB	0.77	0.99	0.02	0.77	0.99	**0.87**
RDA	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**

The GLM model Exhibits an accuracy of 0.78, with perfect sensitivity and recall, indicating a strong ability to predict positive outcomes but a specificity of only 0.03, suggesting it is not effective at identifying negative outcomes (Fail=0). The KNN shows a lower accuracy of 0.74 but improved specificity (0.03) compared to GLM and GBM. However, it maintains high sensitivity and recall, although slightly lower than perfect, with values of 0.95. The GBM matches GLM in terms of accuracy, sensitivity, recall, and F1 Score but shares the same low specificity, indicating similar strengths and weaknesses. DT, RF, SVM, NB), and RDA models share an accuracy of 0.77, perfect sensitivity, and recall. However, their specificity is very low to zero, raising concerns about their ability to identify the negative class in an unbalanced dataset correctly.

When dealing with unbalanced datasets, it is essential to look beyond accuracy since high accuracy can be achieved by simply predicting the majority class. Sensitivity and

recall are critical for positive class prediction, but specificity cannot be ignored as it measures the correct identification of the negative class. The F1 Score becomes especially important as it accounts for both precision and recall, providing a more balanced view of a model's performance. Considering the above, the GLM and GBM stand out with the highest accuracy (0.78) and F1 Score (0.87). However, their very low specificity (0.03) is a significant disadvantage. The KNN, with a slightly lower accuracy (0.74) and F1 Score (0.85), presents a slightly more balanced approach with a marginally better specificity.

In selecting the best model, one must weigh the importance of correctly predicting the positive class against the potential cost of misclassifying the negative class. Suppose the priority is to minimise the risk of failing to identify students who will pass (thus potentially missing out on providing necessary support). In that case, GLM or GBM may be preferred despite their low specificity. On the contrary, if it is equally important to avoid incorrectly predicting that a student will pass when they will not, a model with a higher specificity, such as KNN, might be more desirable despite a slight compromise on sensitivity and recall. Hence, with an unbalanced dataset where the cost of false negatives and false positives needs careful consideration, the KNN model may be considered the best model due to its relatively higher specificity while still maintaining commendable sensitivity and recall metrics.

Confusion Matrix: Upon analysing the confusion matrices for the student experience – unbalanced outcome model with the academic outcome, where the 'positive' class is 1 (Pass), several insights into the performance of each model can be deduced.

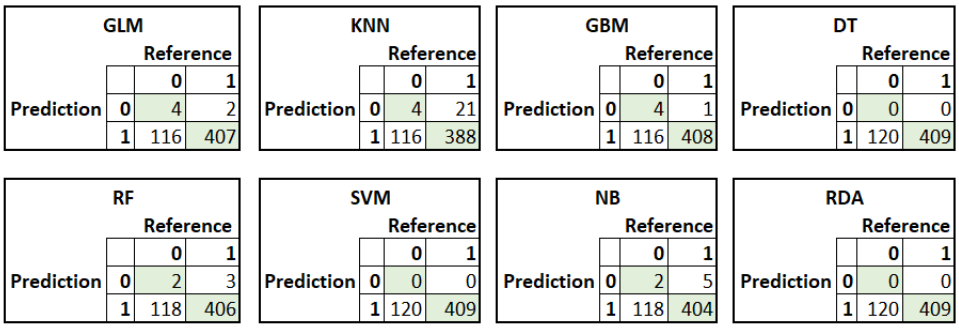


Figure 5.19: Experience Unbalanced Dataset Confusion Matrix

The GLM correctly predicted 407 passes but incorrectly classified 116 failures as passes, suggesting a potential for overestimation of pass outcomes. The KNN algorithm showed a similar true positive rate but with a substantially higher false positive rate, with 21 actual failures being classified as passes. This indicates a sensitivity towards predicting passes, potentially at the expense of precision. The GBM model had an improved balance with

408 true positives and a lower false positive rate than KNN, indicating a better distinction between pass and fail outcomes. The DT model showed the highest number of true positive predictions (409). However, it did not correctly identify any of the true failures, raising concerns about its ability to generalise and identify failures accurately.

Moving to the RF model, it revealed a slightly better performance in correctly identifying failures (2 true negatives). Still, it had a considerable number of failures being misclassified as passes (118 false positives). The SVM model exhibited an extreme bias towards predicting passes, with no failures being correctly identified, similar to the DT model. The NB model demonstrated a slightly better balance in terms of false positives compared to RF but still showed a tendency to misclassify failures as passes. Finally, the RDA model reflected similar results to SVM, with all outcomes predicted as passes, indicating a severe bias and a failure to recognise any of the actual failures.

Considering the balance between sensitivity (true positive rate) and specificity (true negative rate), the GBM model emerges as the most balanced, with a strong true positive rate and a relatively low false positive rate compared to the other models. While the DT and SVM models show the highest number of true positives, their complete inability to predict any true negatives (failures) correctly significantly undermines their effectiveness for this particular outcome prediction. Therefore, the GBM model could be considered the best-performing model due to its ability to maintain a high true positive rate while also preserving a greater accuracy in predicting true negatives compared to other models.

5.3.2.2 Random Over Sampling Examples (ROSE) Models

Table 5.14 shows performance measures for the Student Experience – Random Over Sampling Examples (ROSE) - Outcome Model highlights the predictive capabilities of various data mining techniques to predict academic outcomes where a 'positive' class denotes a pass (1).

The DT model has the highest accuracy (0.68) and F1 score (0.78), indicating a balanced trade-off between the ability to identify true positives and the precision with which predictions are made. Notably, the DT model demonstrates the highest sensitivity (0.75), signifying a superior ability to identify actual passes correctly. However, this is somewhat compromised by a lower specificity (0.44), suggesting a propensity to misclassify actual fails as passes. The RF model follows closely, with a balance across all measures, achieving an F1 score of 0.75. Its sensitivity and specificity measures suggest a more balanced classification capability than the DT model.

The GLM, GBM, and RDA models have identical accuracy and F1 scores, yet the GBM's specificity (0.66) is superior, reflecting a more consistent performance in correctly

Table 5.14: ROSE Model Performance Measures for Student Experience and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.62	0.62	0.60	0.84	0.62	0.71
KNN	0.59	0.60	0.54	0.82	0.60	0.69
GBM	0.61	0.61	0.66	**0.86**	0.61	0.71
DT	**0.68**	**0.75**	0.44	0.82	**0.75**	**0.78**
RF	0.66	0.68	0.57	0.84	0.68	0.75
SVM	0.65	0.66	0.60	0.85	0.66	0.74
NB	0.36	0.20	**0.89**	**0.86**	0.20	0.32
RDA	0.62	0.62	0.61	0.84	0.62	0.71

predicting negative cases (fails). The SVM shows an F1 score of 0.74 with a precision rate (0.85). This indicates a high proportion of true passes among all pass predictions, although its slightly lower specificity compared to RF and GBM suggests a potential for false positives. The KNN algorithm shows moderate performance with a lower specificity (0.54), indicating it may not be as reliable in identifying true fails. The NB model, with the lowest accuracy (0.36) and F1 score (0.32), performs significantly poorer than other models. Despite a high specificity (0.89), its extremely low sensitivity (0.20) reveals a critical deficiency in detecting true passes.

The DT model is determined to be the best model, given its highest accuracy and F1 score. However, it is crucial to consider the low specificity when applying this model, as there may be a higher chance of false positives. The Random Forest (RF) model may be considered a close alternative, offering a more balanced performance across all measures. The ROSE approach is designed to mitigate the effects of class imbalance in the dataset. The evident improvement in the performance measures, particularly sensitivity, suggests that ROSE has been beneficial in enhancing the models' ability to detect the minority class, which in this case is the 'positive' class representing academic success.

Confusion Matrix: The confusion matrices for the Student Experience – Random Over Sampling Examples (ROSE) - Outcome Model presents a comparative view of various classification algorithms applied to an academic outcome with a 'positive' class representing a pass (1).

The GLM shows a relatively balanced distribution of predictions across the confusion matrix, with 61 true negatives and 289 true positives. However, there are a number of false positives (120), indicating that while the model is inclined to predict a pass, it does so at the risk of incorrectly classifying some fails. The KNN algorithm shows

GLM		Reference	
		0	1
Prediction	0	72	155
	1	48	254

KNN		Reference	
		0	1
Prediction	0	65	163
	1	55	246

GBM		Reference	
		0	1
Prediction	0	79	161
	1	41	248

DT		Reference	
		0	1
Prediction	0	53	102
	1	67	307

RF		Reference	
		0	1
Prediction	0	68	130
	1	52	279

SVM		Reference	
		0	1
Prediction	0	72	139
	1	48	270

NB		Reference	
		0	1
Prediction	0	107	327
	1	13	82

RDA		Reference	
		0	1
Prediction	0	73	156
	1	47	253

Figure 5.20: Experience ROSE Dataset Confusion Matrix

higher false positives (77) compared to GLM, suggesting that it may be overfitting to the majority class. With 93 true negatives, it demonstrates a reasonable ability to identify fails. However, the high number of false positives is a concern for over-prediction of passes. The GBM model shows an improvement in predicting true negatives (37) over KNN, which implies better specificity. However, with 83 false positives, the risk of misclassification is still present, although it shows a good number of true positives (303), indicating a strong sensitivity.

The DT model shows a good balance with the highest number of true negatives (55) among the models and a high number of true positives (306). The false positive rate (103) is lower than GLM and KNN, indicating a more balanced approach to classification. The RF model presents a higher number of false positives (72) than true negatives (44), which might be indicative of a bias towards predicting the majority class. However, it also correctly identifies a significant number of true positives (337), showing strong predictive power for passes. The SVM shows true negatives (53) with fewer false positives (108) than KNN and GLM, suggesting a more conservative approach to predicting passes. It maintains a robust number of true positives (301), indicating a well-rounded predictive capability.

The NB model stands out with a high number of true negatives (99), suggesting excellent specificity. However, it has the lowest number of true positives (136), indicating that it is highly conservative and more likely to predict a fail, potentially at the expense of missing true passes. The RDA model has a balanced number of true negatives (61) and a moderate number of false positives (122), suggesting a reasonable specificity. The number of true positives (287) is also commendable, though not the highest among the models.

The DT model appears to offer the best balance between sensitivity and specificity, with a strong ability to predict both true negatives and positives. It has the highest number of true negatives, which is crucial in the context of academic outcomes where

the cost of false positives (incorrectly predicting a pass) can be significant. However, the choice of the best model may also depend on the particular cost function or the specific balance between precision and recall that an educational institution prioritises.

Variable of Importance: The Decision Tree model is the best Student Experience - ROSE (Random Over-Sampling Examples) Outcome Model with Academic Outcome, and it has several variables of importance that contribute to the prediction of academic success (Pass=1). The variables of importance identified by the Decision Tree model are as follows: APS, female, Bursary, FA4 (Access to Information) and FA1 (Effective Learning)

The APS is a critical indicator of academic outcomes. Its importance in the model's variable importance suggests that higher APS scores are likely associated with a higher likelihood of passing. The Decision Tree likely uses this variable to create a threshold that significantly separates the passing and failing students. The inclusion of the gender variable, particularly being female, implies that there may be a gender-based pattern in academic outcomes within the dataset. The model has identified that female students either outperform or underperform compared to male students to a statistically significant degree, which influences the prediction of pass rates. The presence of a bursary can be a proxy for both financial stability and recognition of academic competence, as bursaries are often awarded based on financial need or academic merit. Students with bursaries have different pass rates compared to those without, which could be due to a variety of factors, including socio-economic status, motivation, and available resources for educational support.

Access to information (FA4) is an important factor in student success, as it reflects the ability to obtain necessary academic resources and learning materials. This variable's importance in the model suggests that students who have better access to information tend to pass more frequently than those who do not, possibly due to enhanced learning opportunities and support. Effective learning (FA1) is a qualitative measure of the learning process. Its significance in the model indicates that students who score higher on this variable are more likely to pass, emphasising the importance of effective learning strategies and environments.

It is important to acknowledge the potential biases inherent in the dataset and be cautious of overfitting when interpreting the results. While the model has shown effectiveness in classifying students based on the confusion matrix, it is important to note that the variables identified as significant are contextualised within the broader academic environment and should not be considered causative without further analysis. For instance,

although having a bursary may be associated with passing, it does not necessarily imply a causal relationship. It could be that the qualities leading a student to obtain a bursary are also those contributing to their academic success. Therefore, additional analysis is needed to establish causation.

This model's usefulness lies in its ability to identify key factors associated with academic outcomes. This information can inform targeted interventions and support strategies for improving student performance. For example, the APS is found to be a strong predictor of success. In that case, the university can focus on providing preparatory courses aimed at enhancing APS scores. Similarly, gender is identified as a significant variable, and further investigation into gender-specific academic support systems may be warranted. The findings regarding bursaries, access to information, and effective learning can guide policy-making and resource allocation efforts to strengthen these areas for student success. These findings provide valuable insights that can inform evidence-based interventions and policy decisions aimed at enhancing student success rates.

5.3.2.3 Oversampling Models

The table 5.15 shows the different predictive models of the Student Experience – Oversampling - Outcome Model, which generated a range of performance measures. These models were assessed based on their accuracy, sensitivity, specificity, precision, recall, and F1 score, considering a 'positive' class for students who pass.

Table 5.15: Oversampling Model Performance Measures for Student Experience and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.66	0.71	0.51	0.83	0.71	0.76
KNN	0.68	0.81	0.23	0.78	0.81	0.80
GBM	0.64	0.74	0.31	0.79	0.74	0.76
DT	0.68	0.75	0.46	0.83	0.75	0.79
RF	**0.72**	**0.82**	0.37	0.82	**0.82**	**0.82**
SVM	0.67	0.74	0.44	0.82	0.74	0.78
NB	0.44	0.33	**0.83**	**0.87**	0.33	0.48
RDA	0.66	0.70	0.51	0.83	0.70	0.76

The RF model appears as the greater classifier with the highest accuracy (0.72) and F1 score (0.82). The F1 score, which is a balance between precision and recall, indicates that the RF model is both accurate and reliable. The RF model also shows the highest sensitivity (0.82), suggesting it is particularly proficient at correctly identifying students

who will pass. The KNN model also demonstrates strong performance with an F1 score of 0.80 and impressive sensitivity (0.81) and recall (0.81). However, it is somewhat let down by its lower specificity (0.23). This suggests that while KNN is good at identifying students who will pass, it may also incorrectly predict a pass for students who will fail. The DT and GLM models present similar F1 scores (0.79 and 0.76, respectively), with the DT model showing a slightly better balance between sensitivity and specificity. These models are strong contenders, offering a reasonable trade-off between the various measures.

The SVM and GBM models have comparable F1 scores (0.78 and 0.76, respectively) but differ in their specificity and precision, with the SVM model showing a better ability to predict true passes out of all positive predictions. The NB model shows the highest specificity (0.83) and precision (0.87), suggesting that when it predicts a pass, it is highly likely to be correct. However, its overall performance is hindered by low accuracy (0.44) and an F1 score (0.48), indicating a significant number of false negatives, where it fails to identify students who will pass. The RDA model has moderate measures across the board, with its performance being consistent but not outstanding compared to the other models.

In conclusion, the RF model stands out as the best model, given its superior performance across all measures, particularly in terms of balance between sensitivity and precision. This suggests that the RF model benefits from the oversampling approach, which aims to balance the dataset and improve the model's ability to predict minority class outcomes accurately. This model would be advisable for deployment in predicting academic outcomes where the cost of false negatives and false positives needs to be minimised.

Confusion Matrix: The confusion matrices provided for the Student Experience – Oversampling Outcome Model with Academic Outcome (Pass=1 and Fail=0) show the performance of various predictive models when applied to an imbalanced dataset that has been adjusted using oversampling techniques.

The GLM model shows a reasonable balance between true positives (289) and true negatives (61) but with a significant number of false positives (120), indicating a tendency towards predicting a pass, which may not always be accurate. The KNN model has a higher sensitivity, as seen by the substantial number of true positives (332), but at the cost of a considerable number of false positives (77). KNN seems to be more aggressive in predicting passes but less discerning in its classification of fails. The GBM presents a better balance with fewer false positives (106) compared to KNN, suggesting a more moderate approach. With 303 true positives, it shows a strong ability to predict passes

GLM		Reference	
		0	1
Prediction	0	61	120
	1	59	289

KNN		Reference	
		0	1
Prediction	0	27	77
	1	93	332

GBM		Reference	
		0	1
Prediction	0	37	106
	1	83	303

DT		Reference	
		0	1
Prediction	0	55	103
	1	65	306

RF		Reference	
		0	1
Prediction	0	44	72
	1	76	337

SVM		Reference	
		0	1
Prediction	0	53	108
	1	67	301

NB		Reference	
		0	1
Prediction	0	99	273
	1	21	136

RDA		Reference	
		0	1
Prediction	0	61	122
	1	59	287

Figure 5.21: Experience Oversampling Dataset Confusion Matrix

accurately. The DT model offers a good number of true positives (306) with a moderately high number of true negatives (55), showing a balanced approach. However, it also displays a relatively high number of false positives (103). The RF model has the highest number of true positives (337), indicating a strong predictive power for identifying students who will pass. However, it also has a relatively high number of false positives (72), which suggests it may be likely to to over-predicting passes. The SVM model has a commendable number of true positives (301) and is more conservative with false positives (108) than KNN, indicating a careful balance in predicting passes. The NB model demonstrates a high number of true negatives (99), indicating a strong specificity. However, it has the lowest number of true positives (136), suggesting it may be overly cautious in predicting passes. The RDA model shows a balanced approach with a good number of true positives (287) and a moderate number of true negatives (61). However, like the GLM, it has a considerable number of false positives (122).

In summary, while all models offer varying strengths, the Random Forest model stands out with the highest true positive rate, making it potentially the most suitable for identifying students likely to pass. However, its false positive rate needs to be taken into account depending on the cost associated with incorrectly predicting a pass.

Variable of Importance: Table 5.16 display the variables used by the Random Forest model to predict academic success, along with their associated Mean Decrease Gini, which is a measure of variable importance.

The APS stands out with the highest Mean Decrease Gini score of 29.62, highlighting its critical role as an indicator of academic outcome. A higher APS is strongly associated with a student's likelihood of passing, suggesting that academic preparation before university has a substantial impact on subsequent success. Social well-being (FA2) and Access to Information (FA4) factors are nearly equivalent in importance, with scores of 18.52 and 18.46, respectively. Social well-being summarises the student's social environment

Table 5.16: Experience Oversampling Variable Importance in the Random Forest Model

Variable	Mean Decrease Gini
APS	29.62
FA2 (Social well-being)	18.52
FA4 (Access to Information)	18.46
FA3 (Academic Support)	16.84
Female	15.44
FA1 (Effective Learning)	15.34
Mainstream	9.26
African	6.67
Bursary	6.65
Residence	2.87

and personal circumstances, which evidently play a significant role in academic performance. Access to information, similar to resources and learning materials, is similarly crucial, reinforcing the notion that material support is vital for academic achievement.

Academic Support (FA3) with a Mean Decrease Gini score of 16.84, the level of academic support available to students, such as tutoring and mentoring programs, is identified as a significant predictor of academic outcomes. The model attributes a substantial importance score of 15.44 to being female, indicating that gender may influence academic outcomes. This reflects broader social factors or institutional dynamics. The importance of Effective Learning (FA1) strategies is nearly on par with gender, with a score of 15.34. This suggests that how students learn and the methods they employ are as influential as their social and academic contexts.

Mainstream programmes and bursaries have lower importance scores of 9.26 and 6.65, respectively. Being in a mainstream program as opposed to an extended one may affect a student's progress while having a bursary could relate to both financial security and motivation. Ethnicity (African), with an importance score of 6.67, implies there are differences in the pass rates between African or coloured students and others, which could be due to a variety of complex socio-economic and educational factors. The residence is the least important variable, with a score of 2.87, indicating that living in a university residence has a relatively insignificant direct impact on academic success compared to the other factors listed.

The reliance on these variables highlights the multidimensional nature of academic success. Both inherent student characteristics (like gender and ethnicity) and external

factors (such as academic support and social well-being) are crucial in determining outcomes. These insights are instrumental in designing targeted interventions and support systems to help students succeed.

5.3.2.4 Undersampling Models

Table 5.17 explores the usefulness of various predictive models using an undersampled dataset to address class imbalance. The 'positive' class, representing students predicted to pass (1), was analyzed across multiple models, with performance metrics such as accuracy, sensitivity, specificity, precision, recall, and F1 score serving as evaluation benchmarks.

Table 5.17: Undersampling Model Performance Measures for Student Experience and Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.60	0.61	0.58	0.83	0.61	0.70
KNN	0.61	0.66	0.45	0.80	0.66	0.72
GBM	0.59	0.61	0.52	0.81	0.61	0.70
DT	**0.66**	**0.74**	0.40	0.81	**0.74**	**0.77**
RF	0.62	0.64	0.54	0.83	0.64	**0.77**
SVM	0.61	0.64	0.49	0.81	0.64	0.71
NB	0.37	0.22	**0.88**	**0.86**	0.22	0.35
RDA	0.60	0.60	0.57	0.83	0.60	0.70

The DT and RF models achieve the highest F1 scores (0.77), which suggests a balanced trade-off between precision and recall. The DT model shows a particularly high sensitivity (0.74), indicating a strong ability to identify students who will pass correctly. However, it has a lower specificity (0.40), which means it is more likely to predict passes for students who will fail falsely. The KNN model shows performance with a good balance between sensitivity (0.66) and F1 score (0.72), albeit with lower specificity (0.45), indicating a tendency to over-predict passes. The GLM, GBM, and RDA models display moderate accuracy and F1 scores around 0.70. These models offer a middle ground in terms of sensitivity and specificity.

The SVM model has an F1 score (0.71) slightly higher than that of GLM, GBM, and RDA, with a reasonable balance across all measures. Notably, the NB model has the highest specificity (0.88) and precision (0.86). However, its low sensitivity (0.22) and F1 score (0.35) indicate that while it is accurate in predicting the students who will fail, it struggles to identify those who will pass.

The findings show that the DT and RF models exhibited superior performance, with

F1 scores of 0.77, indicating a robust balance between precision and recall. The DT model, in particular, showed high sensitivity, demonstrating an ability to identify a high number of students who would pass correctly. However, this came at the expense of specificity, where the model was prone to false positives. The RF model, while slightly less sensitive, presented a more balanced profile, suggesting a trade-off between identifying true positives and avoiding false positives.

Confusion Matrix: The confusion matrices provided for the Student Experience – Undersampling Outcome Model with Academic Outcome shows a comparative analysis of how different predictive models perform in classifying students into 'pass' (1) and 'fail' (0) categories.

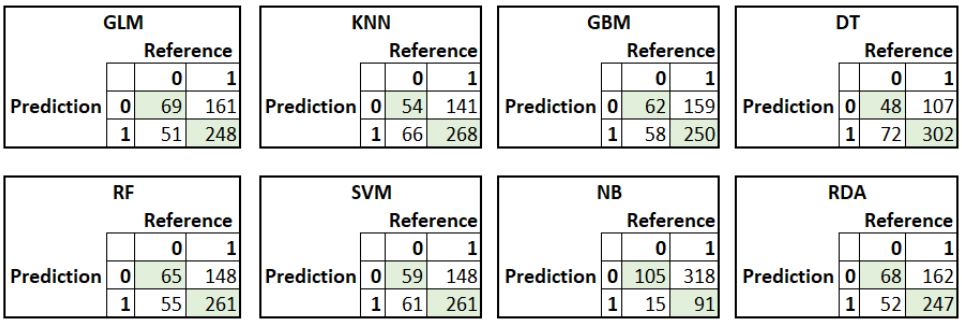


Figure 5.22: Experience Undersampling Dataset Confusion Matrix

The GLM shows a moderate predictive accuracy with a true positive rate of 248. It does, however, show a propensity for false negatives, with 51 students who pass being predicted to fail. Its ability to correctly identify those who fail is slightly better, with 69 true negatives. The KNN model shows a higher true positive rate (268) compared to GLM, suggesting a better sensitivity towards identifying passing students. Nevertheless, it also has a higher false negative count (66), and the number of true negatives (54) is lower than that of the GLM. The GBM offers a balanced approach between identifying true positives (250) and true negatives (62). However, similar to the KNN model, it has a relatively high number of false negatives (58). The DT model shows a high number of true positives (302), indicating strong sensitivity and a lower number of false negatives (72). This model also has the lowest count of true negatives (48), suggesting a lower ability to identify students who will fail the academic outcome correctly.

The RF model closely mirrors the DT model in terms of true positives (261), with a slightly higher true negative rate (65) and fewer false negatives (55). The RF model appears to achieve a better balance between sensitivity and specificity. The SVM model's performance is similar to that of the RF, with an equal number of true positives (261).

It also has a comparable number of true negatives (59) and false negatives (61). The NB model demonstrates the highest specificity with a significant number of true negatives (105). However, it underperforms in identifying true positives (91), resulting in a high number of false negatives (15), indicating a considerable lack of sensitivity. The RDA has a balanced number of true positives (247) but exhibits a lower specificity with a moderate count of true negatives (68).

The DT and RF models are good at identifying students who are likely to pass. However, they do not perform as well when trying to detect students who may not succeed. On the other hand, the Naive Bayes model tends to correctly spot students who might fail but does not predict passing students as effectively. This situation highlights a common issue in predictive modelling: it is challenging to find a model that is equally good at predicting both successes and failures. Therefore, it is important for making decisions to carefully consider which errors of predicting a pass where there is a fail, or vice versa, are more critical to avoid and choose the model that best serves their needs and goals for student outcomes.

Variable of Importance: This part explores the variables of importance for the DT and RF models that have been identified as the best models by focusing on their capacity to explain the factors that most significantly influence whether first-year students pass or fail. Understanding these key variables is critical for developing targeted support strategies that can lead to improved academic success.

The Decision Tree model has highlighted a set of variables as significant predictors of student success, defined as achieving a pass (class 1). The variables of importance, as determined by the model, offer insights into the factors that may influence academic performance. These were the variables: APS, mainstream programmes, Bursary Status, FA2 (Social Well-being), FA1 (Effective Learning) and FA3 (Academic Support). The APS emerges as an important variable, indicating that the scores students achieve upon application hold significant predictive power for their academic success. A higher APS may correlate with better preparedness for university-level studies, suggesting that this metric can be an effective tool for the early identification of students likely to succeed.

The distinction between students enrolled in mainstream programmes versus those in extended programmes is identified as a key factor. The model's focus on this variable suggests that the structure and intensity of the academic programme play a role in student outcomes, potentially reflecting differences in curriculum rigour or student readiness. Whether a student has received a bursary is another variable of importance. This could indicate that financial support impacts academic success, possibly by reducing financial

stress or by serving as a motivational factor, as bursaries are often merit-based.

The model's emphasis on Social Well-being (FA2) highlights the influence of students' social environments on their academic performance. Social well-being can encompass a range of factors, including social support networks, engagement in campus life, and overall mental health, all of which can significantly affect a student's ability to focus on and achieve in their studies. Effective learning (FA1) practices, as denoted by FA1, are highlighted as predictors of success. This suggests that the model recognises the importance of how students approach learning, including their study habits, time management, and the usage of learning resources. The inclusion of academic support (FA3) reflects the model's valuation of institutional support mechanisms such as tutoring services, mentorship programmes, and academic advising. The prominence of this variable indicates that the support students receive from the institution can be a determinant of their academic outcomes.

The Decision Tree model's identification of these variables indicates a relationship between individual student characteristics, such as their APS and effective learning strategies, and institutional factors, like the type of academic programme and support provided. This understanding can help in developing comprehensive strategies that not only enhance student preparedness at the point of entry but also provide ongoing support tailored to students' social and academic needs.

Turning to the Random Forest model, its ensemble approach typically corroborates the importance of these variables while potentially offering additional robustness against overfitting. The Random Forest model is likely to validate the significance of these predictors and may provide a more generalised predictive power across diverse student populations. Its ability to handle a large number of input variables and to model complex interactions makes it a powerful tool for capturing the multifaceted nature of academic success.

The APS, with the highest Mean Decrease Gini score, is the most influential predictor. This highlights the important role of students' scores upon application in predicting their academic performance, suggesting that initial academic preparedness is a strong indicator of future success. Social well-being (FA2) and academic support (FA3) are also key factors with importance scores. This highlights the multifaceted interplay between students' social context and the academic support they receive and how these factors collectively impact academic outcomes. The access students have to information and resources (FA4) was the next in importance, reinforcing the idea that having the right tools and information is crucial for academic achievement.

Effective Learning (FA1) indicates that the strategies and habits students use in their learning process are essential predictors of their academic performance. Mainstream Pro-

Table 5.18: Experience Undersampling Variable Importance in the Random Forest Model

Variable	Mean Decrease Gini
APS (Admission Points Score)	11.38
FA2 (Social Well-being)	8.54
FA3 (Academic Support)	8.51
FA4 (Access to Information)	7.33
FA1 (Effective Learning)	7.07
Mainstream	3.31
African	3.17
Female	2.68
Bursary	1.97
Residence	1.067

gramme, African Ethnicity, Gender, Bursary, and Residence have lower importance scores but are still significant. Participation in a mainstream programme versus an extended one, a student's ethnicity and gender, whether they have financial aid, and their living situation all provide additional context to a student's likelihood of passing. However, their influence is less than the variables above.

The Random Forest model's identification of these variables provides actionable insights. For instance, institutions may focus on strengthening initial academic readiness as indicated by the APS and provide targeted support to enhance social well-being and effective learning practices. Additionally, understanding the role of demographics such as ethnicity and gender can inform the development of tailored support systems to address the unique challenges faced by different student groups.

The insights gathered from the Decision Tree and Random Forest models highlight the relationship between individual student attributes and institutional factors in determining academic success. Both models concur on the importance of Admission Points Score as a key predictor, suggesting that initial academic readiness is paramount. This alignment indicates a clear path to not only enhance student preparedness from the beginning but also to continuously support their journey with resources that address both their social and academic needs. The models also highlight the need for tailored strategies that consider demographics such as ethnicity and gender, thereby ensuring that support systems are sensitive to the diverse challenges of the student body. These findings offer a strategic framework that can guide the university in creating a more personalised and effective educational experience that fosters student success across various backgrounds.

5.4 Expectation & Experience Gap - Models

This section explores the results of the Student Expectation and Experience Gap - predictive models. It covers the measures that underpin the efficacy of various prediction models, each tailored to anticipate academic success with increasing precision. Subsequent subsections will explain the results derived from confusion matrices, offering a transparent view of each model's performance in classifying outcomes correctly. Finally, a discussion is presented on the variables of importance as they form the cornerstone of the predictions, shaping interventions that strengthen student success.

5.4.1 Student Performance - Models

The table 5.19 below contains a Performance Dataset model and its corresponding measures for Accuracy, Sensitivity, Specificity, Precision, Recall, and F1 Score.

Table 5.19: Balanced Model Performance Measures for Gap and Student Performance

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	**0.58**	0.51	0.65	0.59	0.51	**0.55**
KNN	0.55	**0.56**	0.55	0.55	**0.56**	**0.55**
GBM	**0.58**	0.50	0.65	0.59	0.50	0.54
DT	0.57	0.48	0.66	0.58	0.48	0.52
RF	0.57	0.50	0.64	0.58	0.50	0.53
SVM	0.57	0.53	0.60	0.57	0.53	**0.55**
NB	**0.58**	0.26	**0.89**	**0.71**	0.26	0.38
RDA	**0.58**	0.51	0.65	0.59	0.51	0.54

The Balanced-Performance Dataset provides a comprehensive overview of various predictive models' performance distinguishing students above and below the median first-year average mark.

The accuracy of these models hovers around the moderate range of 0.55 to 0.58, with GLM, GBM, NB, and RDA achieving the highest accuracy at 0.58. However, Naive Bayes notably falls behind in Sensitivity (0.26), indicating many false negatives. In contrast, the Specificity of NB is the highest (0.89), suggesting it is best at identifying true negatives. Precision varies less dramatically among the models, with NB leading (0.71), potentially indicating many true positive predictions relative to false positives. The F1 Score, which balances Precision and Recall, is relatively consistent across models, with most scoring around 0.55. This suggests none of the models are particularly strong in

balancing Precision and Recall.

Given these measures, the Generalized Linear Model (GLM) and the Regularized Discriminant Analysis (RDA) emerge as the top performers, providing a balanced performance across all metrics. However, when choosing the best model, one should consider the specific application and the costs associated with false positives and false negatives. For balanced performance in both identifying students who are above and below the median first-year average mark, GLM and RDA would be recommended based on this dataset.

5.4.1.1 Confusion Matrix

The confusion matrices 5.23 for the various Expectation and Experience Gap Performance Predictive Models provide a detailed view of each model's ability to classify students' first-year average marks as either above or below the median.

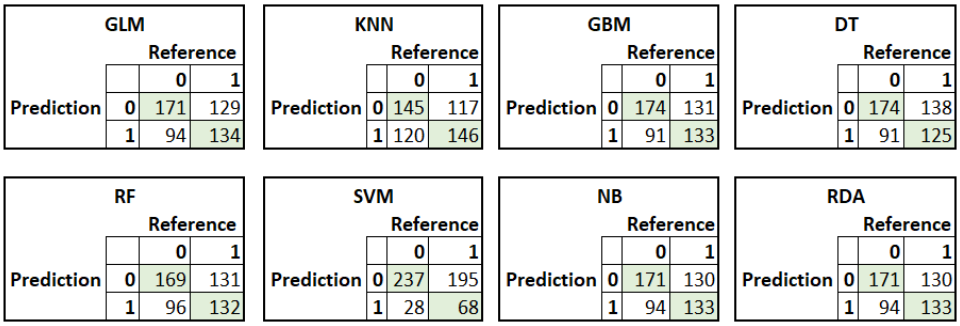


Figure 5.23: Gap Balanced Dataset Confusion Matrix

The Generalized Linear Model (GLM) showed a balanced capability in prediction with 171 true negatives and 134 true positives. This balance suggests a good trade-off between sensitivity (ability to identify positive cases) and specificity (ability to identify negative cases), with the model correctly identifying many students in each category. However, the presence of 129 false positives and 94 false negatives indicates potential areas for improvement.

The K-Nearest Neighbors (KNN) model tended to have higher sensitivity, with 146 true positives. However, it also had a relatively high number of false negatives (120) and false positives (117), indicating that the model may benefit from parameter tuning to better differentiate between the classes.

The Gradient Boosting Machine (GBM) and the Decision Tree (DT) models demonstrated a strong ability to identify true negatives (174 each). Still, they also presented many false positives (131 for GBM and 138 for DT). This might reflect a propensity of these models to favour the negative class in their predictions.

Random Forest (RF) had similar performance characteristics to the GBM and DT models, with a substantial number of true negatives (169) and true positives (132). Yet, it also suffered from a high rate of false positives (131), indicating a possible overfitting issue or a need for model calibration.

The Support Vector Machine (SVM) model displayed an excellent true negative rate (237), the highest among all models, but at the expense of a high false positive rate (195). This suggests that while SVM effectively identifies students who are not above the median, it struggles to identify those who are accurately.

Naive Bayes (NB) and Regularized Discriminant Analysis (RDA) showed a moderate balance with 171 true negatives and 133 true positives, similar to the GLM model. However, like GLM, they also had a notable number of false positives and false negatives, suggesting that while they are relatively balanced, they could be improved for precision and recall.

In summary, each model exhibits unique strengths and limitations in predicting students' academic performance. Models like GLM, NB, and RDA provide a balanced approach to classification, while others like SVM may be more conservative, minimizing false negatives but increasing false positives. The choice of model should be guided by the specific application needs, considering the implications of false positives and negatives in the context of academic performance prediction.

5.4.1.2 Variables of importance

This section exclusively focuses on the models chosen as the best models based on their performance. Specifically, the Generalized Linear Model (GLM) was selected to analyse the important variable.

- **Logistic Regression or Generalized Linear Model (GLM)**

  Figure 5.24 shows a summary output of the Expectation and Experience Gap Performance Predictive Models that predict whether a student's first-year average mark is above (1) or below (0) the median; the results highlight significant factors correlating with academic performance outcomes.

  The model's coefficients indicate the extent to which each predictor variable is expected to impact the log odds of a student being in the 'Positive' class (Above Median First Year Average Mark). The intercept, at -2.4306, sets a baseline for the log odds of achieving above-median marks without all other factors.

  Bursary1, with an estimate of 1.0021, suggests that having a bursary is positively associated with above-median academic performance, with a significant z-value of

```
> summary(GAP_perf_glm_mod)

Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.353  -1.082  -0.659   1.145   1.669

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.4306     0.5918   -4.11  4.0e-05 ***
Bursary1      1.0021     0.3061    3.27  0.00106 **
APS           0.0708     0.0129    5.47  4.5e-08 ***
Residence1   -0.1205     0.1710   -0.70  0.48091
FA1          -0.1941     0.1003   -1.93  0.05300.
FA2           0.1125     0.0769    1.46  0.14355
FA3          -0.0747     0.0720   -1.04  0.29927
FA4           0.2442     0.0988    2.47  0.01343 *
Female        0.4561     0.1273    3.58  0.00034 ***
African1     -0.6330     0.2678   -2.36  0.01807 *
Mainstream1  -0.3577     0.1492   -2.40  0.01650 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1713.4  on 1235  degrees of freedom
Residual deviance: 1604.6  on 1225  degrees of freedom
AIC: 1627

Number of Fisher Scoring iterations: 4
```

Figure 5.24: Expectation and Experience Gap Balanced Dataset: Logistic Regression Model

3.27 and a p-value of 0.00106, indicating strong evidence against the null hypothesis of no effect.

The Admission Points Score (APS) variable shows a positive relationship with a student's likelihood of achieving above-median grades, with an estimate of 0.0708. The high z-value of 5.47 and a very low p-value (4.5e-08) affirm its predictive strength.

Variables associated with student support and environment, such as Effective Learning (FA1), Social Well-being (FA2), Academic Support (FA3), and Access to Information (FA4), have varying degrees of influence in the gap between expectation and experience. FA4, in particular, shows a positive effect (estimate 0.2442) with statistical significance, indicating the importance of information accessibility in student performance.

The Female variable coefficient (0.4561) is positively associated with the 'Positive' class, suggesting female students are more likely to have above-median grades, as supported by a p-value of 0.00034. Conversely, the African1 variable coefficient (-0.6330) suggests that African students are less likely to have above-median grades, which is statistically significant and warrants further investigation into potential underlying causes.

The Mainstream1 variable, indicating enrollment in the mainstream program, has a negative association with above-median performance, with a coefficient of -0.3577, indicating that students in the mainstream program might be less likely to achieve above-median marks than their counterparts.

The model's deviance residuals range from -2.353 to 1.669, with median residuals closer to zero, which suggests that the model fits moderately well for a significant number of observations. However, the presence of relatively large residuals indicates the potential for model improvement or the existence of outliers.

With a residual deviance of 1604.6 on 1225 degrees of freedom, the model seems to fit the data better than the null model, evident from the lower AIC value of 1627. The number of Fisher Scoring iterations is 4, indicating that the algorithm required a few iterations to converge, which is typical for well-specified GLM models.

In summary, this GLM output indicates that financial support, APS, access to information, and gender are significant predictors of student academic performance. The negative coefficients for African1 and Mainstream1 suggest areas for potential policy intervention and further research to understand the disparities in academic outcomes.

- Gradient Boosting Machine (GBM)

  Table 5.20 shows the relative importance of various variables in predicting whether a student's first-year average mark is above the median (classified as 'Positive' with a value of 1) in the context of a Balanced-Performance Predictive Model.

  At the forefront, the Admission Points Score (APS) is the most influential variable with a relative importance of 38.86%, indicating that it strongly predicts a student's likelihood to have an above-median first-year average mark.

  Following APS, the aspects of Effective Learning (FA1), Social well-being (FA2), and Academic Support (FA3) emerge as significant factors with the relative importance of 14.06%, 12.04%, and 8.34%, respectively. These elements underscore the importance of learning environments, social contexts, and support structures in

influencing academic performance.

Table 5.20: GBM Variable Relevance Information

Variable	Mean Decrease Gini
APS (Admission Points Score)	38.86
FA1 (Effective Learning)	14.06
FA2 (Social Well-being)	12.04
Female	9.93
FA4 (Access to Information)	8.85
FA3 (Academic Support)	8.34
Bursary	3.12
African	2.51
Mainstream	1.42
Residence	0.87

Access to Information (FA4) also plays a notable role (8.85%), suggesting that obtaining necessary academic resources is crucial for student success.

The presence of a bursary holds some predictive power (3.12%), which could imply financial security's impact on academic outcomes.

Being a female student, an African student, enrolled in a Mainstream programme, and residing in student accommodation show varying degrees of influence on academic performance, with the relative importance of 9.93%, 2.51%, 1.42%, and 0.87%, respectively. These demographic and situational variables offer additional insights, although they have less predictive strength than academic-related factors.

While demographic factors contribute to the model, the primary indicators of a student's likelihood to score above the median in their first-year average mark are predominantly related to their academic engagement and resources, with APS being the most significant predictor.

5.4.2 Academic Outcome - Measures

5.4.2.1 Unbalanced Models

This section provides a statistical summary of the performance of various models on the Expectation and Experience Gap with an Unbalanced Outcome in the context of Academic Outcomes. The outcome is binary, with a 'Pass' represented by 1 and a 'Fail' by 0; the 'Positive' class is the 'Pass' outcome.

Table 5.21: Unbalanced Model Performance Measures for Gap and Academic Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	**0.78**	**1.00**	0.03	**0.78**	**1.00**	**0.87**
KNN	0.74	0.95	0.03	0.77	0.95	0.85
GBM	0.77	**1.00**	0.03	**0.78**	**1.00**	**0.87**
DT	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**
RF	0.77	0.99	0.02	0.77	0.99	**0.87**
SVM	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**
NB	0.77	0.99	0.02	0.77	0.99	**0.87**
RDA	0.77	**1.00**	0.00	0.77	**1.00**	**0.87**

Table 5.21 summarises the performance of each model based on various measures. The GLM has the highest accuracy (0.78), indicating it is the most reliable model overall for predicting the outcomes correctly. Except for the KNN, all models reveal perfect sensitivity (1.00), indicating they correctly identify students who pass all the time. However, this also suggests a potential issue with overfitting to the 'Pass' class. All models show low specificity, with DT, SVM, and RDA showing no ability to correctly identify 'Fail' outcomes. This points to a significant issue where the models rarely identify students who fail correctly. Precision indicates the proportion of positive identifications that were correct. Here, the models are consistent, with GLM and Gradient Boosting Machine (GBM) leading slightly (0.78). This suggests that when the models predict a student will pass, they are correct approximately 78% of the time. All models score high on the F1 Score (0.85 to 0.87), with GLM and GBM having the highest (0.87), suggesting a good balance between precision and recall.

The GLM is the best model, given its highest accuracy and F1 Score. However, it is essential to note that the specificity of all models is low, and thus, the models are unreliable for predicting the 'Fail' outcome. While GLM is the best among the compared models, it still struggles with imbalanced classification. It might benefit from techniques to handle unbalanced data, such as resampling or specialised loss functions.

In summary, while the GLM model may be selected as the best model for predicting passes based on the available metrics, caution should be exercised due to the overall low specificity across all models, which could result in a significant number of false positives—predicting a pass when the actual outcome is a fail. This could be problematic in an academic setting where incorrectly assuming a student will pass without intervention might lead to neglecting those at risk of failing. Further investigation into the models'

performance on a more balanced dataset or the application of class-weight adjustments might be warranted to improve specificity.

Confusion Matrix: The confusion matrices presented for the Expectation and Experience Gap – Unbalanced Outcome Model with Academic Outcome provide valuable insights into the performance of each model. The 'Positive' class here is '1', representing a 'Pass'.

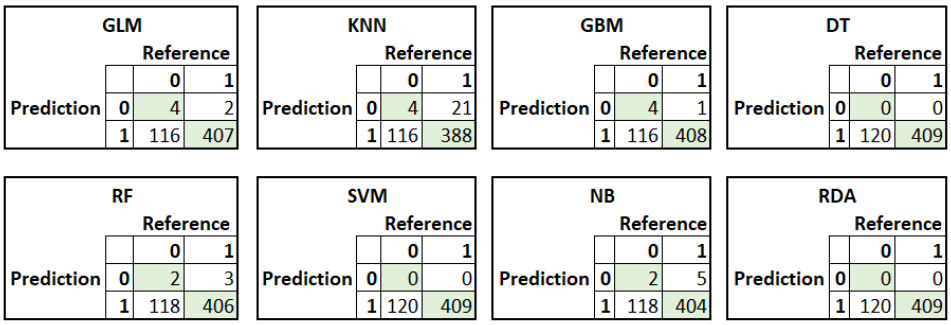


Figure 5.25: Gap Unbalanced Dataset Confusion Matrix.

For the GLM, it predicted 'Pass' (1) correctly 407 times and incorrectly predicted 'Fail' (0) as 'Pass' 116 times. It rarely predicted 'Fail', with four correct predictions and two false negatives. This model has high sensitivity but very low specificity. The KNN model had fewer true positives at 388, indicating it was less successful at identifying 'Pass' outcomes than GLM. It also incorrectly predicted 'Fail' as 'Pass' 116 times. Notably, it has 21 false negatives, suggesting it is more cautious about predicting 'Pass', which results in missing some actual passes. GBM showed a similar pattern to GLM, with many true positives (408) and false positives (116). However, with only one false negative, it indicates a strong bias towards predicting 'Pass'.

The DT model predicted every student would 'Pass', resulting in many true positives (409) and false positives (120). However, it failed to predict any 'Fail' outcomes, indicating an extreme bias towards the positive class and a lack of specificity. For the RF model, there were 406 true positives and 118 false positives, with a slightly better balance of two true and three false negatives, showing a marginal improvement in specificity over DT. The SVM model, like DT, predicted a 'Pass' for all students, which indicates it has the same issue as DT with an extreme bias towards the positive class and no specificity.

The Naive Bayes (NB) model predicted 404 true positives and 118 false positives, with two true negatives and five false negatives. This suggests it has a slightly better balance than GLM, GBM, and DT but still leans heavily towards predicting 'Pass'. Lastly, the RDA, much like DT and SVM, predicted that all students would pass, again indicating

no ability to identify 'Fail' outcomes, reflecting an overly optimistic model.

All models strongly prefer predicting 'Pass', with GLM and GBM showing the best balance between sensitivity and specificity, despite the latter being quite low. DT, SVM, and RDA lack specificity, undermining their usefulness in practical scenarios. The best model would ideally be one that maintains high sensitivity but improves specificity. However, based on these matrices, if one had to choose, the GLM or GBM might be the preferred models due to their higher true positive rates despite their low true negative rates. Additional measures are needed to improve the predictive performance of these models on the negative class.

5.4.2.2 Random Over Sampling Examples (ROSE) Models

Table 5.22 compares various classification models using the Random Over Sampling Examples (ROSE) - Outcome Model, where the 'positive' class is designated as '1' (Pass). The accuracy, sensitivity (also known as recall), specificity, precision, and F1 score are measures used to evaluate the performance of these models.

Table 5.22: ROSE Model Performance Measures for GAP and Academic Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.62	0.62	0.60	0.84	0.62	0.71
KNN	0.59	0.60	0.54	0.82	0.60	0.69
GBM	0.62	0.61	0.66	**0.86**	0.61	0.71
DT	**0.68**	**0.75**	0.44	0.82	**0.75**	**0.78**
RF	0.66	0.68	0.57	0.84	0.68	0.75
SVM	0.65	0.66	0.60	0.85	0.66	0.74
NB	0.36	0.20	**0.89**	**0.86**	0.20	0.33
RDA	0.62	0.62	0.61	0.84	0.62	0.71

Generalised Linear Model (GLM) and Radial Discriminant Analysis (RDA) demonstrate equivalent performance across all metrics, with an accuracy and sensitivity of 0.62 and an F1 score of 0.71, suggesting a balanced trade-off between precision and recall. Notably, the Decision Tree (DT) classifier shows the highest sensitivity of 0.75 and an F1 score of 0.78, indicating a relatively better performance in correctly identifying the positive class. However, its specificity is the lowest at 0.44, which may indicate a higher rate of false positives.

The RF and SVM models exhibit similar patterns in their metric scores, with RF slightly outperforming SVM in sensitivity and F1 scores. Conversely, the NB classifier

significantly underperforms in sensitivity at 0.2 despite a high specificity of 0.89. This suggests that while it effectively identifies true negatives, it fails to identify true positives reliably, resulting in a markedly low F1 score of 0.33. KNN and GBM models have comparable outcomes, with KNN having marginally lower scores in most measures except for precision.

The DT is the best model because it correctly identifying passes is dominant despite a higher false-positive rate risk. The low performance of NB highlights the challenges of handling unbalanced data. The choice of model would ultimately depend on the specific context and the cost associated with misclassifications of either class.

Confusion Matrix: The results presented in the confusion matrices for the various models provide a comprehensive insight into their classification performance in predicting academic outcomes in the context of the Experience and Expectation Gap - Random Over Sampling Examples (ROSE) Outcome Model.

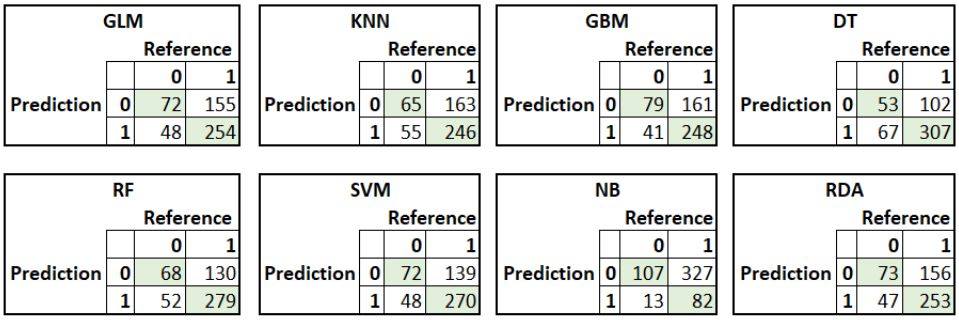


Figure 5.26: Gap ROSE Dataset Confusion Matrix.

The GLM model shows a relatively balanced number of true positives and false negatives. With 254 true positives, the model is quite sensitive to the positive class, but with 155 false positives, it suggests a modest precision. This balance may make the GLM suitable for scenarios where both classes are of similar predictive importance. The KNN model has fewer true positives (246) than the GLM and a slightly higher false positive rate (163). This indicates that KNN may be less effective than GLM in this context, potentially due to overfitting or sensitivity to the imbalanced data.

The GBM model balances sensitivity and specificity better, with fewer false positives (161) and more true positives (248). This might suggest that GBM is effectively leveraging the patterns in the data, possibly due to its iterative approach to minimising errors. The DT model has the highest number of true positives (307), indicating a high sensitivity, but also has a significant number of false negatives (67). This could be due to the decision tree potentially overfitting to the positive class or the model being more tuned to identify

the pass outcomes. With a true positive count of 279 and false negatives at 52, the RF model appears to strike a reasonable balance, suggesting that the ensemble method might capture the underlying patterns in the data more effectively than single decision trees.

The SVM model shows a similar pattern to RF, with a true positive rate of 270 and false negatives at 48. This indicates that SVM may be well-suited for this classification task, possibly due to its capacity to handle high-dimensional data. The NB model exhibits a stark contrast with a high false negative rate (327) and a low true positive rate (82), indicating a high specificity but very low sensitivity. This could imply that NB is overly conservative, potentially missing out on identifying many students likely to pass. RDA shows a moderate number of true positives (253) against the false negatives (47), suggesting that it has a balanced sensitivity and may serve as a reasonable model for predicting academic outcomes.

The models exhibit various behaviours, with some favouring sensitivity (e.g., DT) and others specificity (e.g., NB). An ideal model in an educational context should have a high true positive rate to correctly identify students who will pass, coupled with a low false negative rate to avoid missing students at risk of failing. However, one must also consider the false positive rate as it could lead to the misallocation of resources to students not actually in need. Thus, a model like RF or SVM might be preferable, as they exhibit a good balance between true positives and false negatives, indicating a more reliable performance across both classes. The models' tendencies to predict true positives over true negatives or vice versa highlight the inherent trade-offs between sensitivity and specificity. These trade-offs must be carefully considered in academic outcome predictions, where the cost of a false negative (failing to identify a student at risk of failing) may be more significant than a false positive. The right balance would ensure that at-risk students receive the necessary support without overburdening resources on those who may not need it.

Variable of Importance: The selection of the Random Forest and Decision Tree models for identifying important variables is well-founded. Both models offer distinct advantages and can provide valuable insights into the factors most predictive of academic success. In the Experience and Expectation Gap - Random Over Sampling Examples (ROSE) - Outcome Model, the objective is to determine the likelihood of academic success (Pass=1) or failure (Fail=0).

With its high sensitivity, the DT model is particularly proficient at identifying students at risk of failing (the positive class). This characteristic makes it a powerful tool for prioritising intervention and support for struggling students. The DT model's structure

also allows for easy interpretation of the results, providing a clear and hierarchical representation of how different variables contribute to the predictive outcome. The variables of importance for the model include the Admission Points Score (APS), gender (Female), the provision of a bursary, FA4 (Access to Information) and FA1 (Effective Learning). This section will explore the significance of each variable, offering an understanding of their influence on the model's predictive capacity.

Given its numerical nature and direct correlation with academic history, the APS is a critical predictor in the Decision Tree model. The APS captures prior educational performance, thus providing a robust foundation for forecasting future academic success. The inclusion of gender, particularly the variable 'Female,' in the model indicates that female students have distinctive outcomes compared to their male counterparts, possibly due to socio-economic, behavioural, or institutional factors unique to this demographic in the studied academic environment.

The financial aid variable 'Bursary' implies the economic aspect of a student's educational journey. Access to a bursary could alleviate financial stress, potentially contributing to a student's ability to succeed academically by enabling better access to resources or reducing the need for part-time work. The variable FA4 (Access to Information) suggests the influence of information accessibility on academic performance. It reflects the availability and quality of academic resources, such as libraries, internet access, and academic counselling, which are vital for a student's learning and research activities. Effective learning practices, encapsulated by FA1 (Effective Learning), play a pivotal role in a student's academic success. This variable encompasses the methods and environments that facilitate efficient learning, including study groups, teaching quality, and curriculum design.

The Decision Tree model's variable importance analysis for the ROSE - Outcome Model has highlighted the complex nature of academic outcome predictors. The APS emerged as a foundation variable, directly linking historical academic performance. The gender-specific variable 'Female' and a 'Bursary' presence offer insights into the socio-economic factors at play. Meanwhile, FA4 and FA1 highlight the critical role of institutional support and effective learning strategies. Collectively, these variables reflect the complexity of academic achievement and guide targeted interventions to foster educational success.

On the other hand, the RF model offers a more robust performance by aggregating the predictions of multiple decision trees. This ensemble approach improves overall predictive accuracy and control over-fitting, a common pitfall in singular decision trees. While the RF model is more complex and may not provide as straightforward an interpretation as

DT, it often ranks variables by their importance across all the trees in the forest, offering a comprehensive overview of the factors influencing the model's predictions.

Table 5.23: Gap ROSE Variable Importance in the Random Forest Model

Importance Variable	Mean Decrease in Gini
APS (Admission Points Score)	22.5
FA2 (Social Well-being)	16.2
FA4 (Access to Information)	16.0
FA1 (Effective Learning)	16.0
FA3 (Academic Support)	13.6
Female	9.4
Bursary	5.6
African	5.5
Mainstream	4.6
Residence	2.0

The Admission Points Score (APS) emerges as the most significant variable, with the highest mean decrease in Gini (22.5), highlighting its pivotal role in predicting academic outcomes. A high APS typically indicates strong prior academic performance, which aligns with better academic achievement prospects. Variables relating to academic support and learning environment: FA4 (Access to Information), FA2 (Social Well-being), FA3 (Academic Support), and FA1 (Effective Learning; all show substantial importance with a mean decrease in Gini scores ranging from 13.6 to 16.2. These factors are integral to a student's academic journey, as they represent the resources available, the support they receive, and the efficacy of their learning strategies.

The gender(female) has a moderately high importance (9.4), suggesting gender-specific trends within the data that may reflect different educational experiences or outcomes between female and male students. A Bursary and African play a role, albeit to a lesser extent (5.5 and 5.6, respectively). These variables represent underlying socio-economic conditions that can impact a student's academic journey. The Mainstream program variable (4.6) and Residence (2) impact the model's predictions the least. While still relevant, these variables may not be as strong predictors of academic success as the others or their effects may be mediated through interactions with other more dominant variables.

The RF model provides a perspective on the factors influencing academic outcomes. The APS stands out as a crucial indicator, while support mechanisms and learning conditions also play significant roles. Understanding these variables' importance helps tailor

interventions and policies to enhance support systems, aiming to maximise student success and minimise failure rates.

For both models, the variables of importance likely highlight key indicators of academic outcome. Variables such as the APS emerge as significant, and other factors like gender, bursaries, access to academic resources and effective learning environments (FA4 and FA1) may also show strong predictive power. Where the aim is to identify and support students who may fail accurately, the DT model's sensitivity to the positive class is crucial. However, the RF model's balanced approach might be more appropriate for efficiently allocating resources and avoiding unnecessary interventions. When deploying these models, it is recommended to continuously monitor their performance and recalibrate as necessary, considering changes in the student population and academic environment.

5.4.2.3 Oversampling Models

The performance metrics table 5.24 evaluates several predictive models applied to the Expectation and Experience Gap – Oversampling Outcome Model. The academic outcomes are binary, with 'Pass' coded as 1 and 'Fail' as 0, focusing on the 'Positive' class as 1.

Table 5.24: Oversampling Model Performance Measures for GAP and Academic Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.66	0.71	0.51	0.83	0.71	0.76
KNN	0.68	0.81	0.23	0.78	0.81	0.80
GBM	0.64	0.74	0.31	0.79	0.74	0.76
DT	0.68	0.75	0.46	0.83	0.75	0.79
RF	**0.72**	**0.82**	0.37	0.82	**0.82**	**0.82**
SVM	0.67	0.74	0.44	0.82	0.74	0.78
NB	0.44	0.33	**0.83**	**0.87**	0.33	0.48
RDA	0.66	0.70	0.51	0.83	0.70	0.76

The RF model emerges as the most effective, exhibiting the highest accuracy (0.72) and F1 score (0.82) among all models. Its sensitivity (recall) is also the highest at 0.82, indicating its proficiency in correctly identifying students likely to pass. Although not the highest, specificity is acceptable at 0.37, suggesting a reasonable rate of correctly identifying those who will not pass. The KNN model also performs admirably with an F1 score of 0.80 and the best sensitivity of 0.81. However, its specificity is the lowest at 0.23, which points to many false positives, indicating that while it is good at identifying students who will pass, it may also incorrectly predict passing for students who fail.

The DT model shows balanced performance with a good F1 score of 0.79, but it does not excel compared to the RF model. The NB model has the lowest accuracy (0.44), sensitivity (0.33), and F1 score (0.48), making it the least suitable model for this dataset, despite having the highest precision (0.87). Models like the GLM, GBM, and RDA show moderate performance, with F1 scores ranging from 0.76 to 0.76. These models strike a balance between sensitivity and specificity but do not reach the effectiveness of the RF model.

In summary, based on these metrics, the RF model stands out as the best model for predicting academic outcomes in the context of the ROSE Outcome Model. Its superior balance of accuracy, sensitivity, specificity, and F1 score suggests that it can most effectively use the oversampled data to predict which students will pass or fail, thus guiding interventions to support at-risk students.

Confusion Matrix The confusion matrices presented for each model provide a detailed view of their performance in the context of the Expectation and Experience Gap – Oversampling Outcome Model, aimed at predicting academic success or failure, with success labelled as the positive class (1).

GLM		Reference	
		0	1
Prediction	0	61	120
	1	59	289

KNN		Reference	
		0	1
Prediction	0	27	77
	1	93	332

GBM		Reference	
		0	1
Prediction	0	37	106
	1	83	303

DT		Reference	
		0	1
Prediction	0	55	103
	1	65	306

RF		Reference	
		0	1
Prediction	0	44	72
	1	76	337

SVM		Reference	
		0	1
Prediction	0	53	108
	1	67	301

NB		Reference	
		0	1
Prediction	0	99	273
	1	21	136

RDA		Reference	
		0	1
Prediction	0	61	122
	1	59	287

Figure 5.27: Gap Oversampling Dataset Confusion Matrix

The GLM model reveals a relatively balanced prediction with 289 true positives and 59 false negatives. This suggests a moderate level of sensitivity, but with 120 false positives, the specificity is somewhat compromised. The overall performance is reasonable, but the model may be prone to over-predicting the positive class. The KNN model shows high sensitivity, with 332 true positives, but at the cost of many false positives (77), indicating a model favouring the positive class.

The GBM shows a good balance with 303 true positives and fewer false negatives (83) than KNN, but it still struggles with specificity, as indicated by the 106 false positives. The DT model has many true positives (306), indicating strong sensitivity similar to KNN. However, DT has fewer false positives (103) and false negatives (65), suggesting a

more balanced classification ability. The RF model stands out with the highest number of true positives (337) and a moderate number of false negatives (76). It also maintains fewer false positives (72) relative to its true positive rate, indicating a model with high sensitivity and better specificity than most other models.

The SVM model shows a significant number of true positives (301) with a relatively balanced number of false negatives (67) and false positives (108), which demonstrates its capability to predict the positive class while maintaining a moderate false positive rate. The NB model has the highest specificity, with the lowest number of false positives (99). However, it struggles significantly with sensitivity, having the highest number of false negatives (273) and the lowest number of true positives (136). The RDA model shows a reasonable number of true positives (287) but, like the GLM, has a relatively high number of false positives (122), affecting its overall specificity.

The RF model is the best model because it has shown a balance across the various measures, with strong sensitivity and a better handle on specificity than other models. It manages to predict a high number of true positives (students likely to pass) while keeping the number of false positives relatively low, which could make it the best choice for this specific application. It suggests that RF could be the most reliable for efficiently using resources to identify students who need support without over-identifying those who do not.

Variable of Importance The Random Forest model's variable importance for the Expectation and Experience Gap – Oversampling Outcome Model reveals insightful patterns about the factors contributing to academic outcomes.

Table 5.25: Gap Oversampling Variable Importance in the Random Forest Model

Importance Variable	Mean Decrease Gini
APS (Admission Points Score)	26.3
FA2 (Social Well-being)	18.3
FA3 (Academic Support)	17.5
FA4 (Access to Information)	16.6
FA1 (Effective Learning)	16.4
Female	15.1
Mainstream	9.4
African	7.1
Bursary	6.7
Residence	3.5

The Admission Points Score (APS) holds the highest Mean Decrease in the Gini score of 26.3, underlining its significance as a predictor of academic success. APS, representing the student's prior academic performance, strongly indicates future academic achievements. Social well-being (FA2) follows with a score of 18.3, implying that a student's social context is a substantial factor in their academic performance. Academic Support (FA3) and Access to Information (FA4), with scores of 17.5 and 16.6, suggest that students' resources and support are nearly as pivotal as social factors. Effective Learning (FA1) is also critical, with a score of 16.4, emphasising the role of learning strategies and educational practices in student outcomes.

Gender, represented by the variable Female, has a noteworthy importance score of 15.1, indicating potential differences in outcomes based on gender that warrant further investigation. The variables Mainstream, African, and Bursary show lower importance scores (9.4, 7.1, and 6.7, respectively), suggesting they have some impact on academic outcomes but are less dominant than the variables mentioned above. Residence, with the lowest score of 3.5, may have a minor direct impact on academic success compared to other factors.

The Random Forest model points to multidimensional influences on academic outcomes, where individual academic history and socio-environmental factors play significant roles. These insights can inform targeted interventions to support students, emphasising enhancing academic support and addressing social well-being to improve academic outcomes.

5.4.2.4 Undersampling Models

Table 5.26 provides a comprehensive overview of several classification models applied to the Expectation and Experience Gap – Undersampling outcome Model, which addresses an academic outcome prediction problem where the positive class indicates a pass (1).

The DT model outperforms the others regarding sensitivity (0.74) and F1 score (0.77), indicating a robust ability to identify students who will pass correctly. Despite having the lowest specificity (0.40), it suggests a propensity to over-predict the positive class, which may not be detrimental in an academic setting where the cost of missing out on a student who needs help is high. The RF and KNN models both show balanced performance, with RF having a slightly better accuracy (0.62) and KNN exhibiting a higher sensitivity (0.66). The F1 scores for both models are strong (0.72), indicating a good balance between precision and recall.

The SVM model has comparable metrics to RF, with an F1 score of 0.71, and might also be considered for its balance of sensitivity and precision. The GLM, GBM, and RDA

Table 5.26: Undersampling Model Performance Measures for GAP and Academic Outcome

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
GLM	0.60	0.61	0.58	0.83	0.61	0.70
KNN	0.61	0.66	0.45	0.80	0.66	0.72
GBM	0.59	0.61	0.52	0.81	0.61	0.70
DT	**0.66**	**0.74**	0.40	0.80	**0.74**	**0.77**
RF	0.62	0.64	0.53	0.82	0.64	0.72
SVM	0.61	0.64	0.49	0.81	0.64	0.71
NB	0.37	0.22	**0.88**	**0.86**	0.22	0.35
RDA	0.60	0.60	0.57	0.83	0.60	0.70

models show moderate performance, with F1 scores around 0.70. They have reasonable accuracy and sensitivity but do not reach the effectiveness of the DT model. The NB model suffers from very low sensitivity (0.22) and accuracy (0.37) despite having the highest specificity (0.88), making it the least suitable model for this dataset.

In summary, the Decision Tree model is the best candidate for the undersampling outcome model. Its high sensitivity and F1 score suggest that it effectively identifies students who will pass the academic threshold. Since identifying students who may require assistance is critical, the DT's tendency to favour the positive class can be particularly beneficial, as it minimises the risk of failing to offer support to those who need it.

Confusion Matrix The confusion matrices provided offer a detailed comparison of various classification models used in the Expectation and Experience Gap – Undersampling Outcome Model, where the goal is to predict the academic outcomes of students with a 'Pass' labelled as '1'.

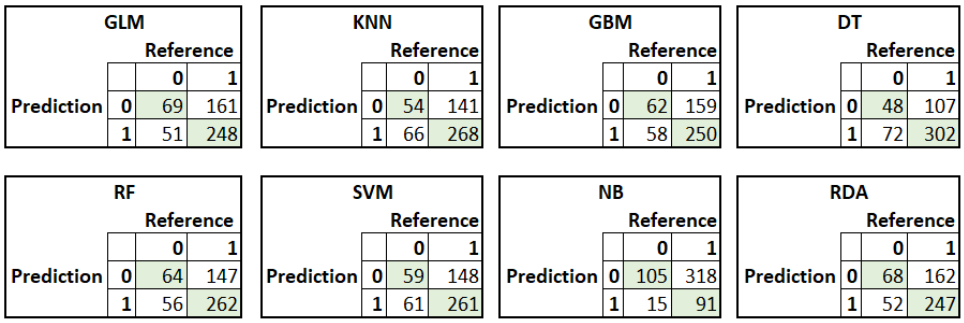


Figure 5.28: Gap Undersampling Dataset Confusion Matrix

The GLM model shows a relatively balanced distribution of predictions with a total of

248 true positives, indicating its capability to identify students who will pass. However, it also has 161 false positives, suggesting a tendency to overestimate the number of students passing. The KNN model has a higher number of true positives (268) compared to GLM, which reflects better sensitivity. The false positive count is lower than GLM's at 141, showing an improvement in specificity. The GBM model presents an intermediate performance with 250 true positives and 159 false positives. It suggests that while GBM is effective in identifying true passes, it also makes a considerable number of errors in over-predicting passes.

The DT model stands out with the highest number of true positives (302) among the models, indicating strong sensitivity. However, it also has a relatively high number of false positives (107), which may affect its overall specificity. The RF reveals a solid balance with 262 true positives and 147 false positives. Its performance indicates a tendency to correctly predict passing students, albeit with a moderate rate of over-prediction. The SVM model has a performance similar to RF, with 261 true positives and slightly more false positives (148). It appears to be competitive with RF in identifying students who are likely to pass. The NB model shows a concerning number of false negatives (318) and the lowest number of true positives (91), which indicates poor sensitivity, making it less suitable for identifying students who will pass. The RDA has a moderate performance with 247 true positives but also has a substantial number of false positives (162), similar to GLM.

In summary, the Decision Tree (DT) model appears to be the most favourable regarding correctly identifying students who will pass (high sensitivity), making it a strong candidate for an academic setting where it is crucial to offer support to as many students in need as possible. However, its higher false positive rate suggests that while it is good at capturing the positive class, it might also misclassify some students as passing when they are not. Thus, where resource allocation based on these predictions is critical, a balance between sensitivity and precision—like that provided by the Random Forest (RF) model—may be more desirable to avoid the misallocation of student support resources.

Variable of Importance This section explores the discussion of variables that hold significant importance for decision trees and random forests. Specifically, with a focus on the Experience and Expectation Gap - Undersampling - Outcome Model with Academic Outcome, where the 'Positive' Class (Pass=1 and Fail=0) is considered. It is essential to keep in mind that the aim is to understand how these variables contribute to decision-making processes; by exploring their significance, the study can gain valuable insights into the factors influencing outcomes related to academic achievement.

The variable importance rankings provided by the Decision Tree (DT) model for the Experience and Expectation Gap - Undersampling - Outcome Model have the factors that are most predictive of academic success. Consistently seen as a crucial predictor, the APS captures a student's accumulated academic performance before university. A higher APS often correlates with a stronger likelihood of academic success, suggesting that past performance is a robust indicator of future outcomes. The distinction between students enrolled in mainstream programs versus those in extended programs is significant. This variable's importance might reflect differences in curriculum difficulty, student preparedness, or resource allocation. It could indicate that students in mainstream programs are more likely to pass, possibly due to a variety of systemic and individual factors. Financial aid, indicated by the presence of a bursary, is another significant variable. This suggests that economic support impacts academic success, potentially by easing financial burdens that might otherwise detract from a student's focus on their studies.

The prominence of Social Well-being (FA2) points to the social aspect of the student experience as a key determinant of academic outcomes. This could encompass a range of factors, from social support networks to engagement in campus life, which contribute to a student's overall well-being and academic performance. The importance of Effective Learning (FA1) strategies is highlighted, emphasising the need for students to engage with course material in a way that promotes understanding and retention. This might relate to the quality of instruction, the learning environment, or the student's study habits. Academic Support (FA3), such as tutoring services, writing centres, or study groups, is identified as an important factor. Its influence highlights the value of institutional resources that directly support the learning process, assisting students in overcoming academic challenges.

The DT model's emphasis on these variables provides actionable insights for educational institutions. By understanding and enhancing the factors that contribute to student success—such as providing adequate financial aid, fostering supportive social environments, and ensuring effective learning and academic support—educational policies and practices can be better tailored to improve academic outcomes.

The Random Forest model's assessment of variable importance for the Experience and Expectation Gap - Undersampling Outcome Model reveals several key factors influencing academic outcomes where the positive class denotes students who have passed.

APS (Admission Points Score) stands at the top with a score of 12.2. APS confirms its status as a critical predictor of academic success. A higher APS, indicative of better prior academic performance, is closely aligned with favourable academic outcomes, reaffirming the value of historical academic data in predicting future success. Social Well-being

Table 5.27: Gap Undersampling Variable Importance in the Random Forest Model

Importance Variable	Mean Decrease Gini
APS (Admission Points Score)	12.2
FA2 (Social Well-being)	9.1
FA1 (Effective Learning)	8.9
FA3 (Academic Support)	7.7
FA4 (Access to Information)	7.6
African	4.1
Female	3.6
Mainstream	3.2
Bursary	2.1
Residence	1.2

(FA2) and Effective Learning (FA1), with scores of 9.1 and 8.9, respectively, point to the integral role of the students' social environment and their learning efficacy. These elements encompass a broad range of factors, from the students' interpersonal relationships and community engagement to their learning strategies and academic practices.

Academic Support (FA3) with a score of 7.7, the importance of academic support underscores the impact of the resources and assistance provided to students, which can include mentoring, tutoring, or access to study materials. Access to Information (FA4), with a score of 7.6, highlights the significance of students being able to access necessary information for their academic work, which is a fundamental component of a supportive learning environment.

Demographic factors such as being African or Female show moderate importance, with scores of 4.1 and 3.6, implying that there may be particular experiences or challenges associated with these identities that influence academic performance. The Program Type (Mainstream) variable holds some importance (3.2), suggesting that the type of academic program a student is enrolled in can affect their performance, potentially due to curriculum differences or the level of academic rigour. Financial Support (Bursary), with a score of 2.1, and Residence status, with a score of 1.2, although less influential, still play a role in academic outcomes. These may reflect the economic and living conditions that can either hinder or facilitate a student's academic journey.

The Random Forest model illuminates the complex nature of academic success predictors. By understanding the relative importance of these variables, institutions can develop targeted interventions and support structures to address the most impactful areas, thus

promoting academic achievement and reducing the likelihood of failure.

In summary, the Decision Tree (DT) and Random Forest (RF) models for the Experience and Expectation Gap - Undersampling Outcome Model both identify the Admission Points Score (APS) as the primary indicator of academic success, highlighting the significance of past performance in predicting future outcomes. The DT model places importance on program type (Mainstream), financial aid (Bursary), and elements of student welfare (FA2: Social Well-being, FA1: Effective Learning). Similarly, the RF model considers student support factors (FA3: Academic Support, FA4: Access to Information) and demographic variables (African, Female) as influential. These findings suggest that educational achievements are influenced by a combination of academic history, socio-economic factors, and support structures, guiding targeted interventions to foster student success.

5.5 Summary

In the Results chapter, an exploratory data analysis of the relationship between first-year university students' expectations, experiences, and subsequent academic performance was conducted. The primary objective was to understand how these features interact and influence student success, specifically focusing on Grade Point Average (GPA) and pass/fail status. The first section of the chapter conducted a data profiling analysis, which involved examining student demographics, categorical academic attributes, and numerical academic attributes. This analysis served as the foundation for developing a predictive model for academic performance. It provided an in-depth examination of various attributes such as gender, first-generation status, population group, bursary recipients, residence type, programme types, Admission Points Score (APS), and GPA.

Key findings from the data profiling analysis revealed important insights regarding demographic and categorical academic attributes. The study found that there was a balanced distribution of academic performance across genders. However, slight variations in pass/fail rates were observed. Additionally, first-generation students displayed marginally lower performance and success rates compared to their counterparts. Significant disparities in academic performance and outcomes were also identified among different population groups. African and Coloured students showed lower performance metrics compared to White and Indian students. The bursary recipients and residence status demonstrated notable differences in academic outcomes. The analysis of numerical academic attributes focused on the Admission Points Score (APS) and GPA. This statistical breakdown highlighted varied performance patterns among students performing above and below the median. Distinct differences were observed in mean, median, and

standard deviation values between these two groups.

Furthermore, the study explored student expectations and experience profiles. High expectations were uncovered regarding social engagement, academic support, and personal welfare. However, challenges related to workload management, financial issues, and academic integrity were also reported. Interestingly, students' actual experiences deviated from their initial expectations, particularly in areas such as social engagement and financial concerns. A significant difference between student expectations and actual experiences was identified, with certain aspects like social involvement and academic support showing a noticeable gap. To better understand the complex interrelations among various factors impacting student performance, an Exploratory Factor Analysis (EFA) was conducted. The EFA revealed four latent factors: Effective Learning, Social well-being, Academic Support, and Access to Information. These factors collectively explained a significant portion of the variance in the data.

The findings from the various Student Expectation predictive models for student performance and academic outcomes were focused on identifying the most effective models based on their accuracy, F1 score and the significance of various influencing factors. For student performance models, the Generalized Linear Model (GLM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Regularized Discriminant Analysis (RDA) were evaluated. The best model overall was found to be RDA, which had the highest Accuracy (0.61) and Specificity (0.73). GLM, GBM, and SVM also showed balanced performance with moderate F1 Scores (ranging from 0.56 to 0.57). Key variables that positively correlated with above-median academic performance included Bursary, APS, and gender, while Access to Information, racial background, and programme type showed negative correlations.

For academic outcome models categorised as Unbalanced, ROSE (Random OverSampling Examples), Oversampling, and Undersampling, different evaluation measures such as Accuracy, Precision, Recall (or Sensitivity), and F1 Score were considered. The Random Forest model consistently outperformed other models in terms of true positive rates across all categories. The strategies tested for handling unbalanced data (ROSE, Oversampling and Undersampling), oversampling resulted in the best performance. In particular, in the ROSE model, RF showed the highest sensitivity (0.69), accuracy (0.65), and F1 score (0.76). In the Oversampling model, RF again stood out with the highest accuracy (0.75) and F1 score (0.84). In the Undersampling model, both RF and DT demonstrated the highest accuracy (0.68), with RF excelling in precision (0.84) and F1 score (0.78). The key variables that emerged as important predictors across different models for academic

outcomes included APS, FA4 (Access to Information), FA3 (Academic Support), FA2 (Social Well-being), FA1 (Effective Learning), gender, program type, financial aid, and racial background. APS consistently appeared as a critical predictor. Based on these findings, it can be concluded that the Random Forest model is the most reliable for predicting student success across various academic outcome models. APS is identified as a crucial factor in predicting academic success, alongside factors such as Access to Information, Academic Support, and Social Well-being.

Lastly, the Expectation and Experience Gap model results were obtained from evaluating various predictive models such as Balanced-Performance, Unbalanced, ROSE, Oversampling, and Undersampling models. The evaluation of the Balanced-Performance Dataset revealed that the GLM and RDA demonstrated a worthy balance across multiple metrics in predicting academic performance. These models showed promise for accurately predicting students' performance. The confusion matrix analysis highlighted each model's ability to classify academic performance. Notably, GLM, NB, and RDA exhibited a balanced classification approach. However, all models highlighted areas for improvement in balancing false positives and negatives. The GLM identified key predictors of academic success, including financial support (Bursary), APS, and access to information (FA4). The GBM model also emphasized APS as a critical predictor, along with factors like Effective Learning (FA1) and Social Well-being (FA2).

Furthermore, academic outcomes measures of the unbalanced models like GLM demonstrated high accuracy and F1 scores but struggled with reliably predicting 'Fail' outcomes due to low specificity. ROSE models had varied performances, with DT excelling in sensitivity but showing a propensity towards false positives. In oversampling scenarios, the RF model emerged as the most effective due to its balanced accuracy, sensitivity, specificity, and F1 score. Conversely, in undersampling scenarios, the DT stood out with high sensitivity and an F1 score for identifying students likely to pass. This analysis highlights that predictive modelling for academic performance is multifaceted. Each model has unique strengths and limitations. The GLM and RF models are notable for their balanced performance across different scenarios. However, the choice of a specific model should consider the unique requirements of the academic context and the implications of false predictions. The variables of importance identified in this study, such as APS, financial support, and access to information, provide critical insights for policy interventions and further research into addressing disparities in academic outcomes.

The data profiling segment provided an essential foundation for understanding the complex aspects influencing first-year university students' academic performance. By examining demographic, categorical, and numerical attributes, this study offered crucial

insights into patterns and disparities in academic performance. By exploring students' expectations and experiences alongside the power of EFA analysis, a view of the factors contributing to academic success or failure was obtained. These insights are invaluable for developing targeted interventions and strategies to strengthen student performance and enhance their overall university experience. Ultimately, this detailed analysis lays a foundation for informed decision-making by aligning student support mechanisms with their actual needs and expectations.

Chapter 6

Conclusion

6.1 Introduction

The initial transition into university life is a pivotal phase for first-year students, marked by numerous changes and challenges. Transition theory, as articulated by Gardner [38], offers a valuable framework for understanding the various challenges. It suggests that successful adaptation during the transition is influenced by three core factors: the individual's perception and the inherent characteristics of the transition itself, the traits of the surrounding environment, and the personal attributes of the individual changing.

Central to the success of the transition is the academic and environmental context and the extent of the student's academic and social involvement. Astin and Magolda et al. [3, 76] highlight the significance of student involvement theory, which suggests that the degree of physical and psychological effort a student invests in their academic and extracurricular activities directly impacts their learning, development, and likelihood of graduating.

Tinto's theory of student departure [101, 102, 103] further expands on these concepts, highlighting that student attrition is often a result of academic difficulties, challenges in social and intellectual integration, or a diminished commitment to the institution. This theory lays a foundation for understanding student retention and departure.

To address these challenges, first-year experience programs, including outdoor or orientation initiatives, are designed to resonate with the key elements of transition, student involvement, and departure theories. These programs aim to enhance student retention and persistence by fostering a smoother transition into university life.

Adding to this theoretical landscape is Lizzio's Five Senses of Success Framework, developed by Alf Lizzio at Griffith University [72]. This framework presents a holistic approach to addressing potential gaps in the early university experience. It is composed

of five key principles that are integral to aiding students in their successful transition into university life: the importance of academic preparedness, self-awareness, connectedness, academic strategies, and a sense of purpose. By addressing these principles, universities can tackle various barriers to student success, including those arising from disruptive environments.

The aim of the study was to investigate the influence of first-year expectations and experiences on the academic performance of first-time entering students in a formal qualification at the University of the Western Cape, South Africa. The study used the Cross-Industry Standard Process for Data Mining (CRISP-DM) and aimed to predict the academic outcome and performance, thereby enabling the early implementation of strategic interventions.

The study was structured around four specific objectives designed to create a comprehensive framework for understanding and enhancing student success. These objectives were:

1. Conduct a comprehensive data profiling of key areas: student demographics, academic attributes, expectations, and experiences.

2. Implement exploratory factor analysis to determine the factors within first-year expectations and experiences that influence student academic performance.

3. Investigate the feasibility of predicting student academic performance based on the factors identified through the analysis and demographic data.

4. Formulate recommendations for developing a student intervention strategy based on the findings derived from the first three objectives.

Chapters 1, 2, and 3 presented the study's introduction to the topic, background on the concepts related to data mining techniques, the first-year experience, various definitions of student academic performance, and the review of the literature. Chapter 4 discussed the study's methodology, including the data collection method, variables of interest, the data analysis plan, and ethical considerations. Chapter 5 discussed the results and an evaluation of the predictive models developed to address the aim and objectives of the study. The study's findings and conclusions of how the study has addressed the research aim and objectives, as well as the recommendations for further research, are discussed in this chapter.

6.2 Analysis of the Results

6.2.1 Sub-Objective 1: To perform data profiling of student academic performance, student expectations, and student experience data

This study reflects on the data profiling conducted to analyse student academic performance, expectations, and experiences. The profiling process began with a thorough examination of student demographics, including gender, first-generation status, and population group, in conjunction with categorical academic attributes like bursary recipients, residence type, and program types. Numerical academic attributes such as Admission Points Score (APS) and Grade Point Average (GPA) were also examined. The aim was to determine how these factors correlate with academic performance.

The analysis generated several key findings. Firstly, there was a relatively balanced distribution of academic performance across genders. However, subtle variations in pass and fail rates suggested underlying complexities, with female students outperforming males.

A marginally lower performance among first-generation students highlighted the potential challenges faced by this demographic, which could stem from a range of socio-economic factors. These students, the first in their families to attend university, may confront a variety of socio-economic hurdles. They often lack the familial guidance and academic support systems that their peers with college-educated parents might take for granted [34]. This absence of a familial academic background can lead to difficulties in navigating the complex landscape of higher education, from administrative processes to academic expectations. Consequently, these students may require additional support and resources to bridge this experiential gap and achieve academic success.

The observed differences in academic performance across different population groups, particularly among African and Coloured students, point towards deeper systemic issues. These groups often face unique challenges that stem from a complex interplay of historical, socio-economic, and cultural factors [34]. The legacy of inequality and limited access to quality pre-university education can result in a preparedness gap for these students upon entering higher education. Additionally, these students might confront subtle biases and a lack of representation within the academic environment, which can further impact their academic engagement and performance. This situation highlights the need for universities to acknowledge these systemic disparities and actively work towards creating inclusive and equitable educational environments. This could involve implementing targeted support programs, enhancing diversity within faculty, and fostering a campus culture that embraces and supports all student demographics.

Bursary recipients generally achieved better academic outcomes, hinting at the impact of financial aid on educational success. Research consistently shows that financial aid, particularly in the form of bursaries, positively impacts student engagement, academic performance, and persistence [16]. Bursary recipients are more likely to engage with peers, participate in community service, and demonstrate higher levels of dedication to their studies [16]. They also show higher levels of retention and success, with positive attitudes towards their studies and institutions [7].

Similarly, residence status showed differences in outcomes, possibly reflecting the influence of living conditions on academic engagement. Hountras et al. [51] found that students living in residence halls generally had higher GPAs, while Turley et al. and Sikhwari et al. [75, 94] both found that living on campus was associated with better academic performance. However, Turley et al. [75] noted that this effect was particularly noticeable for Black students and those attending liberal arts institutions. Schudde [91] further supported the positive impact of campus residency on student retention. These findings suggest that residence status can indeed influence academic outcomes, with living on campus potentially providing a more conducive environment for academic success.

The variation in Admission Points Score (APS) and Grade Point Average (GPA) across students reflects a wide academic preparedness and achievement spectrum. This diversity in performance measures indicates that students enter university with varying readiness levels, which could be influenced by factors such as prior educational experiences, socio-economic backgrounds, and access to preparatory resources. The differences in APS and GPA highlight the importance of personalised academic support and the need for universities to tailor their teaching and support services to cater to this wide range of academic abilities and backgrounds.

The profiling of student expectations and experiences indicates instances where students enter university with high expectations, particularly in terms of social engagement, academic support, and personal well-being. These high expectations often reflect an idealised view of university life, where students anticipate vibrant social interactions, robust academic guidance, and a supportive environment that nurtures their overall well-being. However, these expectations frequently clash with reality as students encounter the actual challenges of university life. Among these are the daunting tasks of managing significant workloads, grappling with financial constraints, and navigating the complexities of academic integrity. This divergence between expectation and experience can lead to disillusionment and stress, affecting students' academic performance and personal satisfaction.

This difference is most evident in social engagement and financial management, where the actual experiences of students often fall short of their initial expectations. The so-

cial aspect, a critical component of the university experience, is frequently less engaging and supportive than anticipated, potentially leading to feelings of isolation or alienation. Similarly, financial concerns, which might not have been fully anticipated or understood before entering university, emerge as a significant source of stress and distraction. These gaps highlight the need for universities to actively work towards bridging this divide. By implementing more comprehensive orientation programs, enhancing financial support systems, and creating more inclusive social environments, universities can significantly improve the alignment between student expectations and their actual experiences. Such interventions are vital for improving student satisfaction and crucial for fostering an environment conducive to academic success and personal growth.

A range of factors has been found to influence student performance, including academic stress and well-being [85], social support and adaptation to university life [4], and student engagement in academic activities [110]. These factors are interconnected, with social support and self-compassion mediating the relationship between academic stress and well-being [85], and student-student and teacher-student relationships central to student engagement. These findings highlight the importance of addressing student expectations and experiences to support their academic success.

In conclusion, data profiling has revealed the multifaceted nature of first-year students' academic journeys. It has laid bare the disparities based on demographic factors and the expectation-experience gap, highlighting the necessity for targeted support and improvement in university systems. By understanding these critical elements, stakeholders are better positioned to craft strategies that foster equitable educational opportunities and improve overall academic outcomes, ultimately contributing to a more robust and supportive academic environment for all students.

6.2.2 Sub-Objective 2: Use factor analysis to identify the factors from first-year expectations and experiences that impact student academic performance

This section addressed Sub-Objective 2 by using factor analysis to group first-year student expectations and experiences into underlying factors that impact academic performance. Exploring student expectations and experiences offers a window into the potential determinants of academic success and areas where student support may be optimised. The application of Exploratory Factor Analysis (EFA) has led to the identification of Effective Learning (FA1), Social well-being (FA2), Academic Support(FA3), and Access to Information (FA4) as four of the latent factors influencing academic performance or academic

outcome.

Effective Learning includes the strategies and habits that facilitate acquiring and applying knowledge. At the same time, Social well-being reflects the quality of students' social interactions and personal satisfaction. The gap in students' expectations versus experiences in these areas could affect their academic outcomes and their overall university tenure. Additionally, Academic Support and Access to Information emerged as crucial factors. Academic support involves the institution's resources and services to help students in their academic journey. Access to Information refers to the availability and quality of information resources that students require for their studies. The significant variance explained by these factors within the EFA highlights their importance in the academic ecosystem.

The factor analysis conducted offers a strategic advantage. By understanding these latent factors, universities can develop targeted interventions to close the gap between expected and actual experiences. For instance, strengthening academic support services and ensuring equitable access to Information can directly address the areas where students' experiences fell short of their expectations.

6.2.3 Sub-Objective 3: Investigate whether there is a possibility to predict student academic performance based on identified factors and demographic data

The development of the Student Expectations, Experience, and Gap Predictive model within the South African public higher education context provides valuable input into factors that can potentially predict academic performance. The investigation into the potential for predicting student academic performance based on identified factors and demographic data has produced compelling evidence by applying various predictive models. This section discusses the viability of predicting student performance or academic outcomes based on identified factors and demographic data, aiming to culminate in a strategic framework for student intervention.

Moreover, the study leverages predictive modelling to identify key factors such as the Admission Points Score (APS), financial aid, and gender, aligning with findings on the predictive potential of various student attributes. These factors, along with Access to Information, Academic Support, Social Well-being, and Effective Learning, have been identified as significant through the application of models across different models including Generalized Linear Model (GLM), K-nearest neighbours (KNN), Gradient Boosting Machine (GBM), Decision Trees (DT), Random Forest (RF), Support Vector Machine

(SVM), Naive Bayes (NB), and Regularized Discriminant Analysis (RDA).

Table 6.1 summarises the different student performance (Above/Below Median) models. The accuracy levels were below 70% for the student performance models, suggesting that these models hold minimal potential and are not definitive solution models for student performance. This reinforces the notion that predictive models should be part of a broader decision-making framework rather than standalone tools.

Table 6.1: Summary of the Student Performance (Above/Below Median) Models

Model	**Expectation**		**Experience**		**Gap**	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
GLM	0.60	**0.57**	0.57	0.56	**0.58**	**0.55**
KNN	0.58	0.56	0.56	0.55	0.55	**0.55**
GBM	0.60	**0.57**	0.57	0.55	**0.58**	0.54
DT	0.58	0.52	0.58	0.52	0.57	0.52
RF	0.59	**0.57**	0.54	0.53	0.57	0.53
SVM	0.60	0.56	**0.60**	**0.60**	0.57	**0.55**
NB	0.58	0.38	0.57	0.38	**0.58**	0.38
RDA	**0.61**	0.55	0.55	0.54	**0.58**	0.54

The RDA model also demonstrated high accuracy in the expectation and gap datasets, suggesting its potential effectiveness in certain academic settings. The low-valued F1 scores also suggest the need for an approach to model selection and application. The models analysed did not show a strong predictive relationship, indicating that additional factors and more sophisticated modelling techniques may be required to predict academic performance accurately.

Table 6.2 summarises the different academic outcome (pass/fail) models. The academic outcome models have highlighted the Random Forest model as the best model with superior performance in multiple scenarios, especially in handling oversampled and undersampled datasets. An F1-score above 0.70 suggests the model has a good balance between precision (the model's ability to identify positive instances correctly) and recall (the model's ability to find all positive instances). This implies that the model is fairly reliable in correctly predicting whether a student will pass or fail, making it a potentially useful tool for educational institutions. While an F1-score or accuracy above 0.70 is good, there is always room for improvement. This benchmark can serve as a baseline for further model refinement, incorporating more data or exploring more complex modelling techniques.

Table 6.2: Summary of the Academic Outcome (pass/fail) Models Performance

Type	Accuracy	F1 Score	Specificity	Best Model
Expectation				
Unbalanced	0.78	0.87	0.02	GLM
ROSE	0.65	0.76	0.52	DT
Oversampling	0.75	0.84	0.37	RF
Undersampling	0.68	0.78	0.52	RF
Experience				
Unbalanced	0.78	0.87	0.03	GLM, GBM
ROSE	0.68	0.78	0.44	DT
Oversampling	0.72	0.82	0.37	RF
Undersampling	0.66	0.77	0.40	DT, RF
Expectation and Experience Gap				
Unbalanced	0.78	0.87	0.03	GLM
ROSE	0.68	0.78	0.44	DT
Oversampling	0.72	0.82	0.37	RF
Undersampling	0.66	0.77	0.40	DT

For models based on expectation, the GLM shows the highest accuracy and F1 score (0.78 and 0.87, respectively) but shows minimal specificity (0.02), indicating a strong ability to predict pass outcomes but a limited capability in identifying fail outcomes. The RF model is identified as the best model under both oversampling and undersampling techniques, with the oversampling technique showing slightly better performance in accuracy and F1 score (0.75 and 0.84) compared to undersampling (0.68 and 0.78).

For models based on experience, the GLM and GBM models share the top spot in unbalanced datasets, both showing the same accuracy and F1 score as in the expectation scenario. However, the DT and RF models are highlighted under the ROSE and undersampling techniques, respectively, indicating variability in model performance based on the sampling technique applied.

When analysing the gap between expectation and experience, the performance measures remain consistent with those observed in the experience scenario, highlighting the similarity in model performance when this gap is considered.

Overall, Table 6.2 indicates that the GLM performs exceptionally well in unbalanced datasets across all three criteria. However, the performance of models varies significantly with the application of different sampling techniques, with the RF model frequently emerg-

ing as the best model in scenarios involving oversampling and undersampling, suggesting its robustness in handling imbalanced data.

Assessing the variables of importance, the Admission Points Score (APS) has emerged as a significant predictor of student success. Universities should consider implementing preparatory programs designed to bridge the gap between high school and university, enhancing students' readiness for higher education. These could include summer bridge courses, during which students could strengthen their understanding of core subjects, or early assessment programs to identify and address learning gaps before the semester begins. Other important predictors included Access to Information, Academic Support, Social Well-being, Effective Learning, gender, program type, financial aid, and racial background.

The models highlighted the importance of academic support and effective learning strategies. Universities should develop robust support systems beyond traditional tutoring, encompassing mentorship programs, academic advising, study groups, and workshops that teach effective study techniques. These systems should be easily accessible and actively promoted to ensure that students know the resources available to them.

Social well-being was also a critical predictor of student success. Universities should foster a campus environment that promotes social integration and connection. This could involve facilitating student clubs, extracurricular activities, and community service programs that connect students with peers who share similar interests and challenges. Additionally, providing resources for mental health support, such as counselling services and wellness seminars, can also play a vital role.

The importance of variables such as ethnicity and gender in the models suggests that a one-size-fits-all approach to education and support is insufficient. Universities should strive to understand the unique experiences of different student demographics and tailor support accordingly. This could involve creating targeted high-impact programmes, or establishing diversity and inclusion centres that provide a safe space for underrepresented students.

Access to information was identified as a key variable in some models. Universities should leverage educational technology to provide personalised learning experiences. This could include learning platforms that adjust to individual student's progress, online resources for flexible learning, and data analytics to track student engagement and performance, allowing for timely intervention.

The presence of a bursary as a variable indicates the impact of financial support on academic success. Universities should ensure that bursaries and scholarships are both merit-based and need-based, helping to alleviate the financial burden on students who

may otherwise struggle to afford higher education.

Residence life plays a role, albeit a smaller one, in student success. Universities could enhance residential life programs to ensure that campus living is conducive to learning. This includes providing quiet study areas, promoting a community atmosphere, and offering residence-based academic support.

The findings suggest a strong potential for employing predictive models to predict academic outcomes of whether a student will pass or fail their first year. The Random Forest model, in particular, offers robust predictive capabilities, effectively managing the trade-offs between sensitivity and specificity. This model can be instrumental in identifying students who may require additional support, enabling the institution to allocate resources more efficiently. While the potential of predictive modelling is evident, it is not without its limitations. The complexity of academic outcomes and the dynamic nature of student experiences necessitate a cautious approach. A predictive model can be a valuable tool. However, it must be integrated with a broader strategy considering the holistic educational environment.

6.3 Recommendation

6.3.1 Sub-objective 4: To make a recommendation based on the findings of sub-objectives 1 through 3 by developing a student intervention strategy

In response to Sub-objective 4, which seeks to formulate recommendations drawn from the findings of the preceding sub-objectives, this section aims to develop a strategic framework for first-year student interventions. The insights gathered from an analysis of student performance and academic outcome predictors ranging from individual academic preparedness to the broader spectrum of social and institutional support mechanisms inform these recommendations.

The primary goal is to develop a comprehensive intervention strategy that identifies at-risk students early and implements tailored support systems to enhance their academic journey. This strategy should be grounded in the data and data analytics provided by the predictive models and refined through the empirical outcomes observed in sub-objectives 1 through 3. These recommendations serve as a blueprint for proactive engagement, ensuring that students receive the necessary resources and guidance to thrive in their academic journey.

Based on the findings of the study, it is evident that first-year students entering uni-

versity often have unrealistic expectations regarding effective learning, academic support, social well-being, and access to information. These unrealistic expectations are influenced by their prior experiences and perceptions of university education. The study recommends further qualitative investigation of student expectations and experiences and the exploration of intentional and targeted interventions to manage these expectations when necessary [83].

The literature supports the notion that students' expectations of university education are often shaped by their prior experiences and perceptions. For instance, it has been reported that a significant proportion of first-year students expect university teaching styles to be similar to those experienced in school [46]. These findings underscore the need for interventions that address the misalignment between students' expectations and the reality of their university experiences.

There is an emphasis on the significance of student development [102], highlighting the need for intentional and targeted interventions to support students in managing their expectations [72]. Additionally, It is essential to create the conditions that foster student success, which is important to address the unrealistic expectations of first-year students [17].

The recommendation aims to provide universities with a timeline and structure for support systems that enhance the educational experience without overwhelming students with data playing a pivotal role. Data analytics is key in the early identification of students at risk of academic underperformance by analysing admissions data, such as Academic Performance Score (APS), first-generation status, residence, and bursary information. Institutions can differentiate patterns that highlight students potentially facing challenges characterised by lower APS, first-generation attendance, lack of university residence housing, and financial aid needs by employing predictive models. This approach can enable targeted interventions tailored to individual needs around academic support, financial advice, and student support services.

Figure 6.1 presents a framework designed to improve the effectiveness of the strategy through the continuous refinement of interventions grounded in data-driven insights and empirical results. This approach ensures that students receive timely and targeted support during key phases of their academic journey, thereby promoting success across a diverse range of backgrounds. It is essential that the rollout of these support services is staggered and accompanied by clear communication to avoid overwhelming students. This strategy will help students to be aware of the available resources without being burdened with an excess of options at any one time.

Support should begin even before students set foot on campus by offering support

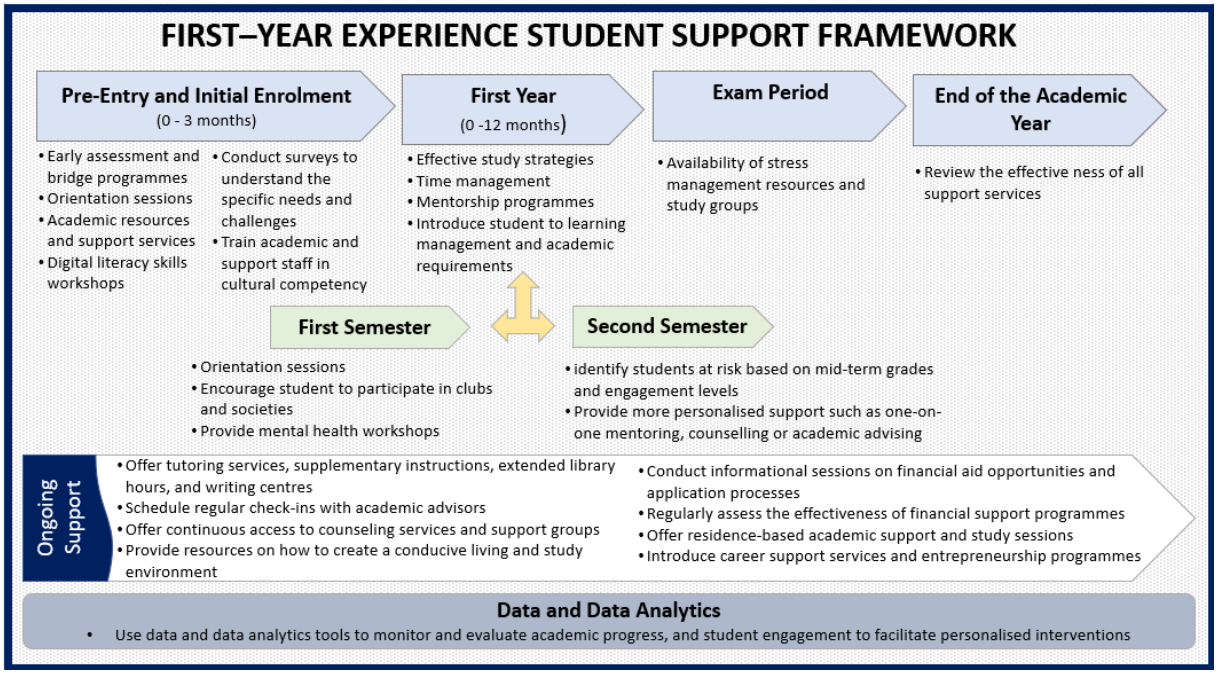


Figure 6.1: Proposed First-Year Student Support Strategic Framework

during open days, school visits,and developing relationships with some schools. Once students are on campus campus programmes can then be offered such as orientation, academic support, and other programmes. Orientation programmes should include surveys to understand the specific needs, academic preparedness, and challenges of diverse student groups academic preparedness, with particular attention to the Admission Points Score (APS) and other academic indicators. Pre-admission workshops focusing on bridging educational gaps and providing an overview of university resources can be instrumental. Conduct and provide digital skills workshops, as well as train faculty and staff in cultural competency to support the creation of an inclusive learning environment. Introduce students to the learning management platforms before the lecturers start. This period is crucial for setting the stage for ongoing support and familiarising students with the resources available to them.

Academic support services should be most intensive during the first six weeks of the semester, as this is a critical period for student adjustment and habit formation. Workshops on effective learning strategies and time management can be offered during this window. Peer mentoring programs should also be initiated early in the semester to provide continuous academic and social support. Implement early assessment and bridge programs for incoming students, focusing on core academic skills and introducing orientation sessions that highlight available academic resources and support services.

In the first semester, focus on establishing a solid foundation of support services geared

towards the transition into university life by encouraging participation in student clubs and organisations early in the university experience to foster a sense of community and provide mental health workshops within the first month to help students cope with the transition to university life.

Support services should shift towards more personalised interventions by the second semester. Data analytics should be used to identify students at risk based on mid-term grades and engagement levels. Tailored support, such as one-on-one tutoring, counselling or academic advising, is to be provided to address specific needs.

During the exam periods, the focus should shift to revision strategies and stress management. This phase should see an increase in the availability of tutors and study groups, ensuring that students can effectively prepare for exams without excessive pressure.

At the end of the academic year, there is a need to review the effectiveness of all support services and make necessary adjustments for the following year. The university should conduct debriefing sessions to review performance. This should include reflective activities that help students identify successful strategies and areas needing improvement, informing support strategies for the following semester.

On an ongoing basis, it is important to continue providing the following support:

- Offer tutoring services, extended library hours, and writing centres throughout the academic year.
- Schedule regular check-ins with academic advisors to monitor progress and address concerns.
- Continuously evaluate and adapt scholarship programs to ensure they meet the changing needs of the student body.
- Offer continuous access to counselling services and support groups throughout the academic journey.
- Use data analytics tools to monitor student engagement and facilitate personalised academic interventions.
- Conduct informational sessions on financial aid opportunities and application processes.
- Regularly assess the effectiveness of financial support programs in aiding student success and make adjustments as needed.
- Provide resources on creating a conducive living and study environment within residence halls.

- Offer residence-based academic support and study sessions throughout the academic year.

The study highlights that the timing of interventions during the first year is important for student success and retention. Early interventions, particularly in the first weeks of the academic term, are essential to help students adjust to university life and can significantly impact their retention and success rates [63, 104]. For example, Tinto (2006) emphasises that interventions during the initial phase of university can address challenges before they become impossible [104]. Similarly, Kuh et al. (2005) argue that timely academic support and engagement opportunities are pivotal in promoting student success and preventing early withdrawal [63].

Aligning interventions with critical milestones and potential stress points throughout the academic year, such as mid-exams and finals, can be particularly beneficial. Yorke and Longden (2008) suggest that providing support during these high-stress periods can help students manage their workload more effectively and reduce the likelihood of dropout [112]. This highlights the importance of not only the nature of the interventions but also their timing in effectively addressing student needs.

Focusing on critical stages allows universities to align their support services with periods when students are most receptive and in need. This strategic timing ensures that interventions are not only available but also optimally deployed to enhance their effectiveness on student success.

In conclusion, a university's commitment to fostering a supportive educational environment is best reflected in the deliberate planning and implementation of its services. The proposed strategic framework is designed to improve student readiness, provide continuous academic and social support, meet individual needs, and safeguard financial stability. By thoughtfully distributing these services across the academic timeline, the university can effectively and sustainably promote student success. Continuous evaluation and responsive adjustments will ensure that these services evolve with student needs, fostering a dynamic and supportive academic community.

6.4 Further Research

Future research should aim to broaden the various predictors used in modelling student academic performance. This could involve incorporating diverse data sources, such as learning management system engagement, in-depth analysis of study habits, student motivation levels, and other personal attributes that may influence academic performance.

Exploring advanced modelling techniques, including ensemble methods and deep learning algorithms, has the potential to enhance the accuracy and depth of the insights generated significantly. To further refine and validate these predictive models, future studies should extend the range of predictive factors under consideration. Integrating a temporal component into the analysis will allow for the examination of trends and patterns over time, offering a dynamic perspective on the data.

Time series analysis and forecasting methods can be used to capture temporal dependencies and seasonal variations, which are critical to understanding the evolution of key metrics. This approach will enable the development of models that can predict future outcomes based on historical data, thus improving their robustness and applicability in real-world scenarios.

Leveraging cutting-edge analytics, such as machine learning and artificial intelligence, will be crucial in advancing the precision and customisation of these models. These technologies can be used to analyse large volumes of time series data, identifying intricate temporal patterns that traditional methods could overlook. By incorporating temporal dimensions, the models will be not only more accurate but also more adaptive to changes over time, thereby providing deeper and more actionable insights.

Another key area of future research should focus on gender dynamics to understand the nuances of how gender may influence academic outcomes. This investigation is crucial in addressing potential disparities and ensuring academic equity.

Developing and implementing technological solutions also present a productive ground for future research. This includes creating a Decision Support System (DSS) that merges an academic performance prediction classifier with a user-friendly interface. Such a system could be designed for increased efficiency, allowing for batch processing of student data and offering strategic intervention suggestions tailored to individual or group needs. The DSS should be developed considering sustainability and accessibility, aligning with operational research techniques that emphasise academic solutions for academic challenges.

In conclusion, the path forward for research in this domain includes expanding predictive factors, exploring innovative modelling techniques, and the development of practical technological tools. These efforts will collectively contribute to an academic environment that is more supportive, responsive, and conducive to student success.

Bibliography

[1] Ashraf Abazeed and Moaiad Khder. A classification and prediction model for students performance in university level. *Journal of Computer Science*, 13(7):228–233, July 2017. doi: 10.3844/jcssp.2017.228.233.

[2] Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, Inc., July 2002. doi: 10.1002/0471249688.

[3] Alexander W. Astin. Student involvement: A development theory for higher education. *Journal of College Student Development*, 25(4):297–308, 1984.

[4] Mohd Mahzan Awang, Faridah Mydin Kutty, and Abdul Razaq Ahmad. Perceived social support and well being: First-year student experience in university. *International Education Studies*, 7(13), Dec 2014. ISSN 1913-9020. doi: 10.5539/ies.v7n13p261.

[5] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, Feb 2011. doi: 10.1002/mpr.329.

[6] E Babbie, JH Vorster, and C Payze. *Practice of business and social research*. Oxford University Press Southern Africa, Goodwood, South Africa, February 2001.

[7] Ghada Badr, Afnan Algobail, Hanadi Almutairi, and Manal Almutery. Predicting students' performance in university courses: A case study and tool in KSU mathematics department. *Procedia Computer Science*, 82:80–89, 2016. doi: 10.1016/j.procs.2016.04.012.

[8] CJ Bamforth. Improving undergraduate performance via an embedded generic skills program. In *Proceedings of HERDSA Conference*, pages 49–59. Research and Development in Higher Education: Reshaping

Higher Education, 6-9 July 2010 2010. URL https://www.herdsa.org.au/research-and-development-higher-education-vol-33.

[9] Albert Bandura. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2):191–215, 1977. ISSN 0033-295X. doi: 10.1037/0033-295x.84.2.191.

[10] Albert Bandura. Human agency in social cognitive theory. *American Psychologist*, 44(9):1175–1184, 1989. ISSN 0003-066X. doi: 10.1037/0003-066x.44.9.1175.

[11] Albert Bandura. Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, 28(2):117–148, March 1993. ISSN 1532-6985. doi: 10.1207/s15326985ep2802_3.

[12] Susan R. Barclay. *Schlossberg's Transition Theory*, pages 23–34. Springer Publishing Company, May 2017. ISBN 9780826118165. doi: 10.1891/9780826118165.0003.

[13] Betsy O. Barefoot. The first-year experience. *About Campus: Enriching the Student Learning Experience*, 4(6):12–18, January 2000. doi: 10.1177/108648220000400604.

[14] Maurice S. Bartlett. A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):296–298, 1954.

[15] Johannes Berens, Kerstin Schneider, Simon Gortz, Simon Oster, and Julian Burghoff. Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining*, 11(3):1–41, 2019. doi: 10.5281/ZENODO.3594771.

[16] Angela Boatman and Bridget Terry Long. Does financial aid impact college student engagement?: Evidence from the Gates Millennium Scholars Program. *Research in Higher Education*, 57(6):653–681, February 2016. ISSN 1573-188X. doi: 10.1007/s11162-015-9402-y.

[17] JM. Braxton, AS. Sullivan, and RM. Johnson. Appraising Tinto's theory of college student departure. In *Higher education: Handbook of theory and research*, volume 12, pages 107–164. Agathon Press, New York, NY, US, 1997.

[18] April A. Brecht. *A Study of the Factors That Predict Academic Success and Retention of Student-Athletes*. PhD thesis, Old Dominion University, 2014. URL https://digitalcommons.odu.edu/efl_etds/86/.

[19] Kendrick Brown. Coloured and Black relations in South Africa: The burden of racialized hierarchy. *Macalester International*, 9(13), 2000. URL http://digitalcommons.macalester.edu/macintl/vol9/iss1/13.

[20] Marko Bursać, Marija Blagojević, and Danijela Milošević. Early prediction of student success based on data mining and artificial neural network. In *Human Centered Computing*, pages 26–31. Springer International Publishing, 2019. doi: 10.1007/978-3-030-37429-7_3.

[21] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations inr. *Journal of Statistical Software*, 45(3), 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03.

[22] James R. Carpenter, Michael G. Kenward, and Stijn Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(3): 571–584, January 2006. ISSN 1467-985X. doi: 10.1111/j.1467-985x.2006.00407.x.

[23] Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[24] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. CRISP-DM 1.0 Step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000. URL https://the-modeling-agency.com/crisp-dm.pdf.

[25] NV. Chawla, KW. Bowyer, LO. Hall, and WP. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

[26] Arthur W. Chickering and Linda Reisser. *Education and Identity*. Jossey-Bass, London, England, 2 edition, November 1993.

[27] Anna B. Costello and Jason Osborne. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10, 2005. doi: 10.7275/JYJ1-4868. URL https://scholarworks.umass.edu/pare/vol10/iss1/7/.

[28] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, September 1951. doi: 10.1007/bf02310555.

[29] Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D. Lytras, Farhat Abbas, and Jalal S. Alowibdi. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion*. ACM Press, 2017. doi: 10.1145/3041021.3054164.

[30] Ann Boyd Davis and Petrus Venter. The performance and success of postgraduate business students. *Progressio*, 33(2):72–90, 2011.

[31] DHET. 2000 to 2016 first time entering undergraduate cohort studies for public higher education institutions. *Department of Higher Education and Training*, 2019. URL https://www.dhet.gov.za/SitePages/Higher-Education-Management-Information-System.aspx.

[32] Christine DiStefano, Min Zhu, and Diana Mîndrilã. Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, 14:1–11, 2009. doi: 10.7275/DA8T-4G52. URL https://scholarworks.umass.edu/pare/vol14/iss1/20/.

[33] Maria do Carmo Nicoletti. Revisiting the Tintos theoretical dropout model. *Higher Education Studies*, 9(3):52, June 2019. doi: 10.5539/hes.v9n3p52.

[34] Jennifer Engle and Vincent Tinto. Moving beyond access: College success for low-income, first-generation students. *Pell Institute for the Study of Opportunity in Higher Education*, 2008.

[35] Leandre R. Fabrigar, Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272, 1999.

[36] Andy Field. *Discovering statistics using IBM SPSS statistics: North American edition*. SAGE Publications, Thousand Oaks, CA, Nov 2017.

[37] Tierra M. Freeman, Lynley H. Anderman, and Jane M. Jensen. Sense of belonging in college freshmen at the classroom and campus levels. *The Journal of Experimental Education*, 75(3):203–220, April 2007. ISSN 00220973, 19400683. URL http://www.jstor.org/stable/20157456.

[38] John N. Gardner. *Launching the First-Year Experience Movement: The Founder's Journey*. Routledge, June 2023. ISBN 9781003445562. doi: 10.4324/9781003445562.

[39] Atul Garg, Umesh Kumar Lilhore, Pinaki Ghosh, Devendra Prasad, and Sarita Simaiya. Machine learning-based model for prediction of student's performance in higher education. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, August 2021. doi: 10.1109/spin52536.2021.9565999.

[40] Darren George and Paul Mallery. *IBM SPSS Statistics 25 Step by Step*. Routledge, October 2018. doi: 10.4324/9781351033909.

[41] Afshin Gholamy, Vladik Kreinovich, and Olga Kosheleva. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. In *Computer Science, Education, Mathematics*, 2018. URL https://api.semanticscholar.org/CorpusID:7467506.

[42] Camilo Ernesto Lopez Guarin, Elizabeth Leon Guzman, and Fabio A. Gonzalez. A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 10(3):119–125, August 2015. doi: 10.1109/rita.2015.2452632.

[43] Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. *Multivariate data analysis*. Prentice Hall, 7 edition, 2010.

[44] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0123814790.

[45] Brett R. Harris, Brianna M. Maher, and Leah Wentworth. Optimizing efforts to promote mental health on college and university campuses: Recommendations to facilitate usage of services, resources, and supports. *The Journal of Behavioral Health Services &; Research*, 49(2):252–258, January 2022. ISSN 1556-3308. doi: 10.1007/s11414-021-09780-2.

[46] Stefanie Hassel and Nathan Ridout. An investigation of first-year students' and lecturers' expectations of university education. *Frontiers in Psychology*, 8, January 2018. URL https://doi.org/10.3389/fpsyg.2017.02218.

[47] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2nd edition, 2009. doi: 10.1007/978-0-387-84858-7.

[48] James C. Hayton, David G. Allen, and Vida Scarpello. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods*, 7(2):191–205, April 2004. doi: 10.1177/1094428104263675.

[49] James Honaker, Gary King, and Matthew Blackwell. *AMELIA II: A Program for Missing Data*, 2011. URL https://cran.r-project.org/package=Amelia. Version 1.7.5.

[50] John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[51] Peter T. Hountras and Kenneth R. Brandt. Relation of student residence to academic performance in college. *The Journal of Educational Research*, 63(8):351–354, April 1970. ISSN 1940-0675. doi: 10.1080/00220671.1970.10884029.

[52] Shaobo Huang and Ning Fang. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, 61:133–145, February 2013. doi: 10.1016/j.compedu.2012.08.015.

[53] Faeeqa Jaffer and James Garraway. Understanding gaps between student and staff perceptions of university study in South Africa: A case study. *Journal of Student Affairs in Africa*, 4(1), June 2016. doi: 10.14426/jsaa.v4i1.145.

[54] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

[55] Henry F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, March 1974. doi: 10.1007/bf02291575.

[56] Annisa Uswatun Khasanah and Harwati. A comparative study to predict student's performance using educational data mining techniques. *IOP Conference Series: Materials Science and Engineering*, 215:012036, June 2017. doi: 10.1088/1757-899x/215/1/012036.

[57] Ralph Kimball and Margy Ross. The data warehouse toolkit: The definitive guide to dimensional modeling. In *Computer Science, Business*, 2013. URL https://api.semanticscholar.org/CorpusID:113497835.

[58] Rex B. Kline. *Principles and practice of structural equation modeling, fifth edition.* Methodology in the Social Sciences. Guilford Press, London, England, 5 edition, June 2023.

[59] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'95)*, pages 1137–1143. Morgan Kaufmann Publishers Inc., 1995.

[60] Zlatko J. Kovačić. Early prediction of student success: Mining students enrolment data. In *Proceedings of Informing Science & IT Education Conference*, volume 10, pages 647–665, 2010.

[61] Zlatko J. Kovačić and John Steven Green. Predictive working tool for early identification of 'at risk' students. Research report, School of Information and Social Sciences, Open Polytechnic, July 2010.

[62] George D. Kuh. What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning*, 35(2):24–32, March 2003. doi: 10.1080/00091380309604090.

[63] George D. Kuh, Jillian Kinzie, John H. Schuh, and Elizabeth J. Whitt. *Student Success in College: Creating Conditions that Matter*. Jossey-Bass, 2005.

[64] Max Kuhn. *caret: Classification and Regression Training*, 2023. URL https://cran.r-project.org/web/packages/caret/. R package version 6.0-94.

[65] Gloria Ladson-Billings. From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, 35(7):3–12, October 2006. ISSN 1935-102X. doi: 10.3102/0013189x035007003.

[66] Daniel T. Larose. *Discovering Knowledge in Data*. John Wiley & Sons, Inc., November 2004. doi: 10.1002/0471687545.

[67] Jehan Latief. Report on transformation at SA universities: It is about more than numbers - HSRC, Oct 2023. URL https://hsrc.ac.za/news/latest-news/report-on-transformation-at-sa-universities-it-is-about-more-than-numbers/.

[68] Rubén Daniel Ledesma and Pedro Valero-Mora. Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research, and Evaluation*, 12:1–11, 2007. doi: 10.7275/WJNC-NM63. URL https://scholarworks.umass.edu/pare/vol12/iss1/2/.

[69] Moeketsi Letseka and Mariette Visser. Student retention & graduate destination: higher education & labour market access & success. *Poverty, race and student*

achievement in seven higher education institutions, pages 25–40, 2010. URL http://repository.hsrc.ac.za/handle/20.500.11910/4385.

[70] Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data*. Wiley, Aug 2002. doi: 10.1002/9781119013563.

[71] Han Liu and Mihaela Cocea. Semi-random partitioning of data into training and test sets in granular computing context. *Granular Computing*, 2(4):357–386, August 2017. ISSN 2364-4974. doi: 10.1007/s41066-017-0049-2.

[72] Alf Lizzio. Fives senses of success: Designing effective orientation and engagement processes. Research report, Griffith University: First Year Experience Project, 2006.

[73] A. Lourens and D. Bleazard. Applying predictive analytics in identifying students at risk: A case study. *South African Journal of Higher Education*, 30(2), June 2016. doi: 10.20853/30-2-583.

[74] Tom Lowe and Yassein El Hakim, editors. *A handbook for student engagement in higher education*. SEDA Series. Routledge, London, England, March 2020.

[75] Ruth N. López Turley and Geoffrey Wodtke. College residence and academic performance: Who benefits from living on campus? *Urban Education*, 45(4):506–532, June 2010. ISSN 1552-8340. doi: 10.1177/0042085910372351.

[76] Marcia B. Baxter Magolda and Alexander W. Astin. What "doesn't" matter in college? *Educational Researcher*, 22(8):32, November 1993. ISSN 0013-189X. doi: 10.2307/1176821.

[77] C. Nel, C. Troskie de Bruin, and E. Bitzer. Students' transition from school to university: Possibilities for a pre-university intervention. *SAJHE*, 23(5):974–991, 2009. ISSN 1011-3487.

[78] Robert Nisbet, Ken Yale, and Gary Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier, 2018. doi: 10.1016/c2012-0-06451-4.

[79] Jason Osborne. Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation*, 8(6), 2002. doi: 10.7275/4VNG-5608. URL https://scholarworks.umass.edu/pare/vol8/iss1/6/.

[80] Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration*

and Policy in Mental Health and Mental Health Services Research, 42(5):533–544, November 2013. ISSN 1573-3289. doi: 10.1007/s10488-013-0528-y.

[81] Subethra. Pather. Social and academic integration of first-year at-risk students in a mathematics intervention programme. In *Proceedings of the Canada International Conference in Education*, pages 66–71, 2016.

[82] Subethra Pather and Elizabeth Booi. First-year undergraduate students' unmet university expectations and experience could influence academic performance: A South African university case study. In *ICERI2019 Proceedings*, 12th annual International Conference of Education, Research and Innovation, pages 3967–3974. IATED, 11-13 November, 2019 2019. ISBN 978-84-09-14755-7. doi: 10.21125/iceri.2019.0997.

[83] Subethra Pather and Nirmala Dorasamy. 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), July 2018. doi: 10.24085/jsaa.v6i1.3065.

[84] Denise F. Polit and Cheryl Tatano Beck. Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, 47 (11):1451–1458, November 2010. ISSN 0020-7489. doi: 10.1016/j.ijnurstu.2010.06.004.

[85] Amy Poots and Tony Cassidy. Academic expectation, self-compassion, psychological capital, social support and student wellbeing. *International Journal of Educational Research*, 99:101506, 2020. ISSN 0883-0355. doi: 10.1016/j.ijer.2019.101506.

[86] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In *Encyclopedia of Database Systems*, pages 532–538. Springer, 2009.

[87] RStudio. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020. URL http://www.rstudio.com/.

[88] SAS. *Getting Started with SAS Enterprise Miner 14.1®*. Cary, NC: SAS Institute Inc, July 2015. URL https://manualzz.com/doc/8812070/getting-started-with-sas-enterprise-miner-14.1-%C2%AE.

[89] Maggi Savin-Baden and Claire H. Major. *Qualitative research: The essential guide to theory and practice*. Routledge, London, 2013.

[90] Nancy K. Schlossberg. A model for analyzing human adaptation to transition. *The Counseling Psychologist*, 9(2):2–18, 1981. doi: 10.1177/001100008100900202.

[91] Lauren T. Schudde. The causal effect of campus residency on college student retention. *The Review of Higher Education*, 34(4):581–610, 2011. ISSN 1090-7009. doi: 10.1353/rhe.2011.0023.

[92] Ian Scott. Designing the South African higher education system for student success. *Journal of Student Affairs in Africa*, 6(1), July 2018. doi: 10.24085/jsaa.v6i1.3062.

[93] Amirah Mohamed Shahiri, Wahidah Husain, and Nur'aini Abdul Rashid. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422, 2015. doi: 10.1016/j.procs.2015.12.157.

[94] Tshimangadzo Daniel Sikhwari, Nkhangweleni Gloria Dama, Azwitamisi Milton Gadisi, and Tshifhiwa Christinah Matodzi. A comparative study of the academic performance of resident and non-resident students at a rural South African university. *Journal of Student Affairs in Africa*, 8(1), July 2020. ISSN 2307-6267. doi: 10.24085/jsaa.v8i1.3468.

[95] Craig Smith. A comprehensive first year engagement theory. *Journal of Access, Retention, and Inclusion in Higher Education*, 1(1):5, 2018. URL https://digitalcommons.wcupa.edu/jarihe/vol1/iss1/5.

[96] Susan Virginia Smith, Ruth Pickford, Janice Priestley, and Rebecca Sellers. Developing the inclusive course design tool: a tool to support staff reflection on their inclusive practice. *Compass*, 14(1), January 2021.

[97] J. Strydom, M. Mentz, and G. Kuh. Enhancing success in higher education by measuring student engagement in South Africa. *Acta Academica*, 42(1):259–278, 2010. URL https://journals.co.za/contentacadem/42/1/EJC15471.

[98] Barbara G. Tabachnick and Linda S. Fidell. *Using multivariate statistics*. Pearson, Upper Saddle River, NJ, 7 edition, Jul 2018.

[99] Mohsen Tavakol and Reg Dennick. Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2:53–55, June 2011. doi: 10.5116/ijme.4dfb.8dfd.

[100] Shao Kuang Ting. Predicting academic success of first-year engineering students from standardized test scores and psychosocial variables. *International Journal of Engineering Education*, 17(1):75–80, 2001.

[101] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, March 1975. doi: 10.3102/00346543045001089.

[102] Vincent Tinto. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press, 1993.

[103] Vincent Tinto. Taking retention seriously: Rethinking the first year of college. *NACADA Journal*, 19(2):5–9, September 1999. doi: 10.12930/0271-9517-19.2.5.

[104] Vincent Tinto. *Completing College: Rethinking Institutional Action*. University of Chicago Press, 2006.

[105] Gugu Wendy Tiroyabone and Francois Strydom. The development of academic advising to enable student success in South Africa. *The Academic Advising Issue*, 9(2):1–15, December 2021. ISSN 2311-1771. doi: 10.24085/jsaa.v9i2.3656.

[106] J.P. Vandamme, N. Meskens, and J.F. Superby. Predicting academic performance by data mining methods. *Education Economics*, 15(4):405–419, December 2007. doi: 10.1080/09645290701409939.

[107] Rüdiger Wirth. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.

[108] Brent D. Wolfe and Gregor Kay. Perceived impact of an outdoor orientation program for first-year university students. *Journal of Experiential Education*, 34(1):19–34, July 2011. doi: 10.1177/105382591103400103.

[109] Ziliang Wu, Wei Chen, Yuxin Ma, Tong Xu, Fan Yan, Lei Lv, Zhonghao Qian, and Jiazhi Xia. Explainable data transformation recommendation for automatic visualization. *Frontiers of Information Technology and Electronic Engineering*, Dec 2022. doi: 10.1631/fitee.2200409.

[110] Matthew J. Xerri, Katrina Radford, and Kate Shacklock. Student engagement in academic activities: a social support perspective. *Higher Education*, 75(4):589–605, June 2017. ISSN 1573-174X. doi: 10.1007/s10734-017-0162-9.

[111] David S. Yeager and Gregory M. Walton. Social-psychological interventions in education. *Review of Educational Research*, 81(2):267–301, June 2011. doi: 10.3102/0034654311405999.

[112] Mantz Yorke and Bernard Longden. The first-year experience of higher education in the uk: final report. In *Higher Education Academy*. Higher Education Academy, 2008.

[113] Ying Zhang, Samia Oussena, Tony Clark, and Hyeonsook Kim. Use data mining to improve student retention in higher education - a case study. In *Proceedings of the 12th International Conference on Enterprise Information Systems*, pages 190–197, June 2010.

Appendix A

Student Expectation and Experience Questions

Table A.1 provides a detailed list of all student expectation and experience statements and demographic information [81, 82].

Table A.1: Description of Student Expectation and Experience Datasets.

Field	**Attributes**
ID	Unique identifier or student number.
FirstGen	First generation student.
Computer	level of computer literacy.
Q1	I will join social organisations/clubs on campus this year, e.g. sports club, student union, etc.
Q2	I will attended many social functions, e.g. sports day, student braai, fund-raising events, etc.
Q3	Joining social clubs/organisations at university will distract me from my academic work.
Q4	I hope to make many new friends this year at university.
Q5	I will make many new friends from different racial groups.

Table A.2: Description of Student Expectation and Experience Datasets.

Field	Attributes
Q6	Making new friends on campus will help me to be successful in my studies.
Q7	I will be involved in academic discussions with my peers outside of formal lectures as it will enhance my learning.
Q8	I will make use of Librarians to help me find information for my assignments and projects.
Q9	I will make use of peer tutors to help me with my first-year courses at university.
Q10	In my course, I know what the in-class and out-of-class workload expected of me is.
Q11	The library is a place where I spent a lot of my time outside of formal class.
Q12	I will be comfortable seeking academic support from the institutional support services.
Q13	I will feel comfortable seeking academic support from my tutor/s.
Q14	I will feel comfortable seeking academic support directly from my lecturers.
Q15	I am aware of the role of academic support services, such as the writing centre, peer tutors and tutorials in helping me pass this year.
Q16	I expect that I will have to self-manage and take responsibility for my own learning at university.
Q17	My lecturers will expect me to write well-structured academic essays.
Q18	My lecturers will expect me to know how to correctly reference my assignments and projects.
Q19	My lecturers will expect me to attend all my lectures.

Table A.3: Description of Student Expectation and Experience Datasets.

Field	Attributes
Q20	I am quite resourceful and will be able to find information about university procedure and support on my own.
Q21	I will expect my lecturer/s to make referrals for me to get academic support if I need it.
Q22	I expect my lecturers to make themselves available outside of the formal lecture time to assist and advise me.
Q23	I will receive regular feedback from my lecturers in response to my assignments and tests.
Q24	I will be safe on campus.
Q25	The university cafeteria will sell affordable food.
Q26	I will be able to find my way around campus buildings.
Q27	I will have access to internet, computers and other resources to enhance my learning.
Q28	The university will be well sign-posted so I do not get lost.
Q29	The university cafeteria will sell healthy food.
Q30	Academic and support staff will be respectful and helpful.
Q31	My classmates will be supportive..
Q32	The university will care about me and my welfare
Q33	I was aware of the academic integrity and plagiarism requirements needed for assignments and tests.
Q34	I will be able to balance my first-year university study with other responsibilities.
Q35	Financial issues will distract me from my first-year studies.

Appendix B

Data Analysis Outputs

The following tables B.1, B.2, B.5 and B.6 provide a detailed list of all analysis outputs that could not be included in the main report but are quoted are used in the results and discussion section.

Table B.1: Summary Statistics of Student Expectation Profile

Questions	1:Strongly Disagree	2:Disagree	3:Neutral	4:Agree	5:Strongly Agree	**Mean**
Q1	18 (1%)	58 (3%)	344 (19%)	902 (50%)	478 (27%)	3.98
Q2	83 (5%)	310 (17%)	559 (31%)	528 (29%)	320 (18%)	3.38
Q3	23 (1%)	84 (5%)	547 (30%)	715 (40%)	431 (24%)	3.80
Q4	17 (1%)	64 (4%)	353 (20%)	848 (47%)	518 (29%)	3.99
Q5	32 (2%)	141 (8%)	546 (30%)	676 (38%)	405 (23%)	3.71
Q6	18 (1%)	70 (4%)	316 (18%)	890 (49%)	506 (28%)	4.00
Q7	7 (0%)	6 (0%)	94 (5%)	687 (38%)	1006 (56%)	4.49
Q8	6 (0%)	18 (1%)	144 (8%)	827 (46%)	805 (45%)	4.34
Q9	15 (1%)	39 (2%)	195 (11%)	836 (46%)	715 (40%)	4.22

Table B.2: Summary Statistics of Student Expectation Profile

Questions	1:Strongly Disagree	2:Disagree	3:Neutral	4:Agree	5:Strongly Agree	**Mean**
Q10	176 (10%)	329 (18%)	770 (43%)	342 (19%)	183 (10%)	3.02
Q11	121 (7%)	248 (14%)	782 (43%)	512 (28%)	137 (8%)	3.16
Q12	136 (8%)	409 (23%)	639 (36%)	481 (27%)	135 (8%)	3.04
Q13	45 (3%)	78 (4%)	377 (21%)	654 (36%)	646 (36%)	3.99
Q14	22 (1%)	43 (2%)	280 (16%)	732 (41%)	723 (40%)	4.16
Q15	48 (3%)	142 (8%)	624 (35%)	664 (37%)	322 (18%)	3.59
Q16	20 (1%)	29 (2%)	263 (15%)	837 (47%)	651 (36%)	4.15
Q17	21 (1%)	48 (3%)	314 (17%)	749 (42%)	668 (37%)	4.11
Q18	24 (1%)	49 (3%)	289 (16%)	770 (43%)	668 (37%)	4.12
Q19	43 (2%)	223 (12%)	472 (26%)	698 (39%)	364 (20%)	3.62
Q20	82 (5%)	206 (11%)	649 (36%)	626 (35%)	237 (13%)	3.41
Q21	17 (1%)	36 (2%)	353 (20%)	927 (52%)	467 (26%)	4.00
Q22	18 (1%)	75 (4%)	682 (38%)	743 (41%)	282 (16%)	3.66
Q23	40 (2%)	84 (5%)	669 (37%)	692 (38%)	315 (18%)	3.64
Q24	20 (1%)	59 (3%)	328 (18%)	758 (42%)	635 (35%)	4.07
Q25	22 (1%)	77 (4%)	492 (27%)	843 (47%)	366 (20%)	3.81
Q26	169 (9%)	358 (20%)	630 (35%)	363 (20%)	280 (16%)	3.13
Q27	12 (1%)	26 (1%)	137 (8%)	698 (39%)	927 (52%)	4.39
Q28	45 (3%)	188 (10%)	690 (38%)	613 (34%)	264 (15%)	3.48
Q29	14 (1%)	57 (3%)	339 (19%)	764 (42%)	626 (35%)	4.07
Q30	21 (1%)	92 (5%)	368 (20%)	795 (44%)	524 (29%)	3.95
Q31	17 (1%)	32 (2%)	388 (22%)	769 (43%)	594 (33%)	4.05
Q32	14 (1%)	30 (2%)	462 (26%)	710 (39%)	584 (32%)	4.01
Q33	307 (17%)	270 (15%)	535 (30%)	439 (24%)	249 (14%)	3.03
Q34	19 (1%)	42 (2%)	419 (23%)	837 (47%)	483 (27%)	3.96
Q35	13 (1%)	37 (2%)	345 (19%)	786 (44%)	619 (34%)	4.09

Table B.3: Summary Statistics of Student Experience Profile

Questions	1:Strongly Disagree	2:Disagree	3:Neutral	4:Agree	5:Strongly Agree	**Mean**
Q1	931 (52%)	411 (23%)	188 (10%)	131 (7%)	139 (8%)	1.96
Q2	726 (40%)	368 (20%)	358 (20%)	241 (13%)	107 (6%)	2.24
Q3	884 (49%)	367 (20%)	348 (19%)	145 (8%)	56 (3%)	1.96
Q4	113 (6%)	103 (6%)	344 (19%)	576 (32%)	664 (37%)	3.88
Q5	130 (7%)	165 (9%)	286 (16%)	579 (32%)	640 (36%)	3.80
Q6	87 (5%)	162 (9%)	482 (27%)	598 (33%)	471 (26%)	3.67
Q7	79 (4%)	115 (6%)	360 (20%)	689 (38%)	557 (31%)	3.85
Q8	470 (26%)	435 (24%)	311 (17%)	378 (21%)	206 (11%)	2.68
Q9	212 (12%)	221 (12%)	368 (20%)	637 (35%)	362 (20%)	3.40
Q10	77 (4%)	157 (9%)	542 (30%)	700 (39%)	324 (18%)	3.58
Q11	236 (13%)	343 (19%)	485 (27%)	469 (26%)	267 (15%)	3.10
Q12	423 (24%)	481 (27%)	578 (32%)	232 (13%)	86 (5%)	2.49
Q13	157 (9%)	187 (10%)	419 (23%)	616 (34%)	421 (23%)	3.53
Q14	148 (8%)	249 (14%)	586 (33%)	545 (30%)	272 (15%)	3.30
Q15	158 (9%)	129 (7%)	436 (24%)	727 (40%)	350 (19%)	3.55
Q16	26 (1%)	43 (2%)	206 (11%)	802 (45%)	723 (40%)	4.20
Q17	44 (2%)	66 (4%)	242 (13%)	630 (35%)	818 (45%)	4.17

Table B.4: Summary Statistics of Student Experience Profile

Questions	1:Strongly Disagree	2:Disagree	3:Neutral	4:Agree	5:Strongly Agree	**Mean**
Q18	35 (2%)	85 (5%)	242 (13%)	648 (36%)	790 (44%)	4.15
Q19	27 (2%)	28 (2%)	218 (12%)	507 (28%)	1020 (57%)	4.37
Q20	445 (25%)	457 (25%)	582 (32%)	217 (12%)	99 (6%)	2.48
Q21	57 (3%)	137 (8%)	613 (34%)	684 (38%)	309 (17%)	3.58
Q22	306 (17%)	293 (16%)	598 (33%)	457 (25%)	146 (8%)	2.91
Q23	68 (4%)	46 (3%)	408 (23%)	834 (46%)	444 (25%)	3.86
Q24	109 (6%)	142 (8%)	515 (29%)	655 (36%)	379 (21%)	3.59
Q25	74 (4%)	159 (9%)	658 (37%)	517 (29%)	392 (22%)	3.55
Q26	323 (18%)	275 (15%)	656 (36%)	413 (23%)	133 (7%)	2.87
Q27	42 (2%)	79 (4%)	440 (24%)	792 (44%)	447 (25%)	3.85
Q28	57 (3%)	44 (2%)	406 (23%)	714 (40%)	579 (32%)	3.95
Q29	141 (8%)	361 (20%)	576 (32%)	467 (26%)	255 (14%)	3.19
Q30	264 (15%)	325 (18%)	791 (44%)	330 (18%)	90 (5%)	2.81
Q31	116 (6%)	66 (4%)	679 (38%)	687 (38%)	252 (14%)	3.50
Q32	65 (4%)	163 (9%)	622 (35%)	614 (34%)	336 (19%)	3.55
Q33	202 (11%)	173 (10%)	818 (45%)	457 (25%)	150 (8%)	3.10
Q34	13 (1%)	2 (0%)	199 (11%)	702 (39%)	884 (49%)	4.36
Q35	67 (4%)	135 (8%)	471 (26%)	730 (41%)	397 (22%)	3.70

Table B.5: Student Expectation and Experience Gap Analysis

Questions	**Mean**	**Std Dev**	**Z-Score**
Q1 I will join social organisations/clubs on campus this year, e.g. sports club, student union, etc.	-2.016	1.500	-2.515
Q2 I will attend many social functions, e.g. sports day, student braai, fund-raising events, etc.	-1.143	1.673	-1.134
Q3 Joining social clubs/organizations at university will distract me from my academic work.	-1.847	1.420	-2.248
Q4 I hope to make many new friends this year at university.	-0.117	1.419	0.489
Q5 I will make many new friends from different racial groups.	0.085	1.556	0.809
Q6 Making new friends on campus will help me to be successful in my studies.	-0.329	1.346	0.154
Q7 I will be involved in academic discussions with my peers outside of formal lectures as it will enhance my learning.	-0.638	1.227	-0.336
Q8 I will make use of Librarians to help me find information for my assignments and projects.	-1.662	1.528	-1.956
Q9 I will make use of peer tutors to help me with my first-year courses at university.	-0.823	1.466	-0.627
Q10 In my course, I know what the in-class and out-of-class workload expected of me is.	0.561	1.504	1.562
Q11 The library is a place where I spent a lot of my time outside of formal class.	-0.060	1.562	0.580
Q12 I will be comfortable seeking academic support from the institutional support services.	-0.552	1.568	-0.198
Q13 I will feel comfortable seeking academic support from my tutor/s.	-0.456	1.537	-0.047
Q14 I will feel comfortable seeking academic support directly from my lecturers.	-0.859	1.372	-0.685
Q15 I am aware of the role of academic support services, such as the writing centre, peer tutors and tutorials in helping me pass this year.	-0.049	1.450	0.597

Table B.6: Student Expectation and Experience Gap Analysis

Questions	**Mean**	**Std Dev**	**Z-Score**
Q16 I expect that I will have to self-manage and take responsibility for my own learning at university.	0.046	1.160	0.747
Q17 My lecturers will expect me to write well-structured academic essays.	0.065	1.298	0.777
Q18 My lecturers will expect me to know how to correctly reference my assignments and projects.	0.036	1.295	0.731
Q19 My lecturers will expect me to attend all my lectures.	0.749	1.331	1.859
Q20 I am quite resourceful and will be able to find information about university procedure and support on my own.	-0.923	1.520	-0.786
Q21 I will expect my lecturer/s to make referrals for me to get academic support if I need it.	-0.411	1.217	0.024
Q22 I expect my lecturers to make themselves available outside of the formal lecture time to assist and advise me.	-0.751	1.444	-0.514
Q23 I will receive regular feedback from my lecturers in response to my assignments and tests.	0.212	1.277	1.010
Q24 I will be safe on campus.	-0.487	1.385	-0.096
Q25 The university cafeteria will sell affordable food	-0.256	1.330	0.270
Q26 I will be able to find my way around campus buildings.	-0.261	1.727	0.262
Q27 I will have access to internet, computers, and other resources to enhance my learning.	-0.544	1.174	-0.186
Q28 The university will be well sign-posted so I do not get lost.	0.473	1.358	1.423
Q29 The university cafeteria will sell healthy food.	-0.887	1.403	-0.729
Q30 Academic and support staff will be respectful and helpful.	-1.140	1.369	-1.129
Q31 My classmates will be supportive.	-0.554	1.282	-0.203
Q32 The university will care about me and my welfare	-0.459	1.265	-0.052
Q33 I was aware of the academic integrity and plagiarism requirements needed for assignments and tests.	0.071	1.601	0.786
Q34 I will be able to balance my first-year university study with other responsibilities.	0.399	1.082	1.306
Q35 Financial issues will distract me from my first-year studies.	-0.392	1.259	0.054

Appendix C

Ethical Clearance from the University of South Africa

COLLEGE OF ECONOMIC AND MANAGEMENT SCIENCE RESEARCH ETHICS REVIEW COMMITTEE

01 September 2022

NHREC Registration # : (if applicable)
ERC Reference #: 2021_CRERC_047(FA)
Name: Mrs Elizabeth Mmapholo Booi
Student #: 35009500

Dear Mrs Elizabeth Mmapholo Booi

Decision: Ethics Approval from 2022 to 2025

Researcher(s): Mrs Elizabeth Mmapholo Booi; 35009500@mylife.unisa.ac.za ; 066 303 3366
College of Economic and Management Sciences
Department of Decision Sciences
University of South Africa

Supervisor(s): Kathrine Mary Malan, malankm@unisa.ac.za , 012 433 4729
Dr MT MaseTshaba, Emasetmt@unisa.ac.za , 012 433 4732
College of Economic and Management Sciences
Department of Decision Sciences
University of South Africa

"Predicting Student Academic Performance Using First-Year Expectations and Experiences."

Qualification: Masters Degree

Thank you for the application for research ethics clearance by the Unisa College of Economic and management Sciences Research Ethics Review Committee for the above-mentioned research. Ethics approval is granted for 3 years, from **01 September 2022 until 31 August 2025**.

*The **low risk application** was **reviewed** by the College of Economic and management Sciences Research Ethics Review Committee on **24 November 2021** in compliance with the Unisa Policy on Research Ethics and the Standard Operating Procedure on Research Ethics Risk Assessment.*

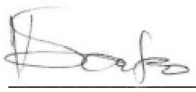
The proposed research may now commence with the provisions that:

1. The researcher(s) will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.
2. Any adverse circumstance arising in the undertaking of the research project that is relevant to the ethicality of the study should be communicated in writing to the College of Economic and management Sciences Research Ethics Review Committee.
3. The researcher(s) will conduct the study according to the methods and procedures set out in the approved application.
4. Any changes that can affect the study-related risks for the research participants, particularly in terms of assurances made with regards to the protection of participants' privacy and the confidentiality of the data, should be reported to the Committee in writing, accompanied by a progress report.
5. The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study. Adherence to the following South African legislation is important, if applicable: Protection of Personal Information Act, no 4 of 2013; Children's act no 38 of 2005 and the National Health Act, no 61 of 2003.
6. Only de-identified research data may be used for secondary research purposes in future on condition that the research objectives are similar to those of the original research. Secondary use of identifiable human research data requires additional ethics clearance.
7. No field work activities may continue after the expiry date **(31 August 2025)** Submission of a completed research ethics progress report will constitute an application for renewal of Ethics Research Committee approval.
8. Permission is to be obtained from the university from which the participants are to be drawn (the Unisa Senate Research, Innovation and Higher Degrees Committee) to ensure that the relevant authorities are aware of the scope of the research, and all conditions and procedures regarding access to staff/students for research purposes that may be required by the institution must be met.
9. If further counselling is required in some cases, the participants will be referred to appropriate support services.

Note:

*The reference number **2021_CRERC_047 (FA)** should be clearly indicated on all forms of communication with the intended research participants, as well as with the Committee.*

Yours sincerely,



Goonasagree Naidoo

Dr Vaola Sambo
Chairperson, CRERC
E-mail: Esambovt@unisa.ac.za
Tel: 012 429 4355

Prof Goonasagree Naidoo
Acting, Deputy Executive Dean: CEMS
E-mail: Naidog@unisa.ac.za
Tel: 012 429 6746

Appendix D

Ethical Clearance from University of the Western Cape

UNIVERSITY OF THE WESTERN CAPE
PERMISSION TO CONDUCT RESEARCH

DEAR **Elizabeth Booi**

This serves as acknowledgement that you have obtained and presented the necessary ethical clearance and your institutional permission required to proceed with the project referenced below:

RESEARCH TOPIC

Predicting student academic performance using first-year expectations and experiences

Name of researcher	:	Elizabeth Booi
Permission valid till	:	31 August 2025
Institution	:	University of South Africa
Ethics reference	:	2021_CRERC_047(FA)
Permission reference	:	UWCRP968878

You are required to engage this office (researchperm@uwc.ac.za) in advance if there is a need to continue with research outside of the stipulated period. The manner in which you conduct your research must be guided by the conditions set out in the annexed agreement: Conditions to guide research conducted at the University of the Western Cape.

Please be at liberty to contact this office should you require any assistance to conduct your research or require access to either staff or student contact information.

Regards
Dr Ahmed Shaikjee
Deputy Registrar Academic Administration

Approval status: **APPROVED** 8 September 2022

To verify or confirm the authenticity of this document please contact the University at researchperm@uwc.ac.za.