

https://dspace.org/

Predicting student success using student engagement in the online component of a blended-learning course

Buraimoh, Eluwumi Folake

2021
https://hdl.handle.net/10539/32382

Downloaded from a DSpace repository. https://dspace.org/

Predicting Student Success Using Student Engagement in the Online Component of a Blended-Learning Course

By

Eluwumi Folake Buraimoh

(2287804)

Master of Science

Research-Report submitted in partial fulfilment for the degree of M.Sc (Coursework and Research Report) Computer Science



School of Computer Science and Applied Mathematics

Faculty of Science

The University of the Witwatersrand, Johannesburg

South Africa

Supervised by

Dr. Ritesh Ajoodha &

Dr. Kershree Padayachee

20th May, 2021

Declaration

I, Eluwumi Folake Buraimoh with student number 2287804, hereby declare that the content of this research report are my own work unless otherwise explicitly referenced. This research report is submitted in partial fulfilment for the degree of Master of Science at the University of the Witwatersrand, Johannesburg, and has not been submitted to any other university, nor for any other degree.

Signed:

Abstract

There has been a surge in student failure rates in blended-learning courses in recent times, which has generated considerable research interests. Engagement is identified as one of the core metrics for measuring students' success or failure in any learning system. This study utilizes machine learning algorithms on students' log-file data collected from an LMS to predict student success and increase their throughput rates.

The machine learning predictive models considered in this study are Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Tree, Random Forests, Gradient Boosting Tree, Multilayer Perceptron Neural Network, and Linear Discriminant Analysis. The study presents the advantage of using SMOTE sampling in handling imbalance class problems over Random Under-Sampling and Random Over-Sampling Techniques.

The Random Forests performance surpassed the other machine learning models in this study with an accuracy value of 91%, AUC of 0.90, and F1-score of 0.98. The results provide an automatic predictive model for timely identification of learners at risk of failing in their courses for instructor early intervention. The significance of this study is to provide a feedback tool on engagement for an increase in student performance.

Keywords: Educational Data Mining, Student Success, Machine Learning, Student Engagement, Predictive Models.

Publications

Some of the work contained in this research report appears in peer-reviewed publications:

- **Eluwumi Buraimoh, Ritesh Ajoodha and Kershree Padayachee**. "Importance of Data Re-Sampling and Dimensionality Reduction in Predicting Students' Success." The 3rd International Conference on Electrical, Communication, and Computer Engineering (ICECCE). IEEE, 2021.
- **Eluwumi Buraimoh, Ritesh Ajoodha, and Kershree Padayachee**. "Application of Machine Learning Techniques to the Prediction of Student Success." 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). IEEE, 2021.
- **Eluwumi Buraimoh, Ritesh Ajoodha and Kershree Padayachee**."Prediction of Student Success using Student Engagement with Learning Management System." International conference on Interdisciplinary Research in Technology and Management, Kolkata, India, 2021.

Acknowledgements

Throughout my study and research report writing, I have received a great deal of support and assistance. I want to acknowledge all the people that made this research report possible.

First of all, I would like to acknowledge my research supervisors, Dr. Ritesh Ajoodha and Dr. Kershree Padayachee, for their priceless advice and guidance throughout the research and writing process. I appreciate their detailed supervisory roles and dedication.

My gratitude goes to my coursework lecturers, Prof. Mumtaz Ali, Prof. Turgay Celik, Prof. Rosman Benjamin, Dr. Hima Vadapalli, Dr. Pravesh Ranchod, Dr. Hairong Bau, Dr. Dmitri Shaktov, and Mr. Steve James. I acknowledge their relentless efforts in impacting positively on me.

I appreciate Ms. Kgomotso Monyepote, Ms. Mpumi Mnaqapu, Mr. Mojalefa Malahlela, and Mr. Brian Maistry for the valuable help and swift responses to my requests during this research and my study.

My sincere appreciation also goes to Temmy, Moyosoreoluwa, and Oluwadarasimi. There is no way I would have accomplished this feat without you guys. I honestly could not be more grateful for all your love, sacrifices, and support both morally and financially. I can not love you guys less; you are my superheroes. The best is yet to come!

My profound gratitude to my friends and support systems for their immeasurable contributions to this study's successful completion: God'sgift Uzor, Benisemeni Zakka, Grace Mabele, Melusi Moyo, Nurcia Mouneyi, Patrick Tshiaba, Jeremiah Olajuwon, and Nouralden Mohammed. Meeting you guys was a blessing, and I do not take your support for granted.

Finally, my heartfelt gratitude to my parents (Mr. and Mrs. Adetoyese Buraimoh), Mrs. Margaret Nihi, and my siblings (Elutope, Abimbola, Babatunde, and Elutunji) for their understanding, kindness, and encouragement during my study.

Contents

Declaration i

Abstract ii

Publications iii

Acknowledgements iv

Contents v

List of Figures viii

List of Tables ix

List of Abbreviations x

1 Introduction 1
- 1.1 Problem Statement . . . 2
- 1.2 Motivation of Study . . . 2
- 1.3 Research Contribution . . . 3
- 1.4 Purpose Statement . . . 3
- 1.5 Aims and Objectives . . . 3
  - 1.5.1 Study Aim . . . 3
  - 1.5.2 Objectives . . . 3
  - 1.5.3 Research Questions . . . 4
- 1.6 Summary . . . 4

2 Literature Review 5
- 2.1 Introduction . . . 5
  - 2.1.1 Impacts of Student Engagement on Performance . . . 6
  - 2.1.2 Engagement Detection Classification . . . 6
  - 2.1.3 Log File Analysis . . . 6
- 2.2 Data Source . . . 7
- 2.3 Feature Selection . . . 7
  - 2.3.1 Feature Selection in Student Success Prediction . . . 8
- 2.4 Predictive Models in Educational Data Mining . . . 8
- 2.5 Evaluation Techniques . . . 10

2.6 Conclusion . . . 11

3 Research Methodology **12**
3.1 Application of Study . . . 12
3.2 Data Acquisition and Collection . . . 13
3.3 Pre-processing and Features Selection . . . 13
3.3.1 Missing Values . . . 14
3.3.2 Data Transformation . . . 14
3.3.3 Data Sampling . . . 15
3.3.4 Feature Selection . . . 15
3.3.5 Libraries, Platforms, Frameworks, and Language . . . 15
3.4 Predictive Models . . . 16
3.4.1 Logistic Regression . . . 16
3.4.2 Support Vector Machines . . . 17
3.4.3 Naïve Bayes Classifier . . . 18
3.4.4 Decision Tree . . . 19
3.4.5 Random Forests . . . 20
3.4.6 Gradient Boosting Tree . . . 20
3.4.7 Multilayer Perceptron . . . 20
3.4.8 Linear Discriminant Analysis . . . 20
3.5 Evaluation Metrics . . . 21
3.6 Summary . . . 22

4 Results and Discussion **23**
4.1 Introduction . . . 23
4.2 Dataset Description . . . 23
4.3 Information Gain and Feature Importance . . . 24
4.3.1 Information Gain . . . 24
4.3.2 Feature Importance . . . 25
4.4 Classification Results with three Sampling Techniques and Imbalanced Dataset . . . 27
4.4.1 Random Under-Sampling Technique with and without Feature Selection . . . 27
4.4.2 Random Over-Sampling Technique with and without Feature Selection . . . 28
4.4.3 SMOTE Technique with and without Feature Selection . . . 29
4.4.4 Classification on Imbalanced Data with Feature Selection and without Feature Selection . . . 30
4.4.5 Feature Selection using the top 5 and 10 Feature Selection with the Re-sampling Techniques . . . 31
4.5 Model Evaluation with Confusion Matrix,Accuracy Precision, Recall, F1-Score,AUC and ROC curve . . . 31
4.5.1 Confusion Matrix . . . 32
4.5.2 Accuracy . . . 33
4.5.3 Precision . . . 34
4.5.4 Recall . . . 35
4.5.5 F1 Score . . . 35
4.5.6 Area Under Curve (AUC) . . . 36

4.5.7 Receiver Operating Characteristic (ROC) 36
4.6 Discussion 38
4.7 Summary 42

5 Conclusion, Further Study, Limitations, and Contribution **44**
5.1 Conclusion 44
5.2 Future Study Recommendation 45
5.3 Limitation 45
5.4 Contributions 46

Bibliography **54**

List of Figures

3.1 Student Success Prediction Flow 13
3.2 Data Transformation using Label Encoding 15
3.3 Logistic Regression Graphical Representation 17
3.4 SVM Hyperplane 18
3.5 Naïve Bayes Graphical Representation 18
3.6 A Graphical Representation of Decision Tree 19

4.1 Data Distribution After Dropping the Missing Values 24
4.2 A Graphical representation of the Information Gain for Set of Attributes to Predict Student Success 26
4.3 Random Forests Feature Importance 26
4.4 Correlation Matrix for Numerical Features 27
4.5 Algorithm Comparison by Training Set 30
4.6 Logistic Regression ROC 37
4.7 Support Vector Machines ROC 37
4.8 Naïve Bayes ROC 38
4.9 Decision Tree ROC 38
4.10 Random Forests ROC 39
4.11 Gradient Boosting Trees ROC 39
4.12 Multilayer Perceptron ROC 40
4.13 Linear Discriminant Analysis ROC 40

List of Tables

2.1 Related Work on Student Success Prediction based on Engagement in Online and Blended Learning Environments . . . 9
2.2 Confusion Matrix . . . 10

3.1 The Students' Attributes Classification. . . . 14

4.1 Features in the Dataset . . . 24
4.2 The Information Gain Ranking for a Set of Features to Predict the Student's Success . . . 25
4.3 Predictive Model Accuracies after 10-fold Cross-Validation for Random Under-Sampling Method . . . 28
4.4 Predictive Model Accuracies after 10-fold Cross-Validation for Random Over-Sampling Method . . . 29
4.5 Predictive Model Accuracies after 10-fold Cross-Validation for SMOTE Method . . . 30
4.6 Predictive Model Accuracies after 10-fold Cross-Validation on the Imbalanced Class . . . 31
4.7 Comparison of the top 5 and 10 Feature Selection with the Re-sampling Techniques . . . 32
4.8 **Logistic Regression** Predictive Model's Confusion Matrix . . . 32
4.9 **Support Vector Machines** Predictive Model's Confusion Matrix . . . 32
4.10 **Naïve Bayes** Predictive Model's Confusion Matrix . . . 33
4.11 **Decision Tree** Predictive Model's Confusion Matrix . . . 33
4.12 **Random Forests** Predictive Model's Confusion Matrix . . . 33
4.13 **Gradient Boosting** Predictive Model's Confusion Matrix . . . 33
4.14 **Multilayer Perceptron** Predictive Model's Confusion Matrix . . . 33
4.15 **Linear Discriminant Analysis** Predictive Model's Confusion Matrix . . . 34
4.16 Accuracy Values for the Predictive Models . . . 34
4.17 Precision Table for the Predictive Models . . . 35
4.18 Recall Table for the Predictive Models . . . 35
4.19 F1-Score Table for the Predictive Models . . . 36
4.20 AUC Score for the Predictive Models . . . 36
4.21 Summary of the Evaluation Metrics for Predictive Models . . . 42

List of Abbreviations

ANN	Artificial Neural Network
AUC	Area Under Curve
BC	Bayesian Classifier
BN	Bayesian Network
CART	Classification and Regression Tree
DNN	Deep Neural Network
DT	Decision Tree
EDDA	Extreme-scale Distribution-based Data Analysis
HPSO	Hybrid Particle Swarm Optimisation
IBk	Instance Based Learner
KNN	K−Nearest Neighbour
LDA	Linear Discriminant Analysis
LR	Logistic Regression
NB	Naïve Bayes
NN	Neural Network
PSO	Particle Swarm Optimisation
RF	Random Forests
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SMO	Sequential Minimal Optimisation
SVM	Support Vector Machines

Chapter 1

Introduction

Educational Data Mining (EDM) is a prominent study area, serving a variety of instructional goals within web-based educational systems [Romero *et al.*, 2014]. The purpose of EDM includes evaluation of learning and instructional design efficacy, designing adaptive environments for students based on their actual behaviour, providing input to both students and teachers, and detecting abnormal learning habits in the system [Castro *et al.*, 2007]. EDM is a data mining subset that utilises machine learning methods in various kinds of academic data to address academic research concerns [Romero and Ventura, 2013]. Academic success is one such concern for tertiary education institutions around the world.

Data analytics is identified for detecting students' academic failure and increasing the throughput rate in educational institutions around the globe [Ajoodha *et al.*, 2020; Kumar *et al.*, 2018].

Student engagement is the student's learning involvement or participation in a learning environment [Hu and Li, 2017]. The concept of student engagement is derived from conventional classroom teaching studies [Hu and Li, 2017]. Though online, blended, and classroom learning processes are different, student engagement theory is the same. That is, they all need students' active involvement [Hu and Li, 2017].

The blended course is a mix of the traditional classroom and online learning [Porter *et al.*, 2014; Margulieux, 2015]. The possible advantages of using technology to complement instructors include improving instruction quality and accessibility of instruction [Margulieux, 2015]. The blended-learning objectives include active learning, accessibility to information at a convenient time, student-teacher physical interaction, improvement in academic performance, learner's independence and improved cost-efficiency [Osguthorpe and Graham, 2003].

The objectives of using the blended-learning method rely on the teacher, learners, curriculum, and school, but adequately implemented technology-based instruction should not adversely impact learning outcomes [Margulieux, 2015]. A large number of learners are participants in technology-mediated learning in recent times because it is a convenient way of learning [Hu and Li, 2017], and the present COVID-19 pandemic has necessitated it.

However, the blended or online learning platform is faced with numerous challenges. Some of the challenges associated with the students in a technology-mediated environment are lack of self-drive in students towards various course materials and activities, increase in the rate of course non-completion and high attrition rate [Hussain *et al.*, 2018; Asiah *et al.*, 2019; Moubayed *et al.*, 2018]. The other challenge is the technical challenges such as lack of automatic detection of student engagement or disengagement, lack of efficiency and accurate prediction model.

Despite the challenges with blended and online learning platforms, the blended-learning environment has dramatically reduced the difficulty of student engagement detection experienced in online learning. According to the study of Owston and York [2018], some scholars agreed that students perform decently well in blended programs than those exclusively in technology-mediated or classroom courses across a broad spectrum of academic subjects and educational offerings.

Student engagement is a core metric for measuring a student's success or failure in any learning system. Recently, the reported rate of failure in undergraduate blended-learning courses is alarming, especially in science courses [Macarini *et al.*, 2019]. Thus, there is a need to investigate the key engagement metrics correlated with a student's success in a blended-learning environment.

Student engagement is important to learning but multi-faceted and complex [Bosch, 2016]. The time allotted to online activities in a blended-learning course is proportional to the course performance [Gómez-Aguilar *et al.*, 2015; Romero *et al.*, 2013]. Hence, the study of the online component of blended-learning.

1.1 Problem Statement

Student engagement is the main indicator for measuring students' performance in any learning system. Evaluating student performance in blended-learning using their interaction with the Learning Management System (LMS) is essential in determining the reason behind high failure rates in undergraduate courses, especially in science. The evaluation result will help the instructors identify students at risk of failing and need intervention to avert failure. To this end, this study looks at predicting student success in a blended course using student engagement data obtained from LMS to research key engagement metrics vital for the prediction of students' success. In this research, the term "at-risk " means students on the low-level band on a scale of 100 percent.

1.2 Motivation of Study

This research is motivated by the need to increase the throughput rates of students in a blended-learning course. Some of the earlier studies classified students' performances into outstanding, average, or below average [Asiah *et al.*, 2019]. The students in the bracket of below-average are eventually identified to fail the course. Hussain *et al.* [2018] also emphasised in their research that excellent learning can be achieved by tracking student engagement in an educational programme through different practices, which ultimately helps minimise dropout rates. This study seeks to investigate the vital student engagement metrics that affect students' success in a course.

1.3 Research Contribution

The contributions of this research are:

- A predictive model to determine which features are vital to predict student success in a blended-learning course.
- Feedback tool on learning for possible redesigning of course to increase throughput in a blended-learning course.
- Automatic predictive model for early identification of at-risk students for timely instructor intervention.

1.4 Purpose Statement

This research evaluates the effect of student engagement with the LMS on the students' performance. It will also compare and evaluate the more influencing interactions in predicting the students' success in a blended course.

1.5 Aims and Objectives

1.5.1 Study Aim

This study aims to predict student success by using stored data containing student activities and evaluating the behavioural patterns using student engagement on the LMS.

1.5.2 Objectives

To achieve this aim, the following objectives are to be met:

- To determine and evaluate the vital engagement metrics on student's performances.
- To provide the instructors or lecturers with automatic detectors to identify a student at risk of failing during courses for possible intervention.
- To provide a machine learning model with optimal performance for the prediction of student's success.
- To investigate the impact of student's engagement on their course assessment scores.
- To evaluate the effect of the data sampling methods on the performance of the models in predicting student success.

1.5.3 Research Questions

The research questions for this study are:

- What are the key engagement metrics for early detection of students at risk of failing for the instructors' hypothetical timely intervention in the blended-learning course?
- Which machine learning models used in this study offer optimum performance in predicting student success in the blended-learning course?
- Which engagement features are vital to predicting students' performance in a blended-learning course?
- Which of the data sampling techniques used in this study give the best model performance in predicting student success?

1.6 Summary

This chapter presented the introduction of the study. It also discussed the problem statement, motivation of study, research contribution, statement of purpose and research questions. The purpose of this study is to obtain information that could be used to identify students at risk of failing and provide the instructor's hypothetical intervention to avert failure and boost performance.

This research report is structured as follows: Chapter 2 provides the related work on various researches done in the past on student success predictions in a blended-learning environment, Chapter 3 outlines the research design and methodology associated with the student success predictions based on engagement using eight predictive models: Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Trees, Random Forest, Gradient Boosting Tree, Multi-layer Perceptron, and Linear Discriminant Analysis. Chapter 4 contains the results and discussion, and Chapter 5 is the conclusion, further study, limitation, and contribution.

Chapter 2

Literature Review

2.1 Introduction

Researchers have attempted to study student success using engagement in blended-learning and online learning environments [Hussain *et al.*, 2018; Amrieh *et al.*, 2016; Van Goidsenhoven *et al.*, 2020]. The correlation between students' performance and engagement has also been extensively investigated using data mining, statistical analysis, and machine learning [Dutt and Ismail, 2019; Hussain *et al.*, 2018; Zacharis, 2016]. Studies have shown that educational achievement has been correlated with LMS engagement metrics and has a strong relationship with their performance in the programme [Macarini *et al.*, 2019; Hussain *et al.*, 2018; Amrieh *et al.*, 2015, 2016].

Hussain *et al.* [2018] and Zacharis [2016] in their research used the number of clicks on educational learning platform as one of the significant student interaction features in a technology-mediated/blended-learning environment in the investigation of student success. Decision Tree, Naïve Bayes Classifier, and Neural Networks algorithms were also considered in their research due to the simplicity and high predictive accuracies of these algorithms.

For blended-learning success prediction, the interaction or engagement of the student with the LMS, which is the online component of the blended-learning course, was explored in the study by [Sheshadri *et al.*, 2019]. Their result showed that the online part of blended learning's engagement correlates with student performance and final grades. New studies suggest that blended learning can improve student success[Vo *et al.*, 2017; Bernard *et al.*, 2014].

This section is structured as follows: Section 2.1.1 discusses the impacts of student engagement on performance; Section 2.1.2 shows the classification of engagement detection; and finally, Section 2.1.3 explores log-file analysis for predicting student success.

2.1.1 Impacts of Student Engagement on Performance

The impact of student engagement is broad on student performance. Student engagement stimulates factors such as students' learning experience, learning satisfaction, learning performance, and competency in the course [Soffer and Cohen, 2019]. Wefald and Downey [2009] showed the positive correlation between student engagement with learning performance, promotion of learning outcomes, an increase in professional maturity, and a reduction in dropout propensity.

Hu *et al.* [2016] examined the role of learning engagement in technology-mediated learning, and their result showed a positive impact on learning effect in either classroom learning or technology-driven learning. Essentially, when students are well engaged in learning, they are bound to gain more in their learning activities Soffer and Cohen [2019] and ultimately succeed in the course.

2.1.2 Engagement Detection Classification

Dewan *et al.* [2019] proposed classifications for engagement detection from the existing approach for online learning learners. There are three major types: automatic, semi-automatic, and manual. These classifications depend on the method and the level of user participation in the engagement identification procedure.

In a manual method of engagement, detection requires learners' direct participation in the process of engagement detection. The manual category is broken down into a self-reporting and an observational checklist.

Engagement tracing is a sub-division of semi-automatic engagement detection, where the learner's indirect involvement in the interaction detection process is required. Engagement tracing records the frequency and precision of learner answers to problem practice and questions.

In the automatic category, the system automatically records the learners' activities in the interaction detection process and does not interrupt the learners. The automatic method sub-divisions are "computer vision-based type, sensor data analysis, and log-file analysis", subject to the data these interaction detection systems capture.

2.1.3 Log File Analysis

For the predictions of student success, the behaviours of the learners stored in log files are analysed. In a technology-mediated learning setting, the students' activities are recorded in logs, and this can provide useful insight for the identification of the engagement towards performance prediction [Dewan *et al.*, 2019].

For example, an LMS can store a user's learning history, including the history of browsing, assignment submissions, frequency of login, test results, and other essential events [Wellman *et al.*, 2014]. Sakai is an example of an open-source LMS, which is regarded as a course management system within the open-source culture. Sakai is an accessible learning resource used in most institutions of higher learning [Cavus and Zabadi, 2014]. There are commercial and open-source LMSs. Kalboard 360 is a cloud-based example of commercial LMS that allows a blended-learning form of pedagogy.

2.2 Data Source

Insufficient consideration has been devoted to sources of data for supervised classification analysis [Gardner and Brooks, 2018]. The knowledge of the sources of data is vital for predictions. It provides valuable foundations for potential research that are not explored. It also reduces developmental and computational times for feature extraction [Gardner and Brooks, 2018; Mining, 2012].

Different data sources had been explored in the past for success predictions, as evident in other researchers' past works. Some researchers used qualitative data (questionnaire or interview), quantitative data (student behavioural data from online learning activities), and others used combinations of both data types [Hu and Li, 2017].

Wells *et al.* [2016] explored log-file data of postgraduate students in an online management program at Imperial Business School, London, for the 2014/2015 session. Hussain *et al.* [2018] used freely accessible data collection from the Open University of the United Kingdom in their study. In their research on the predictions of success, Soffer and Cohen [2019] used four online courses from the Moodle LMS log-file data of the undergraduate students at Tel Aviv University, Israel.

Captured images from videos were developed and used for engagement recognition in relation to student performance by [Kamath *et al.*, 2016]. Moubayed *et al.* [2018] investigated the connection between student engagement and academic success in a technology-mediated platform using the LMS data from a North American university undergraduate science course. Alsubhi *et al.* [2019] explored interview questions for student engagement challenges in e-learning platforms at different Saudi universities and their relationship with student performance. Silva *et al.* [2016] applied Moodle LMS data from graduate courses in public management course at the university in Brazil from 2014 to 2015.

Measurement of student disengagement was conducted by Haiyang *et al.* [2018] using data from Open University. Inventory answers and recorded video data for students in Stanford statistical Massive Open Online Course was used by Mongkhonvanit *et al.* [2019] for student engagement tracing.

2.3 Feature Selection

Feature selection is the key element in any successful machine learning project [Domingos, 2012]. Feature selection involves selecting or creating new features in the data set for an improved result of the prediction model and reducing computation time [Deepika and Sathyanarayana, 2019]. Also, feature selection controls the size of a dataset and removes redundant or irrelevant features [Deepika and Sathyanarayana, 2019]. Ramaswami and Bhaskaran [2009] identified some of the feature selection filters commonly used in EDM as follows:
1) Correlation Based Attribute evaluation
2) Chi Square Attribute evaluation
3) Gain Ratio Attribute evaluation
4) Information Gain Attribute evaluation
5) Relief Attribute evaluation and

6) Symmetrical Uncertainty Attribute evaluation

This section is structured as follows: Section 2.3.1 explores feature selections in student success prediction in the past work.

2.3.1 Feature Selection in Student Success Prediction

The investigation of the student success using engagement in online learning using four variables: initial assessment scores, the highest level of education, final examination score, and the total number of clicks on the learning platform was performed by [Hussain *et al.*, 2018]. Student clicks on nine online learning platforms, final grade and assessment score were highly significant to engagement from spearman statistical results. Their findings also showed that learners' clicks on forumng and oucontent are significant in predicting student success. The forum discussion and course content access activities positively impact student engagement and examination final grade.

The association between student involvement and success in the technology-mediated setting was explored using nine engagement metrics with association rule from learners' event logs in the study of [Moubayed *et al.*, 2018]. It was revealed in the researchers' study that student attributes like the frequency of logins, content read, and the number of forum read influenced the quiz performance and eventually resulted in a high final score in the course. Moubayed *et al.* [2018] proposed that student involvement can be a determinant of academic success due to the positive correlation engagement has on performance.

2.4 Predictive Models in Educational Data Mining

A considerable number of predictive models or machine learning algorithms have been implemented on educational data in past studies to evaluate students' success. Authors have investigated student success based on engagement in blended learning and online environments using features extracted from students' activities in the pedagogy's online component. Table 2.1 highlights the summary of predictive models, features, and evaluation metrics from the related work.

The Decision Tree model is a frequently utilised predictive model in EDM. Hashim *et al.* [2020]; Waheed *et al.* [2020]; Hussain *et al.* [2018]; Amrieh *et al.* [2016] and Cocea and Weibelzahl [2009] applied the Decision Tree model in their studies of student performance, and the highest accuracy value of 82% was obtained from the Decision Tree by [Amrieh *et al.*, 2016].

Naïve Bayes Classifier was also implemented by Hashim *et al.* [2020]; Hussain *et al.* [2018]; Abed *et al.* [2020] and GopalaKrishnan and Sengottuvelan [2016] in their studies of student performance. Abed *et al.* [2020] had a good accuracy value for the Naïve Bayes Classifier used in their study. The Logistic Regression is an extensively used model in EDM (Waheed *et al.* [2020]; Hashim *et al.* [2020]; Bote-Lorenzo and Gómez-Sánchez [2017]; Silva *et al.* [2016]; GopalaKrishnan and Sengottuvelan [2016]; Sinha *et al.* [2014]; Cocea and Weibelzahl [2009]). The authors in Hashim *et al.* [2020] and Silva

et al. [2016] achieved the optimum accuracy value for Logistic Regression among the algorithms experimented in their researches.

The other models from the literature are : Support Vector Machines (Waheed *et al.* [2020]; Hashim *et al.* [2020]; Dutt and Ismail [2019]; Bote-Lorenzo and Gómez-Sánchez [2017]; Ajoodha *et al.* [2020] and Kamath *et al.* [2016]); and Random Forests (Ajoodha *et al.* [2020] and Dutt and Ismail [2019]), the best accuracies were recorded by Ajoodha *et al.* [2020] and Dutt and Ismail [2019] for Random Forest in their studies.

Authors	Features	Models	Evaluation Metrics
Waheed *et al.* [2020]	Students demographics,clickstream behaviour and assessment performance	ANN, LR and SVM	Accuracy,Precision and Recall
Hashim *et al.* [2020]	Number of students, study year, gender, students' birth year,registration, employment examination points and final points	SMO, DT, NN, NB, LR, KNN and SVM	Precision and Recall
Dutt and Ismail [2019]	Demographic, Academic and behavioral features	SVM, LDA,RF,KNN, and CART	Accuracy and F-measure
Alsubhi *et al.* [2019]	Online discussion,assessment,exercises, learning material	Qualitative approach	Expert view
Mongkhonvanit *et al.* [2019]	Item responses ,video interaction	RNN	Accuracy
Hussain *et al.* [2018]	Highest education level,Final results,test score,Number of clicks	Tree-based Algorithms, Naïve Bayes	Accuracy,Kappa,Recall
Moubayed *et al.* [2018]	Student ID, Quiz,Midterm Exam,Final Exam	Apriori Algorithm, Frequent Pattern (FP) Growth Algorithm, Generalisation Sequential Pattern Algorithm	Support,Confidence,Lift
Alshabandar *et al.* [2018]	Clicks on URL,sum_clicks.subpage	EDDA,KNN,LR	Accuracy,F1,AUC, sensitivity,specificity
Kaur *et al.* [2018]	Video recording	RF, DNN	Mean Square Error, Pearson Correlation Coefficient
Bote-Lorenzo and Gómez-Sánchez [2017]	Engagement Videos, assignment scores	LR, SGD ,RF,SVM	AUC
Wells *et al.* [2016]	Gender,Region, Highest qualification, Employment sector	LR	Correlation
Silva *et al.* [2016]	Dialogue,Autonomy, Profile Data	LR	Accuracy
Kamath *et al.* [2016]	Video recording	SVM	Accuracy
GopalaKrishnan and Sengottuvelan [2016]	Pages read,Learning time,Questions attended,Assessment time,Correct answers,Wrong answers,Clicks used,Scroll wheel used	LR,DT,BC,PSO with BCr, HPSO with NBC	Accuracy ,Precision
Sinha *et al.* [2014]	Clicks	LR	Accuracy
Cocea and Weibelzahl [2009]	Clicks	BN,LR,J48,CART, IBk	Accuracy,Precision

TABLE 2.1: Related Work on Student Success Prediction based on Engagement in Online and Blended Learning Environments

2.5 Evaluation Techniques

Different authors evaluated the performance of predictive models in their studies of student success prediction by comparing actual and predicted values. The Confusion Matrix represented in Table 2.2 holds the value of the information produced by a predictive model, about expected and actual class labels [Hussain *et al.*, 2018]. The terms in the confusion matrix are explained below:

- True positive (TP): Number of correct positive predictions.
- False positive (FP): Number of incorrect positive predictions.
- True negative (TN): Number of correct negative predictions.
- False negative (FN): Number of incorrect negative predictions.

	Predicted	
Actual	TP	FN
	FP	TN

TABLE 2.2: Confusion Matrix

There are different studies of student success prediction using engagement where accuracy evaluation technique has been used to test models performance Hussain *et al.* [2018]; Silva *et al.* [2016]; Durgabai and Bhushan [2014]; Cocea and Weibelzahl [2009]; Sinha *et al.* [2014]; Mongkhonvanit *et al.* [2019]; Kamath *et al.* [2016]; GopalaKrishnan and Sengottuvelan [2016]. Accuracy is a ratio of correct prediction.

Cocea and Weibelzahl [2009]; GopalaKrishnan and Sengottuvelan [2016] evaluated their models of student engagement predictions using precision evaluation method. Precision is the number of accurate positive predictions as a percentage of the total number of positive predictions.

Recall or sensitivity evaluation technique is the number of positive results right and the average number of positives. Hussain *et al.* [2018]; Alshabandar *et al.* [2018] tested their models using recall in their studies of student performance using engagement.

Kappa statistics, which is a metric that compares the observed accuracy with an expected accuracy, was explored by Hussain *et al.* [2018] in testing the performance of their model of study success prediction.

An important metric is the Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curve in the domain of prediction of student success. Hussain *et al.* [2018]; Alshabandar *et al.* [2018]; Bote-Lorenzo and Gómez-Sánchez [2017] evaluated their model using this metric.

F1- Measure, which is the harmonic mean of precision and recall, is another indicator for evaluating feature selection techniques' effectiveness. This metric is more important for evaluating the efficiency of the classifier. Alshabandar *et al.* [2018] used the F1 measure evaluation metric in the study of student success.

2.6 Conclusion

This chapter highlighted the researches done in the domain of student success/performance predictions concerning: impact of student engagement in performance, engagement detection, data source, feature selection, and predictive model. The study investigated the essential metrics for the prediction of student success based on engagement in blended-learning also.

From the literature, authors proposed further research into student success using engagement in a blended-learning environment as it has not been extensively investigated Van Goidsenhoven *et al.* [2020]; Park [2014]. In this study, we propose to examine students' success through student engagement in relation to the students' different behavioural patterns with the LMS and the provision of effective balancing techniques for the data emanating from blended-learning settings as we identified these gaps in the literature.

In this research, Logistic Regression, Support Vector Machine, Naïve Bayes Classifier, Decision Tree, Random Forests, Gradient Boosting Tree, Multilayer Perceptron, and Linear Discriminant Analysis predictive models will be explored. These models' choice is due to their high efficiency and accuracy on student data from the LMS for the impact of student engagement on student performance in a blended-learning (Kalboard 360 LMS) platform.

Chapter 3

Research Methodology

The previous chapter presented related work by researchers in the domain of student success predictions based on engagement in blended-learning environments. This chapter addresses the research approach, theoretical framework and methods to be adopted in this research. Student success prediction flow in Figure 3.1 itemises the process flow and the key steps that will be involved in this study to accomplish the goal of this study.

This study adopted a quantitative research approach on data from undergraduate studies based on the related works. The exploration of the linear and non-linear machine learning models was utilised to forecast students' success and increase their throughput rates in a blended-learning course. This study also used the theoretical framework of Amrieh *et al.* [2016] that categorises log-file data of students from an LMS into Behavioural, Demographical, and Academical categories to analyse the importance of each feature to the prediction.

The sections are structured as follows: Section 3.1 outlines the application of the study; Section 3.2 presents the data acquisition and collection of the dataset; Section 3.3 highlights the pre-processing and feature selection used in this study; Section 3.4 gives the machine learning models explored in this study; Section 3.5 shows the performance metrics for testing the performances of the models and Section 3.6 is the summary.

3.1 Application of Study

This study helps identify the students who are at-risk of failing in a course. Following the identification of at-risk students, the instructor's attention will be triggered to track such students' engagement and provide a suitable timely intervention to eliminate or reduce the failure rates. This instructor's intervention will, on the broader level, also improve the reputation of the university as the instructors' intervention will go a long way to put the student on track.

This research provides automated predictive models to the LMS used by the institution to monitor at-risk students. The predictive models will also produce significant

insight for the educational instructors to improve their teaching material and student performances.

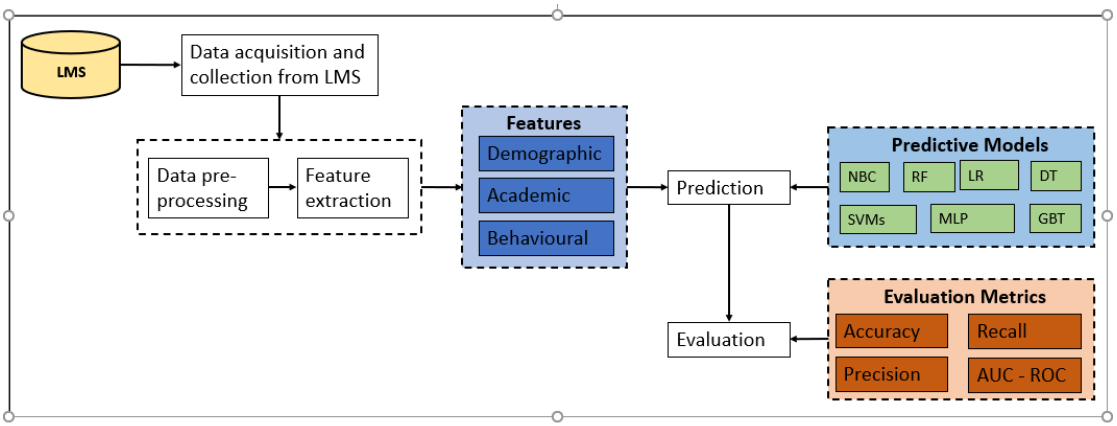

FIGURE 3.1: Student Success Prediction Flow

3.2 Data Acquisition and Collection

In this research, we used the student academic performance dataset that is free and accessible on Kaggle to analyse academic performance. The data was collected from the Kalboard 360 LMS using the xAPI educational activity tracker tool. The xAPI tracks knowledge and teaching tasks, such as login time or page read. The dataset is made up of 500 student records and 16 features as represented in Table 3.1. The attributes were divided into three: (1) Demographic variables, such as gender and nationality. (2) Academic variables, such as the educational stage, the degree level, and the section, and (3) Behavioral variables such as raising hands-on lessons, opening up resources, responding to parent surveys, and parent satisfaction with the school [Amrieh *et al.*, 2016].

3.3 Pre-processing and Features Selection

The performance of any data mining algorithm and the identified trend is essential for the dataset's effectiveness. The pre-processing facilitated the processing of unrefined datasets into a machine learning algorithm-comprehensible format. The emphasis was on identifying and discarding inaccurate or meaningless data such as missing data, anomalies, and inconsistent data. The process reduced the feature space's size, eliminated duplicate, insignificant, or meaningless data, and enhanced the result interpretation. We used information gain to assess the ranking features to determine which features are most necessary to build student success' prediction models. The top ten subsets of features were selected for dimensionality reduction and improved data quality for better model performances. The features were separated into demographic, academic background, and behavioural features for easy determination and evaluation of the vital engagement metrics in predicting student success. The target variable was split into three groups based on the student academic grades: Low, Medium, and High.

TABLE 3.1: The Students' Attributes Classification.

Attribute Classification	Attribute	Explanation
Demographic Attributes	Gender Nationality Place of Birth Relation	Statistical data such as age, gender
Academic Attributes	Stage ID Grade ID Section ID Topic Semester	Data related to student academic activities
Behavioral Attributes	Raise Hands Visited Resources Announcement Views Discussion Parent Answering Survey Parent School Satisfaction Student Absent Days	Student Engagement with LMS

The performance between 0 and 69 implies low; the performance between 70 and 89 means medium, and the performance between 90 and 100 is high.

$$Grade = \begin{cases} 0-69, & \text{Low} \\ 70-89, & \text{Medium} \\ 90-100, & \text{High} \end{cases}$$

3.3.1 Missing Values

There are 20 instances of students who have almost all their values missed,these instances were removed from data in order to clean the data and ensure completeness. The new counts of the dataset was brought to 480 instances with:

$$Class = \begin{cases} 127, & \text{Low} \\ 211, & \text{Medium} \\ 142, & \text{High} \end{cases}$$

3.3.2 Data Transformation

The dataset is the combination of both categorical features and continuous features; the categorical features were transformed for the machine learning models to understand it. Label encoding was used to transform the categorical features to machine intelligible format in this study, as shown in Figure 3.2.

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisITedResources	AnnouncementsView	Di
0	1	4	4	2	1	0	7	0	0	15	16	2	
1	1	4	4	2	1	0	7	0	0	20	20	3	
2	1	4	4	2	1	0	7	0	0	10	7	0	
3	1	4	4	2	1	0	7	0	0	30	25	5	
4	1	4	4	2	1	0	7	0	0	40	50	12	

FIGURE 3.2: Data Transformation using Label Encoding

3.3.3 Data Sampling

The data balancing strategies to obtain a more balanced data distribution will be applied to the training set to handle the class imbalance problem. A Synthetic Minority Oversampling Technique (SMOTE) will be employed on the dataset. SMOTE works by initially populating a minority class X_1 instance and determines its closest minority class neighbours. The synthetic sample is then generated by simply choosing one of the k nearest neighbours X_2 as well as linking X_1 and X_2 to create a line in the space of the feature. The synthetic samples will be created as a convex mix of the both X_1 and X_2 samples selected [Guo *et al.*, 2008; Longadge and Dongre, 2013].

The literature encourages SMOTE over Random Under-Sampling (RUS) and Random Over-Sampling(ROS) techniques due to possible loss of valuable data while removing majority class samples in RUS and the reuse of minority class samples in ROS. SMOTE solves ROS and RUS's likely problems by producing fresh synthetic data from minority samples [Longadge and Dongre, 2013]. A dataset is highly biased or skewed if one class sample is more significant than others [Longadge and Dongre, 2013]. This study will be testing RUS and ROS techniques to investigate the claim in [Longadge and Dongre, 2013].

3.3.4 Feature Selection

The information gain attribute evaluation determines the value of the attributes by measuring the entropy in relation to the rank. Information Gain indicates the importance of the attributes. The Information Gain filter will select the important features in building our models to obtain better accuracy and reduce feature dimension.

3.3.5 Libraries, Platforms, Frameworks, and Language

This study will utilise the scikit-learn framework (built on NumPy, SciPy, and matplotlib) to perform most experiments required for building the eight models in this study. The experiments include: sampling, 10-fold cross-validation on the training set, and the models will later be tested using the testing set. The programming language to be used is the python programming language via anaconda IDE application with Jupyter notebook.

3.4 Predictive Models

In this study, eight predictive models will be used in the prediction of student success. The models are Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Tree, Random Forests, Gradient Boosting Tree, Multilayer Perceptron, and Linear Discriminant Analysis Classifier. These models will be evaluated using k-fold cross-validation to test for their performances, where k is 10.

For the k-fold cross-validation, the initial sample was randomly split into k equal-sized subsets. A single subset of k subsets was maintained as validation data for testing, and the remaining k-1 subsets was used as training data. The cross-validation method was then used k times, for each of the k subsets used as validation data only once. The mean of results produced by the repeated k-fold sub-sampling was obtained to a single estimate.

This section is structured as follows: Section 3.4.1 discusses the Logistic Regression; Section 3.4.2 provides information on Support Vector Machines; Section 3.4.3 outlines details on Naïve Bayes Classifier; Section 3.4.4 explains the Decision Tree; Section 3.4.5 presents Random Forests; Section 3.4.6 discusses Gradient Boosting Tree; Section 3.4.7 highlights Multilayer Perceptron Model, and finally, Section 3.4.8 shows how Linear Discriminant Analysis model works.

3.4.1 Logistic Regression

Logistic regression is one of the most widely used in the analysis of EDM, due to its robust likelihood of performance by nature [Chen and Cui, 2020; Kang *et al.*, 2020; Van Goidsenhoven *et al.*, 2020]. It is used for model-dependent variables with the aid of independent variables. It is based on the estimation of the highest likelihood, and according to that likelihood, the data observed should be the most probable. Logistic regression performs better if there is no perfect correlation between the independent variables [Sperandei, 2014; Brownlee, 2016].

The logistic regression model is commonly used because of its flexibility and capacity to make evaluative assumptions concerning model terms [Kuhn *et al.*, 2013]. The application of logistic regression in this research is from [Brownlee, 2016; Kuhn *et al.*, 2013]. Equation 3.1 and Figure 3.3 are the mathematical and graphical representations of Logistic Regression respectively.

$$\Pr(\mathrm{x}) = \frac{1}{1 + \exp^{-(\beta_x)}} \tag{3.1}$$

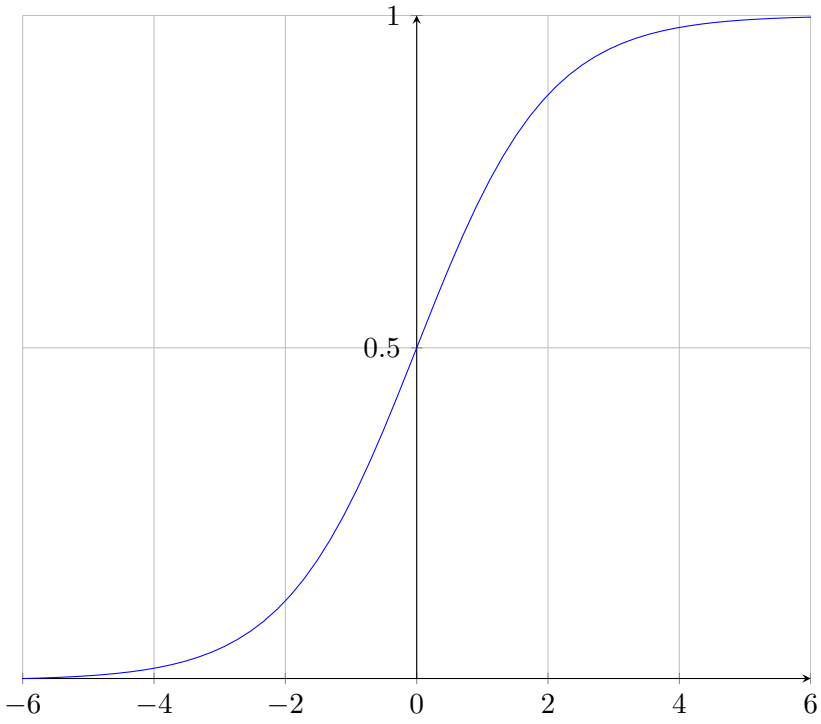

FIGURE 3.3: Logistic Regression Graphical Representation

3.4.2 Support Vector Machines

Support Vector Machines (SVMs) are supervised learning models with learning set of rules that analyse the data used to interpret categorization problems. SVMs are commonly used in EDM because they have high accuracy in prediction [Abed *et al.*, 2020]. SVMs can effectively perform a non-linear classification using the kernel trick, mapping their inputs into high-dimensional function spaces [Brownlee, 2016; Cristianini *et al.*, 2000]. The mathematical representation of Support Vector Machines' hyperplane is in equation 3.2 while Figure 3.4 is the graphical description of the hyperplane.

$$w^T x_z + b = 0, \tag{3.2}$$

such that:

$$w^T x_z + b \geq 1 \quad \text{for} \quad y_z = +1,$$
$$w^T x_z + b \leq -1 \quad \text{for} \quad y_z = -1.$$

where :
w = weight associated with the feature 1 to z,
b = set of points.

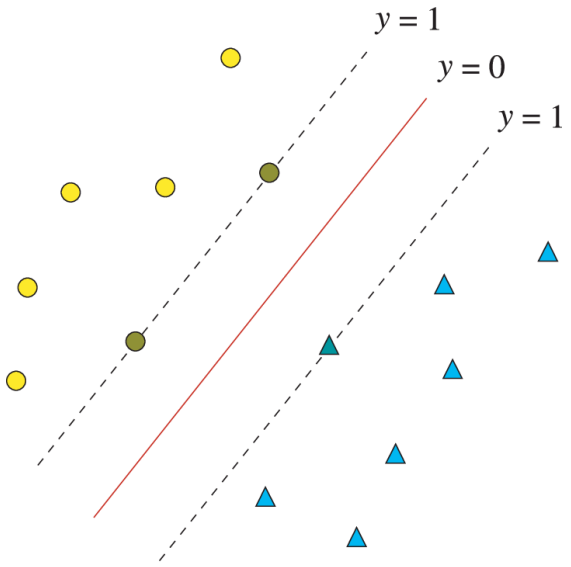

FIGURE 3.4: SVM Hyperplane

Seiffert *et al.* [2008] proposed in their study that SVMs can mitigate the skewed variable's issue without introducing noise. The SVM architecture used in this analysis is from [Park and Kim, 2020; Brownlee, 2016; Kuhn *et al.*, 2013].

3.4.3 Naïve Bayes Classifier

For most predictive problems, this algorithm is the most pragmatic and most straightforward learning approach. The Naïve Bayes Classifier (NBC) is based on Bayes' theorem with strong independence assumptions between the features using a probabilistic approach [Hussain *et al.*, 2018].

Naïve Bayes Classifier is efficient because: it takes less processing time, less training data, low variance, an explicit predictor of posterior likelihood, resilience to noise, and high capacity to handle missing values Webb *et al.* [2010] compared to most machine learning models.

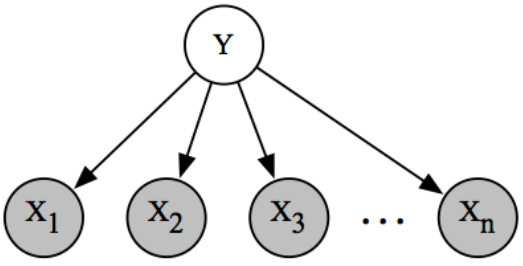

FIGURE 3.5: Naïve Bayes Graphical Representation

The classifier estimates the parameters of the probability distribution P(X/Y) on the training set of features X (16 features represented in table 3.1) given class Y (Low, Medium, or High), and then compute the posterior probability of the testing set. This

leads to classifying the testing set based on the largest computed probability. The probability can be represented mathematically as:

$$P(X/Y) = \frac{P(Y/X) * P(X)}{P(Y)} \tag{3.3}$$

Figure 3.5 is the graphical representation of a simple Naïve Bayes where: $\{X_1, X_2, X_3, \cdots X_n\} \in X$ are features and $\{Y\}$ is the target.

3.4.4 Decision Tree

A Decision Tree has a structural outline architecture graphically depicted in Figure 3.6 in which each internal node tests an attribute. Every division relates to the feature value, and every leaf node allocates a class (Low, High, or Medium). The tree is formed from the dataset by deciding which features at the child nodes better divide input features. The theory of information gain is used to split the nodes. If a node has minimal entropy, then it is used as a split node. "A decision tree fits when a study seeks to determine which features are important in a student prediction model" [Hussain *et al.*, 2018].

The decision tree utilises the training samples to identify the best points to divide the data to reduce the cost metric [Brownlee, 2016]. It is commonly used because of its high interpretability nature and requires lesser data pre-processing, unlike other machine learning algorithms [Brownlee, 2016]. The decision tree's implementation in this research is from [Kuhn *et al.*, 2013; Brownlee, 2016].

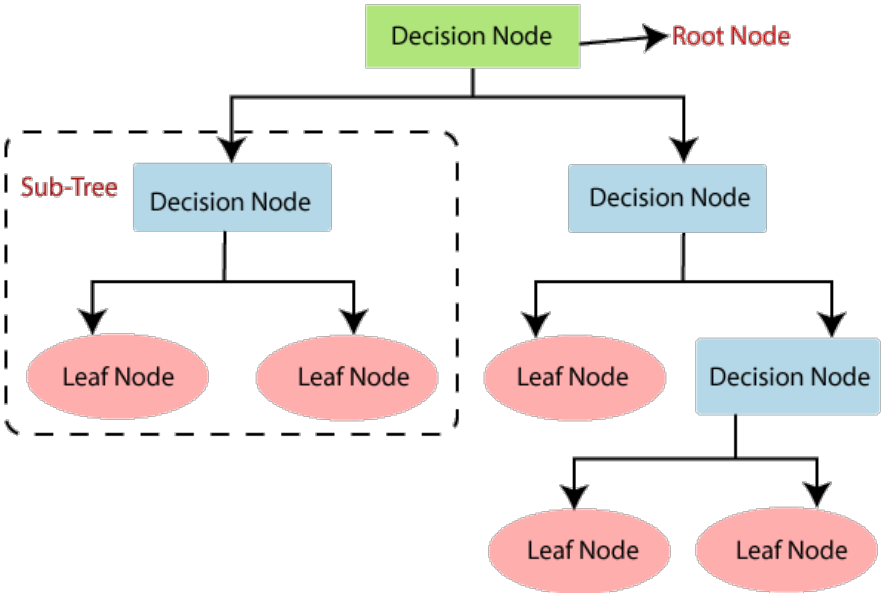

FIGURE 3.6: A Graphical Representation of Decision Tree

3.4.5 Random Forests

The Random Forests is an associative learning algorithm. It consists of several randomly generated decision trees. These randomly generated decision trees are merged to achieve higher accuracy and reliable predictions and inhibit over-fitting challenges likely to experience in a typical decision tree and [Ajoodha *et al.*, 2020]. The performance of random forests is usually better than decision trees Abed *et al.* [2020].

Random Forests is an evolution of the decision tree. Here, training dataset samples are taken with substitution, but the trees are built to eliminate classifiers' association. Principally, instead of rapaciously selecting the best breakup point in each tree's architecture, only an arbitrary set of features is selected for each division [Brownlee, 2016].

3.4.6 Gradient Boosting Tree

Gradient boosting is a machine learning technique that draws recognition to its speed and accuracy of prediction, particularly with massive and complicated data. It reduces the risk of over-fitting. It works by combining a learning algorithm to achieve an efficient learner from several weak learners that are concurrently related Hussain *et al.* [2018].

The gradient boosting concept is to integrate relatively weak predictive models to create a more potent predictive model. It is a very effective technique for creating predictive models. Gradient boosting is appropriate to several different risk functions and maximises these functions' accuracy, which is superior to predictive models [Hew *et al.*, 2020].

3.4.7 Multilayer Perceptron

Multilayer Perceptron (MLP) is an artificial neural network feed-forward class. It uses the supervised learning concept called back-propagation for training and includes a minimum of three processing layers: input, hidden, and output layers Yahaya *et al.* [2020].

The input layer obtains the input signal for analysis. The output layer does the tasks of prediction and classification, and the hidden layers positioned in between input and output layers serve as the computing driver[Raj and Evangeline, 2020]. In MLP, the data flow in the forward direction from the input to the output layer, and the neurons learn with the back-propagation learning algorithm[Raj and Evangeline, 2020]. The architecture of the multilayer perceptron used in this research is from [Raj and Evangeline, 2020; Izenman, 2013]

3.4.8 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a computational method for binary and multiclass classification. It assumes Gaussian distribution for input data for dimensionality reduction. It is easy, reliable statistically, and sometimes generates models whose performance is as excellent as methods that are more complex [Brownlee, 2016].

LDA follows the principle of finding a linear combination set of attributes that appropriately separates classes distinctly [Izenman, 2013]. The implementation of the Linear discriminant analysis classifier used in this research is from [Izenman, 2013].

3.5 Evaluation Metrics

The performance of the eight models used in this study will be measured using evaluation metrics: confusion matrix, accuracy, recall, precision, F1 score, and AUC - ROC Curve, which were mostly used in other related studies. The confusion matrix holds the value of the information produced by a predictive model about predicted and actual class labels. The data in the matrix is used to evaluate our models.

- Accuracy score is a very common evaluation metric for the classification models. It can be expressed as the number of precise prediction based on number of predictions as shown below.

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

- Recall is the number of correct positive predictions and the ratio of the total number of positives. This is also known by rate of true positive.

$$Recall = \frac{TP}{TP + FN}$$

- Precision is the number of correct positive predictions as a ratio of the total number of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

- The F1 score is the mean value for recall and precision. It offers an indicator of mistakenly graded results.

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision}$$

- The Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) will be used to measure the performances of the three classification models. AUC curve is calculated by this formula :

$$AUC = \frac{1}{2}(TPR + TNR)$$

The terms are used in AUC and ROC curve as follows.

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

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

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

3.6 Summary

In this chapter, the study's approach, theoretical framework and methodology were discussed step-by-step, splitting the data into 70 percent for training and 30 percent for testing. Thereafter, the model evaluation will be done using 10-fold cross-validation on the training set. Pre-processing of the data will also be carried out and classifying the data to the various classes available. The models will perform the predictions of students into high, low, and medium classes. The next chapter discusses the results of the models on the testing set.

Chapter 4

Results and Discussion

4.1 Introduction

In the previous chapter, the methodology utilised in the research was presented. This section offers and discusses the results obtained from running the Linear Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Tree, Random Forests, Gradient Boosting Tree, Multilayer Perceptron, and Linear Discriminant Analysis Classifier. Three major sampling experiments: Random Under-Sampling, Random Over-Sampling, and Synthetic Minority Oversampling Technique (SMOTE), were conducted as in this study.

The first section presents the description of the dataset. The second section addresses important information gain and feature importance results. The third section highlights three sampling methods and the models trained with the imbalanced class dataset. The last section is the summary of the chapter.

4.2 Dataset Description

The data acquisition was made using the xAPI learner activity tracker tool from the Kalboard 360 LMS. This form of teaching on the Kalboard 360 LMS is blended-learning. The total dataset contains 500 instances of student records and 16 feature [Amrieh *et al.*, 2016]. There were 20 instances of missing values in the dataset [Amrieh *et al.*, 2015, 2016]. If no information is entered for an attribute, missing values exist [Romero *et al.*, 2014]. In this research, students who have all or nearly all their values omitted were deleted/dropped from the data to clean up and guarantee data correctness. Figure 4.1 represents the dataset after dropping the missing value.

S/No	Features	Interpretations
1.	Gender	Gender of the student, either Female or Male
2.	Nationality	Citizenship of the studebt
3.	Place of birth	Country of birth
4.	Educational Stages	Student educational level
5.	Grade Levels	Student grade levels
6.	Section ID	Student's classroom
7.	Topic	Courses
8.	Semester	School semester
9.	Parent	Student relation
10.	Raised Hand	Number of times student raised a hand in class
11.	Visited Resources	Number of times student visited class content
12.	Viewing Announcements	Number of times student viewed announcement
13.	Discussion groups	Number of times student join the discussion
14.	Parent Answering Survey	Parent's response to the school survey
15.	Parent school Satisfaction	Parents' level of satisfaction with the school
16.	Student Absence Days	Number of times student is absent from class

TABLE 4.1: Features in the Dataset

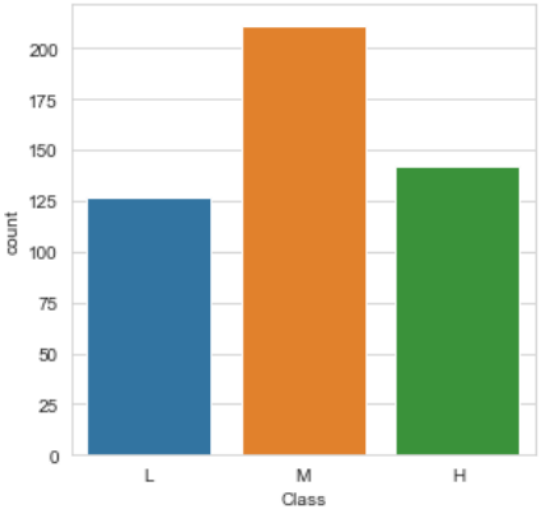

FIGURE 4.1: Data Distribution After Dropping the Missing Values

4.3 Information Gain and Feature Importance

4.3.1 Information Gain

In this phase, we used the information gain filter to assess the features' value by calculating the information gain (entropy) on a class basis. The features that contribute more

Rank	Entropy	Attribute Name
1	0.45801	Visited Resources
2	0.39745	Student Absence Days
3	0.37337	Raised Hands
4	0.2578	Announcements View
5	0.1504	Parent Answering Survey
6	0.12773	Nationality
7	0.1261	Relation
8	0.12292	Place of Birth
9	0.11393	Discussion
10	0.10676	Parent School Satisfaction
11	0.07611	Topic
12	0.05178	Gender
13	0.04748	Grade ID
14	0.01182	Semester
15	0.01058	Stage ID
16	0.00703	Section ID

TABLE 4.2: The Information Gain Ranking for a Set of Features to Predict the Student's Success

information have higher entropy, while those that do not add any information have a lower entropy. Table 4.2 shows the ranking of the important features. The information gain (entropy) value is $0 \leq e \leq 1$. That is, the entropy of features range from 0 to 1.

The features coloured blue are behavioural, the red-coloured features are demographic, and the yellow-coloured features are academic. The top five features belong to the student behavioural category, suggesting that behavioural features are significant in predicting student success in a blended-learning environment. The other feature categories in order of importance are: demographic and academic.

The outcome of this experiment shows the importance of student engagement in academic performance, especially as far as blended or any other online learning is concerned.

4.3.2 Feature Importance

To further investigate the vital information in predicting the student's success, Random Forests feature importance was also considered. Random Forests is utilised because of the tree-like techniques in Random Forests, which automatically calculates and gives a higher rating to the importance of the features [Brownlee, 2016]. Feature importance can provide a clear understanding of the dataset, models, and features to boost the predictive models [Brownlee, 2016].

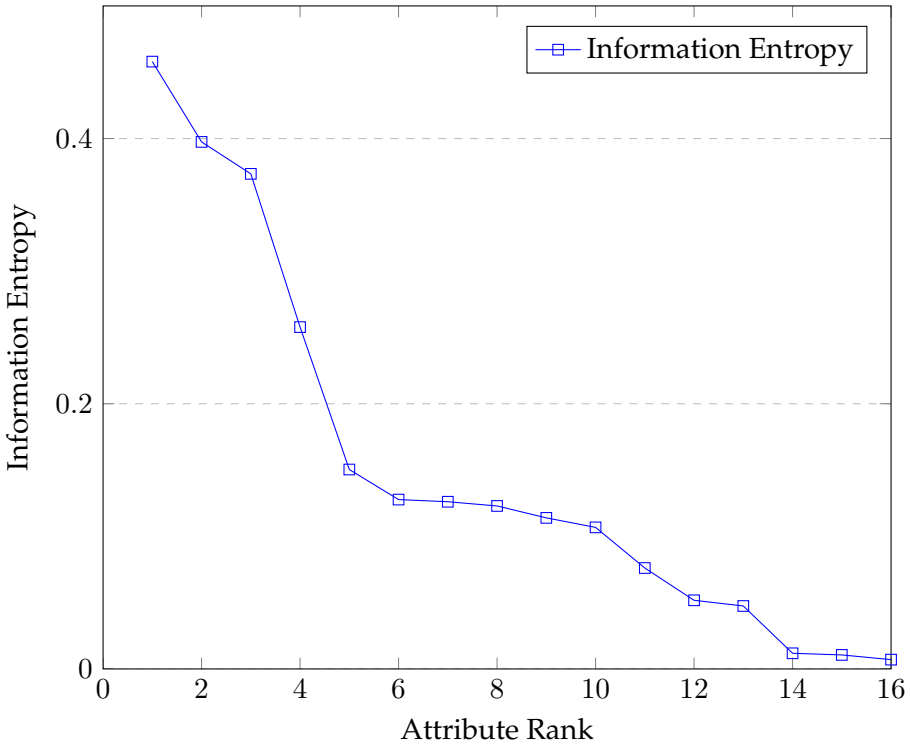

FIGURE 4.2: A Graphical representation of the Information Gain for Set of Attributes to Predict Student Success

Feature importance from Random Forests is represented in Figure 4.3 which corroborates the result in Table 4.2. The top five features in the ranking are: Student absence days, raised hands, visited resources, announcement view, and parent answering a survey. The topmost features in feature importance also belong to the behavioural category, and this gives insight into the features that are important in predicting student success.

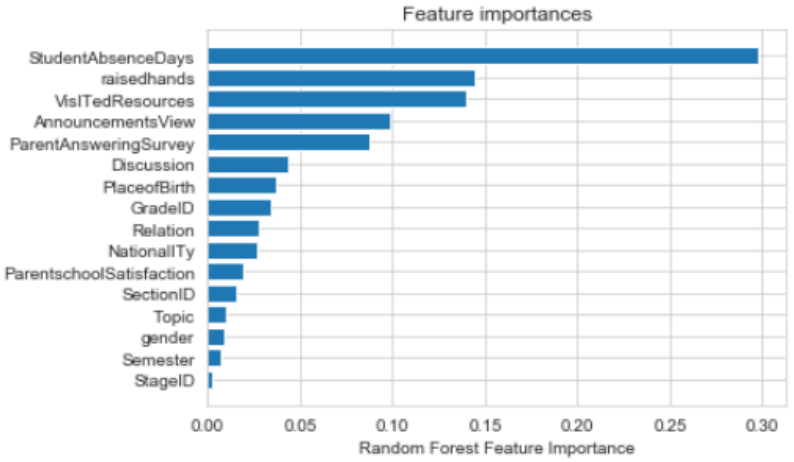

FIGURE 4.3: Random Forests Feature Importance

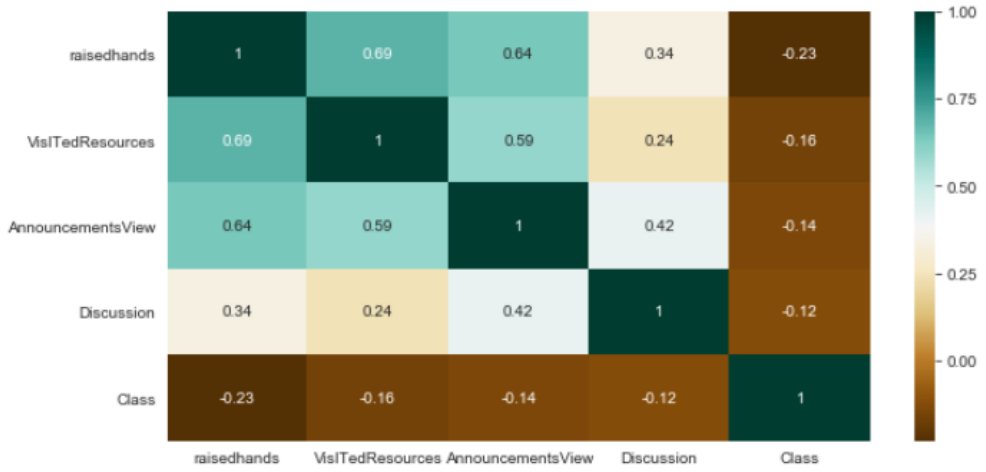

FIGURE 4.4: Correlation Matrix for Numerical Features

4.4 Classification Results with three Sampling Techniques and Imbalanced Dataset

The data was split into training and testing sets, with 70 percent of the data used for 10-fold cross-validation and 30 percent for testing. In order to resolve the problem of imbalanced class in the dataset, three sampling techniques were experimented with to ascertain the predictive model with optimal performance. Feature Selection (FS) was also employed to see the dimensionality reduction effect on the models used in this study.

4.4.1 Random Under-Sampling Technique with and without Feature Selection

Random Under-Sampling (RUS) technique of the data to equal counts was employed to avoid bias and the predictive models' poor performance. 10-fold cross-validation was performed on the training set. The first experiment uses the data's exclusive features, while the second experiment used Feature Selection (FS) to reduce feature dimensions. Table 4.3 shows the accuracy results of the two experiments conducted under the RUS.

The accuracies colour-coded green in Table 4.3 shows the model accuracies with higher performance in sampled data with the RUS and RUS plus FS. It shows that

five models (Decision Tree, Logistic Regression, Gradient Boosting Trees, Linear Discriminant Analysis, and Support Vector Machines) achieved better performances without feature selection. On the other hand, three models (Multilayer Perceptron, Random Forests, and Naive Bayes) improved their performances with the feature selection. From this RUS experiment, the Decision Tree model is the highest performing model with an accuracy value of 0.75.

The drawback of using the random under-sampling technique on the majority class to solve the imbalanced class problem in data mining is the loss of insightful facts from the dataset that can help in model prediction [Longadge and Dongre, 2013].

TABLE 4.3: Predictive Model Accuracies after 10-fold Cross-Validation for Random Under-Sampling Method

Predictive Model	RUS Accuracy	RUS + FS Accuracy
Logistic Regression	0.74	0.72
Support Vector Machines	0.58	0.55
Naïve Bayes	0.65	0.67
Decision Tree	0.75	0.67
Random Forests	0.72	0.74
Gradient Boosting Trees	0.74	0.72
Multilayer Perceptron	0.73	0.74
Linear Discriminant Analysis	0.74	0.69

4.4.2 Random Over-Sampling Technique with and without Feature Selection

Random Over-Sampling (ROS) technique of the data to equal counts was employed to avoid majority class bias and prevent predictive models' poor performance. The two experiments carried out here were on the full features and Feature Selection (FS). Table 4.4 shows the two experiments' accuracy results on the ROS.

Under ROS, the accuracies colour-coded green in Table 4.4 were better in performance with or without feature selection. The result indicates that five models (Gradient Boosting Trees, Logistic regression, Linear Discriminant Analysis, Naïve Bayes, and Multilayer Perceptron) performance were better without feature selection. The orange-coded accuracy shows no change in the decision tree's accuracy with the application of feature selection. Simultaneously, two models (Support Vector Machines and Random Forests) had a boost in performance with the feature selection.

The best performing model is Gradient Boosting Trees with an accuracy value of 0.75 without feature selection. The disadvantage of using ROS to replicate the minority

class in solving the class imbalance problem is that certain models overfit with the ROS application [Longadge and Dongre, 2013].

TABLE 4.4: Predictive Model Accuracies after 10-fold Cross-Validation for Random Over-Sampling Method

Predictive Model	ROS Accuracy	ROS + FS Accuracy
Logistic Regression	0.73	0.70
Support Vector Machines	0.56	0.58
Naïve Bayes	0.70	0.67
Decision Tree	0.68	0.68
Random Forests	0.70	0.72
Gradient Boosting Trees	0.75	0.74
Multilayer Perceptron	0.66	0.43
Linear Discriminant Analysis	0.72	0.71

4.4.3 SMOTE Technique with and without Feature Selection

In Synthetic Minority Over-Sampling Technique (SMOTE), new experimental samples are created to over-sample the disadvantaged minority group using the neighbourhood of every instance of the minority group [Longadge and Dongre, 2013]. SMOTE was applied to the data to balance the imbalanced class problem. Table 4.5 shows the two experiments' accuracy results on the SMOTE.

The SMOTE technique experiment demonstrates that five models (Decision Tree, Multilayer Perceptron, Gradient Boosting Trees, Logistic Regression, and Naïve Bayes) in green colour had higher accuracies without the implementation of feature selection. The orange-coloured accuracies are indifferent with or without feature selection, and two models (Random Forests and Support Vector Machines) performances increased with the application of feature selection.

The Random Forests model obtained the optimal performance with an accuracy value of 0.91 with feature selection. The challenge of using SMOTE for data pre-processing is that the technique takes a long time in learning processes [Longadge and Dongre, 2013].

TABLE 4.5: Predictive Model Accuracies after 10-fold Cross-Validation for SMOTE Method

Predictive Model	SMOTE Accuracy	SMOTE + FS Accuracy
Logistic Regression	0.84	0.83
Support Vector Machines	0.70	0.72
Naïve Bayes	0.83	0.80
Decision Tree	0.89	0.87
Random Forests	0.90	0.91
Gradient Boosting Trees	0.84	0.83
Multilayer Perceptron	0.89	0.84
Linear Discriminant Analysis	0.81	0.81

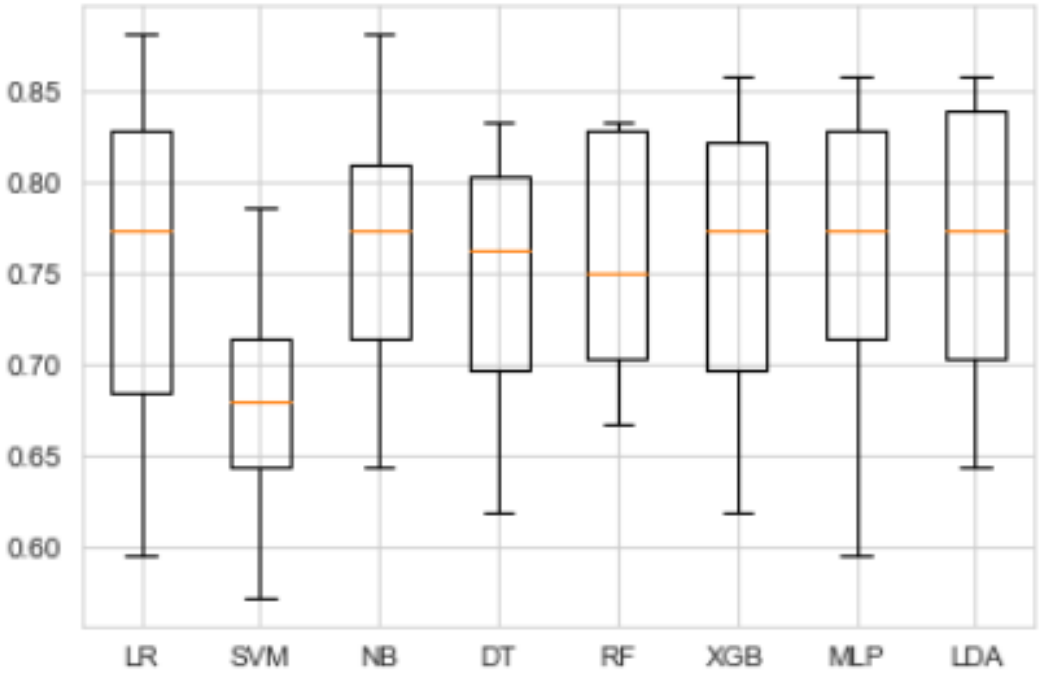

FIGURE 4.5: Algorithm Comparison by Training Set

4.4.4 Classification on Imbalanced Data with Feature Selection and without Feature Selection

In this phase, the dataset with an imbalanced class was used in training the models using 10-fold cross-validation like the other previous experiments. This experiment seeks to investigate the performances of the model when the sampling method is not in use. Two experiments were conducted in this phase that also includes the use of feature selection and without feature selection. Table 4.6 represents the implementation of 10-fold cross-validation on the imbalance class dataset with sampling technique.

The result in Table 4.6 the green-coloured accuracies shows that four models (Logistic Regression, Linear Discriminant Analysis, Naïve Bayes and Support Vector Machines performed better without the feature selection. The orange-coloured accuracy of Random Forests indicates that feature selection does not affect the performance of the model, and three models' (Gradient Boosting Trees, Decision Tree, and Multilayer Perceptron) performances improved with the feature selection.

TABLE 4.6: Predictive Model Accuracies after 10-fold Cross-Validation on the Imbalanced Class

Predictive Model	Imbalanced Class Accuracy	Imbalanced Class + FS Accuracy
Logistic Regression	0.75	0.74
Support Vector Machines	0.63	0.60
Naïve Bayes	0.69	0.67
Decision Tree	0.70	0.73
Random Forests	0.72	0.72
Gradient Boosting Trees	0.72	0.74
Multilayer Perceptron	0.59	0.72
Linear Discriminant Analysis	0.75	0.74

4.4.5 Feature Selection using the top 5 and 10 Feature Selection with the Resampling Techniques

In this phase of the experiment, the top 5 features in table 4.2 and figure 4.3 were selected from the re-sampled dataset with the SMOTE, RUS, and ROS. Table 4.7 shows that the top five features improved the performances of Logistic Regression, Gradient Boosting Trees, Linear Discriminant Analysis, and Naïve Bayes models slightly under SMOTE re-sampling technique.

The Random Forests, Logistic Regression, and Linear Discriminant Analysis models' performances boosted with the top five features' selection under the RUS re-sampling method. The ROS re-sampling technique and the top 5 features increased the Random Forests, Logistic Regression, Linear Discriminant Analysis, Naïve Bayes, and the Multilayer Perceptron models' performances. We can conclude that a reduction in feature dimensionality improved the model performances slightly.

4.5 Model Evaluation with Confusion Matrix,Accuracy Precision, Recall, F1-Score,AUC and ROC curve

On average, the SMOTE technique without feature selection presents the eight models used in this research with the higher performances across the board above the SMOTE

TABLE 4.7: Comparison of the top 5 and 10 Feature Selection with the Re-sampling Techniques

Predictive Model	**SMOTE + FS**		**RUS +FS**		**ROS + FS**	
	5 FS	**10 FS**	**5 FS**	**10 FS**	**5 FS**	**10 FS**
Decision Tree	0.86	**0.87**	0.67	0.67	0.63	**0.68**
Random Forests	0.85	**0.91**	**0.76**	0.74	**0.74**	0.72
Logistic Regression	**0.84**	0.83	**0.74**	0.72	**0.73**	0.7
Gradient Boosting	**0.84**	0.83	0.69	**0.72**	0.72	**0.74**
Linear Discriminant Analysis	**0.83**	0.81	**0.72**	0.69	**0.72**	0.71
Naïve Bayes	**0.83**	0.8	0.67	0.67	**0.68**	0.67
Multilayer Perceptron	0.79	**0.84**	0.74	0.74	**0.76**	0.43
Support Vector Machines	0.7	**0.72**	0.55	0.55	0.55	**0.58**

and feature selection implementation. Therefore, the performance evaluation metrics were implemented on the SMOTE technique without feature selection. In addition to the use of 10-fold cross-validation and testing accuracy, other set of measurement metrics were implemented to also assess the performance of the model in subsections 4.5.1, 4.5.2, 4.5.3, 4.5.4, 4.5.5, 4.5.6 and 4.5.7.

4.5.1 Confusion Matrix

The confusion matrix holds the value of the information produced by a predictive model about predicted and actual class labels. It is a table structure that enables the visualisation of the algorithm's outcome.

The confusion matrices in Figures 4.8,4.9, 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15 were computed from the models trained with SMOTE technique without feature selection due to high accuracy obtained across the eight predictive models in this study. The prediction of low-class was predicted better than medium and high classes.

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**95%**	**5%**	**0%**
	Medium	**11%**	**72%**	**17%**
	High	**0%**	**16%**	**84%**

TABLE 4.8: **Logistic Regression** Predictive Model's Confusion Matrix

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**93%**	**2%**	**5%**
	Medium	**36%**	**36%**	**28%**
	High	**6%**	**18%**	**76%**

TABLE 4.9: **Support Vector Machines** Predictive Model's Confusion Matrix

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**93%**	**7%**	**0%**
	Medium	**8%**	**69%**	**23%**
	High	**0%**	**14%**	**86%**

TABLE 4.10: **Naïve Bayes** Predictive Model's Confusion Matrix

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**100%**	**0%**	**0%**
	Medium	**6%**	**69%**	**25%**
	High	**0%**	**6%**	**94%**

TABLE 4.11: **Decision Tree** Predictive Model's Confusion Matrix

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**95%**	**5%**	**0%**
	Medium	**5%**	**83%**	**12%**
	High	**0%**	**10%**	**90%**

TABLE 4.12: **Random Forests** Predictive Model's Confusion Matrix

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**95%**	**5%**	**0%**
	Medium	**11%**	**69%**	**20%**
	High	**0%**	**14%**	**86%**

TABLE 4.13: **Gradient Boosting** Predictive Model's Confusion Matrix

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**95%**	**5%**	**0%**
	Medium	**6%**	**88%**	**6%**
	High	**0%**	**16%**	**84%**

TABLE 4.14: **Multilayer Perceptron** Predictive Model's Confusion Matrix

4.5.2 Accuracy

Accuracy value is a widespread evaluation metric for classification models. It can be calculated as the number of correct predictions divided by the total number of predictions. For a balanced dataset, accuracy is a fair indicator of the efficiency of the model. An imbalanced dataset requires other evaluation metrics for the test of how well the model performs.

In this study, the predictive models' best accuracy values were achieved with an

		Predicted		
		Low	**Medium**	**High**
Actual	**Low**	**95%**	**5%**	**0%**
	Medium	**17%**	**61%**	**22%**
	High	**0%**	**16%**	**84%**

TABLE 4.15: **Linear Discriminant Analysis** Predictive Model's Confusion Matrix

experiment using SMOTE sampling technique as represented in Table 4.16 in descending order after 10-fold cross-validation. Random Forests reached the highest accuracy value of 90%. Decision Tree, Multilayer Perceptron, Gradient Boosting Trees, Logistic Regression, Naïve Bayes, Linear Discriminant Analysis, and Support Vector Machines obtained accuracy values 89%,89%,84%,84%,83%,81%, and 70% respectively.

Model	Accuracy (%)
Random Forests	**90**
Decision Tree	89
Multilayer Perceptron	89
Gradient Boosting Trees	84
Logistic Regression	84
Naïve Bayes	83
Linear Discriminant Analysis	81
Support Vector Machines	70

TABLE 4.16: Accuracy Values for the Predictive Models

4.5.3 Precision

Precision, which is Positive Predictive Value (PPV), tests the models for the number of the expected positive results that are positive from all positive results. Table 4.17 demonstrates the precision evaluation metric results for this research's predictive models. Decision Tree, Random Forests, and Multilayer Perceptron have higher precision scores with precision values of 0.89 each. Other models with reasonably good precision values are Logistic Regression, Naïve Bayes, and Gradient Boosting Trees with 0.83 precision value each. Linear Discriminant Analysis follows the order with 0.80. The least precision value is the Support Vector Machines with 0.67.

The greater the precision score, the stronger the model at predicting the classes. The low-class has the highest precision score across the board, followed by the high-class, and the medium-class is the least. The implication of this lower precision score for the medium-class is that the models are not predicting the medium class appropriately.

Models	Low	Medium	High	Average
Logistic Regression	**0.90**	0.72	0.88	0.83
Support Vector Machines	0.70	0.57	**0.76**	0.67
Naïve Bayes	**0.93**	0.71	0.84	0.83
Decision Tree	**0.95**	0.89	0.84	0.89
Random Forests	**0.95**	0.81	0.92	0.89
Gradient Boosting Trees	**0.90**	0.74	0.86	0.83
Multilayer Perceptron	**0.95**	0.76	**0.95**	0.89
Linear Discriminant Analysis	**0.86**	0.69	0.84	0.80

TABLE 4.17: Precision Table for the Predictive Models

4.5.4 Recall

The recall calculates the correctly estimated positive results from all the observed positive results. The order of correctly predicted classes as depicted in Table 4.18 are: Random Forests (0.89), Multilayer Perceptron (0.89), Decision Tree (0.88), Logistic Regression (0.84), Naïve Bayes(0.83), Gradient Boosting(0.83), Linear Discriminant Analysis (0.80) and Support Vector Machines which is the least with recall value 0.68.

Models	Low	Medium	High	Average
Logistic Regression	**0.95**	0.72	0.84	0.84
Support Vector Machines	**0.93**	0.36	0.76	0.68
Naïve Bayes	**0.93**	0.69	0.86	0.83
Decision Tree	**1.00**	0.69	0.94	0.88
Random Forests	**0.95**	0.83	0.90	0.89
Gradient Boosting Tress	**0.95**	0.69	0.86	0.83
Multilayer Perceptron	**0.95**	0.89	0.84	0.89
Linear Discriminant Analysis	**0.95**	0.61	0.84	0.80

TABLE 4.18: Recall Table for the Predictive Models

4.5.5 F1 Score

The F1 score is vital as a bridge for recall and precision. It offers a measure of the findings that are incorrectly graded. F1 score is the best metric for measuring the performance of models on an imbalanced dataset.

The higher values of the F1 score specifies the excellent performance of the model. Generally, the range of the F1 score is from 0 to 1. The models with the best F1 scores in our study are Random Forests and Multilayer Perceptron with 0.89 F1 scores each.

Models	Low	Medium	High	Average
Logistic Regression	**0.93**	0.72	0.86	0.84
Support Vector Machines	**0.80**	0.44	0.76	0.67
Naïve Bayes	**0.93**	0.70	0.85	0.83
Decision Tree	**0.98**	0.78	0.89	0.88
Random forestss	**0.95**	0.82	0.91	0.89
Gradient Boosting Tress	**0.93**	0.71	0.86	0.83
Multilayer Perceptron	**0.95**	0.82	0.89	0.89
Linear Discriminant Analysis	**0.90**	0.65	0.84	0.80

TABLE 4.19: F1-Score Table for the Predictive Models

4.5.6 Area Under Curve (AUC)

The AUC value returns a holistic rating of a predictive model's performance. AUC shows how well a model can separate the classes from each other. The AUC value of 1 signifies a clear division of classes, and the mid-way (0.5) value indicates that the model randomises the classes in prediction.

In Table 4.20, virtually all the models in this study have AUC values close to 1 except Support Vector machines with an AUC value of 0.71. The results in Table 4.20 indicates that the other seven models, that is, Logistic Regression, Naïve Bayes, Decision Tree, Random Forests, Gradient Boosting Trees, Multilayer Perceptron, and Linear Discriminant Analysis, were able to separate the classification/prediction of classes low, medium and high distinctively.

Model	AUC
Logistic Regression	0.94
Support Vector Machines	0.71
Naïve Bayes	0.92
Decision Tree	**0.98**
Random Forests	**0.98**
Gradient Boosting Trees	0.95
Multilayer Perceptron	0.94
Linear Discriminant Analysis	0.94

TABLE 4.20: AUC Score for the Predictive Models

4.5.7 Receiver Operating Characteristic (ROC)

Receiver Operating Characteristic (ROC) was used to measure the performances of the eight classification models. ROC is a visual and valuable method for measuring the efficiency of classifiers. ROC effectively describes the exchange between the TP and FP rates as a graph. ROC can reveal potential and consider how the model can correctly

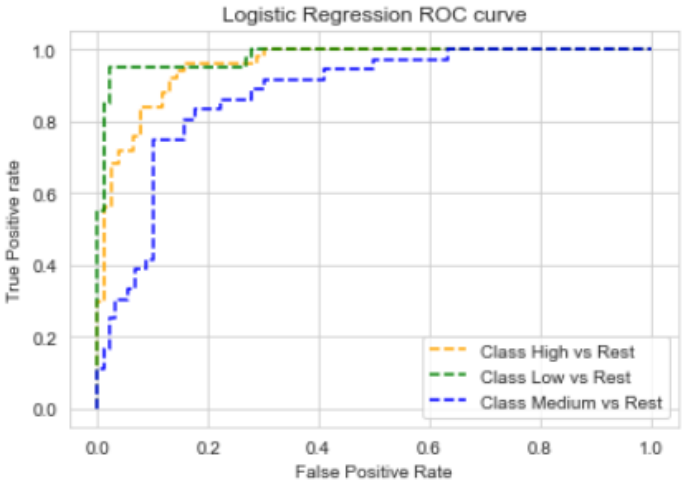

FIGURE 4.6: Logistic Regression ROC

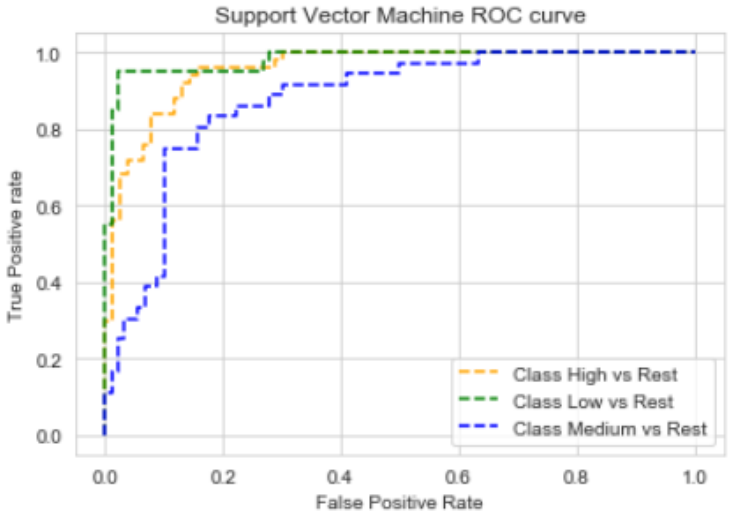

FIGURE 4.7: Support Vector Machines ROC

perform or misidentify negative classes as positive. If the ROC curve is equal to 1, then model is effective in prediction [Hashim *et al.*, 2020; Bowers and Zhou, 2019].

More often than not, it is ideal to have a ROC graph with a significant amount of distance below the curve. ROC closer to 1 is preferable and ROC of 0.5 is really poor model [Bowers and Zhou, 2019]. In this study, the ROCs for all the predictive models are represented in Figures 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12 and 4.13 with good space below the curve and closer to 1.

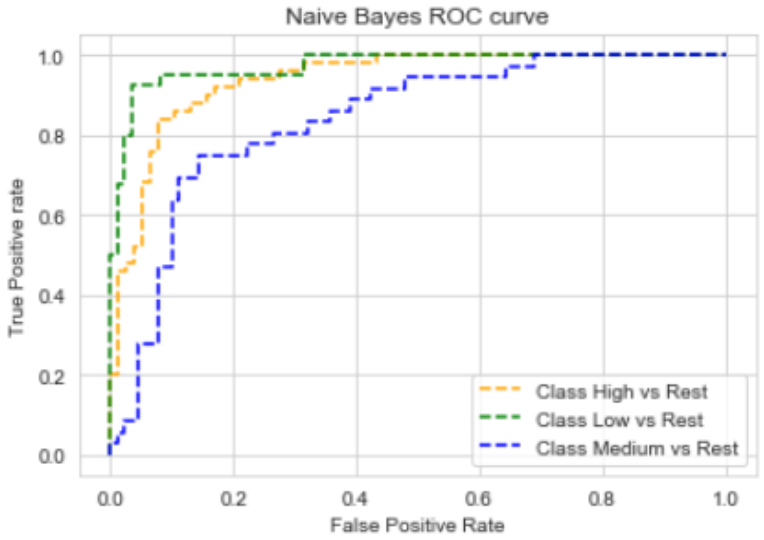

FIGURE 4.8: Naïve Bayes ROC

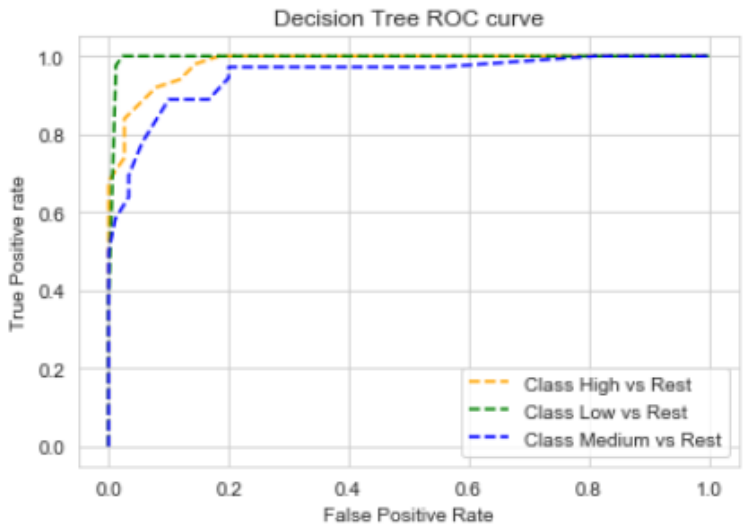

FIGURE 4.9: Decision Tree ROC

4.6 Discussion

This study predicts student success in the right class using the LMS's student engagement.

The class imbalance problem in the data was solved using SMOTE technique after experiments using RUS and ROS techniques. The feature selection effect was also experimented on the data to achieve the best accuracy across all the models used in this research. The accuracy, precision, recall, F1 score, AUC, and ROC obtained from the models proved that the SMOTE technique was better than RUS and ROS methods. The

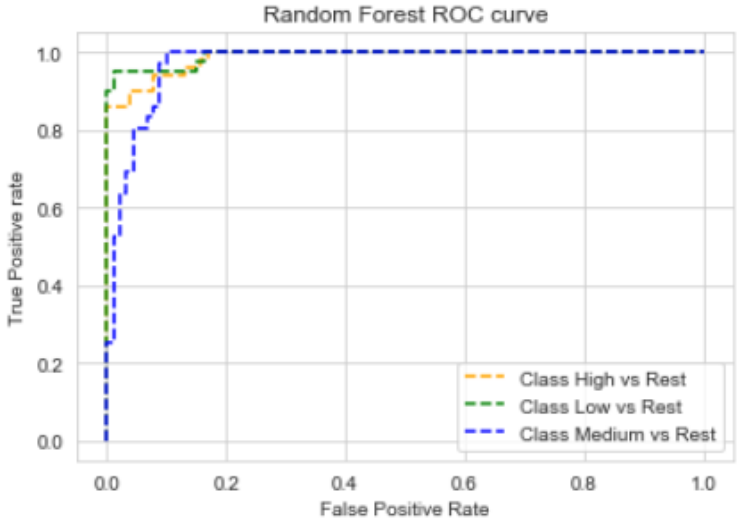

FIGURE 4.10: Random Forests ROC

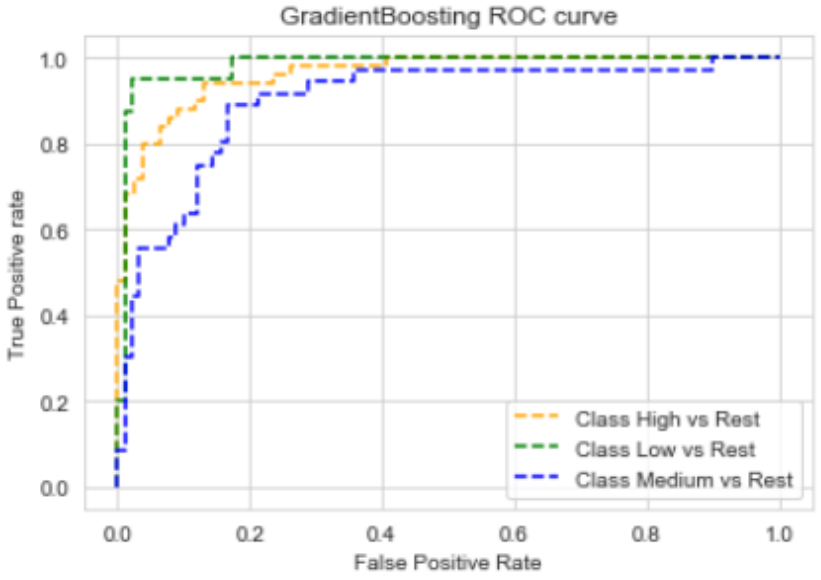

FIGURE 4.11: Gradient Boosting Trees ROC

SMOTE balancing result affirms the claim in Longadge and Dongre [2013] regarding the choice of pre-processing sampling method for handling the class imbalance problem.

The Information Gain filter was used to determine the contributing importance of the features in predicting student success classes. The results in Table 4.2 and Figure 4.3 highlight each feature's contribution to the student classes classification. The feature selection experiment phase demonstrates little or no contribution from the feature selection in the overall models' performances, contrary to the view of [Deepika and

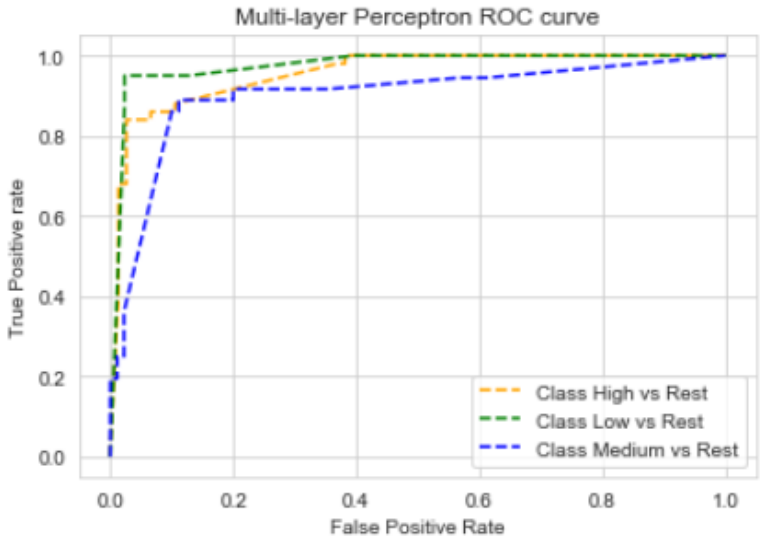

FIGURE 4.12: Multilayer Perceptron ROC

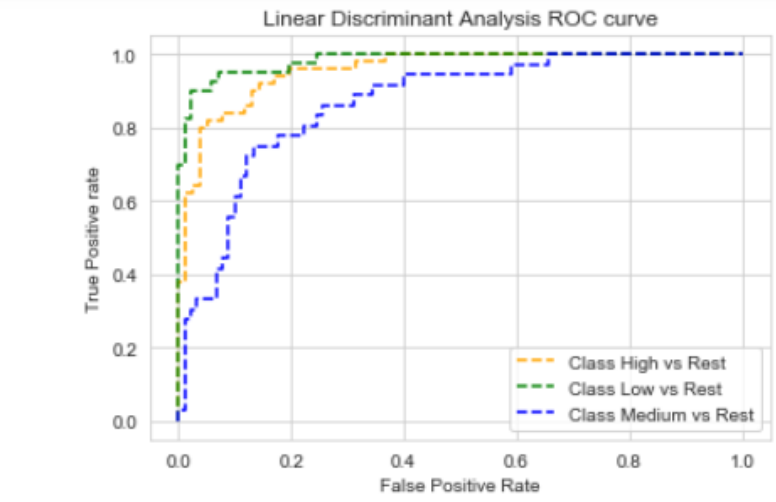

FIGURE 4.13: Linear Discriminant Analysis ROC

Sathyanarayana, 2019]. The contradiction in the literature and the results obtained using feature selection is probably due to the nature and the size of the data used in this study.

The top five features are the most contributing features to the model predictions from the Information Gain ranking. These top five features are under the behavioural attributes of students on LMS, which agrees with the studies by Amrieh *et al.* [2016]; Dutt and Ismail [2019], while the other features from six to sixteen are categorised as demographic and academic, respectively. The result from the Information Gain indicates that the students' engagements on LMS directly correlate with the students' academic performances.

All the machine learning models used in this study predicted students' classes reasonably well, with low-class being the major predicted class and followed by high-class. The least predicted class is the medium class. Table 4.21 is the comparison of the evaluation metrics and the eight predictive models in this study.

Random Forests performed better just like the results from the studies of Ajoodha *et al.* [2020]; Dutt and Ismail [2019] than all the other models with an accuracy value of 90%, AUC of 0.98, precision of 0.89, recall of 0.89, and F1 score of 0.89. The outstanding performance of the Random Forests is due to the randomly generated decision trees that are merged to achieve higher accuracy and reliable predictions. The hyperparameter tuning used for the best accuracy in Random Forests is ten tree numbers in the forest (n_estimator), the function for measurement of entropy (information gain), and zero randomness of the bootstrapping. The other models' accuracy values are 89%, 89%, 84%, 84%, 83%, 81%, and 70% for Decision Tree, Multilayer Perceptron, Gradient Boosting Trees, Logistic Regression, Naive Bayes, Linear Discriminant Analysis and Support Vector Machines respectively. The Least performing model is SVM, with low values for the performance metrics across the models.

The results obtained in this chapter answered the research questions in Chapter 1 as follows:

- **What are the key engagement metrics for early identification of at-risk students for the instructors' hypothetical timely intervention in the blended-learning course?**
  Information gain and feature importance in Table 4.2 and Figure 4.3 show that the top engagement metrics are essential in the identification of at-risk students:
  (i) Visited Resources: the number of times students visited the resources,
  (ii) Student Absence Days: the number of absent days on the LMS,
  (iii) Raised Hands: the number of times the student raised hands to ask questions or make contributions on LMS,
  (iv) Announcement Views: the number of times students viewed the announcement placed on LMS and
  (v) Parent Answering Survey: The parent response to the surveys offered by the school.

  These top five engagement metrics are in the behavioural category of the feature classification.

- **Which machine learning models used in this study offer optimum performance in predicting student success in the blended-learning course?**
  The machine learning model with the optimal performances is Random Forests.

Random Forests predicted low class appropriately by 95%, medium class accurately by 83%, and high class rightly by 90%. Although Decision Tree, Gradient Boosting Tree, Multilayer Perceptron, and Logistic Regression models predicted some classes more accurately but not across the classes.

- **Which features are crucial to predicting students' performance in the blended-learning course?**
  The features required to predict student performance in the blended-learning course are the top five features in Table 4.2 and Figure 4.3. The experiment with the top five features shows a slight improvement in some models' performances as represented in Table 4.7.

- **Which of the data sampling techniques used in this study give the best model performance in the prediction of student success**?
  The results obtained from sampling techniques highlight the best model performance in the prediction of student success as shown in Tables 4.3, 4.4 and 4.5. The models with the SMOTE sampling techniques achieved higher performances compared to the results from ROS and RUS.

Model	**Accuracy** (%)	**Precision**	**Recall**	**F1 score**	**AUC**
Random Forests	90	0.89	0.89	0.89	0.98
Decision Tree	89	0.89	0.88	0.88	0.98
Multilayer Perceptron	89	0.89	0.89	0.89	0.94
Gradient Boosting Trees	84	0.83	0.83	0.83	0.95
Logistic Regression	84	0.83	0.84	0.84	0.94
Naive Bayes	83	0.83	0.83	0.83	0.92
Linear Discriminant Analysis	81	0.80	0.80	0.80	0.94
Support Vector Machines	70	0.67	0.68	0.67	0.71

TABLE 4.21: Summary of the Evaluation Metrics for Predictive Models

4.7 Summary

In this chapter, the results obtained from training the eight predictive models were presented and discussed. The Random Forests model's performance was the best in this study with an accuracy value of 90%, AUC of 0.98, the precision of 0.89, recall of 0.89, and F1 score of 0.89. The top ten features in Table 4.2 highlighted the significant features in predicting student success. The top ten features were also identified to belong to behavioural attributes, which indicates the essentials of the students' behavioural patterns explicitly to their performances.

We experimented with different re-sampling techniques, and the SMOTE technique improved the models' performances across the board. The RUS technique's effect on the models' performances appears better in accuracy than the ROS technique effect. Finally, this current study's comparative analysis of predictive models and re-sampling methods revealed that machine learning techniques have a prospective role in predicting student success and other educational outcomes.

Chapter 5

Conclusion, Further Study, Limitations, and Contribution

The results obtained by the trained models were explained in the previous chapter, together with other findings and future approaches that could enhance the overall models. This chapter presents the study's conclusion on predicting student success using student engagement in a blended-learning course's online component. Other discussions in this chapter are possible future work, the drawbacks of the study, and contributions.

5.1 Conclusion

This study aimed to predict student success using student engagement in the LMS in a blended-learning course. We showed that random forests' performance exceeded other classifiers in this study in terms of evaluation metrics. This study also revealed that feature selection does not significantly affect the performances of the models. However, random forest and support vector machines only had a little leap with feature selection, as shown in 4.5.

The SMOTE technique of data sampling gave promising results in the models ' performances with a major classification of low and high classes. The medium class was disadvantaged across the models with the use of SMOTE and the other two data sampling methods.

The information gain ranking highlighted the behavioural features as the highest determinants in predicting student success. To alleviate the risk of failing or dropout drastically in a blended-learning, irregularities in students' behavioural attributes are

pointers to instructors or institutions for timely intervention to prevent colossal damage in student education trajectory. Based on the predictive models used in this study and their evaluation metrics, we conclude that student engagement can play a vital role in predicting student success.

The predictive models explored in this study offer substantial insight into education for student performance improvement, increase in throughput rate, and the reduction in the dropout rate. This study implies that the instructor or institution's timely intervention will avert possible failure in the students' courses, increase throughput rate, and other academic outcomes.

5.2 Future Study Recommendation

An investigation into better and more accurate data sampling methods for balancing the class imbalance problem is recommended to dig deep into the medium class' least prediction anomaly experienced in this study.

Implementing the predictive models used in this study is also recommended on the University of Witwatersrand undergraduate students' data to ascertain the important features to detect students at risk of failing and dropping out of the program for possible instructor and institution intervention.

An extension of experiments in this study to other models like deep neural networks and other sophisticated machine learning models is also recommended for a holistic view of other useful factors in predicting student success.

5.3 Limitation

One major limitation of this research was that we could not test the predictive models on the university's undergraduate student data as proposed due to concerns surrounding the release of racial demographic information for research, as indicated by the university's ethics committee. This limited our research scope somewhat; however, in this case, we only opted for data from online repositories. The data from online repositories are not as rich as the data from the university's repository.

The size of the data used in this study is also a limitation. A larger data would have given a better insight into the other influencing factors on student success prediction. Another limitation is the non-availability of free and robust educational datasets online due to privacy-related issues in ethics.

5.4 Contributions

The contributions of this study are the provision of the vital engagement metrics on students performances, early detection technique for instructors or lecturer to identify a student at risk of failing during courses for possible intervention based on the LMS behavioural patterns, and the impacts of student engagement on the students' success or performance.

This study also provides optimum data re-sampling methods for the imbalanced class problem associated with data in educational settings. This research presents a foundation for future research into different re-sampling strategies combined with feature space reductions.

Bibliography

[Abed *et al.* 2020] Tasneem Abed, Ritesh Ajoodha, and Ashwini Jadhav. A prediction model to improve student placement at a south african higher education institution. In *2020 International SAUPEC/RobMech/PRASA Conference*, pages 1–6. IEEE, 2020.

[Ajoodha *et al.* 2020] Ritesh Ajoodha, Ashwini Jadhav, and Shalini Dukhan. Forecasting learner attrition for student success at a south african university. In *Conference of the South African Institute of Computer Scientists and Information Technologists 2020*, pages 19–28, 2020.

[Alshabandar *et al.* 2018] Raghad Alshabandar, Abir Hussain, Robert Keight, Andy Laws, and Thar Baker. The application of gaussian mixture models for the identification of at-risk learners in massive open online courses. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2018.

[Alsubhi *et al.* 2019] Mohammed Abdulaziz Alsubhi, Noraidah Sahari Ashaari, and Tengku Siti Meriam Tengku Wook. The challenge of increasing student engagement in e-learning platforms. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 266–271. IEEE, 2019.

[Amrieh *et al.* 2015] Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Preprocessing and analyzing educational data set using x-api for improving student's performance. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5. IEEE, 2015.

[Amrieh *et al.* 2016] Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8):119–136, 2016.

[Asiah *et al.* 2019] Mat Asiah, Khidzir Nik Zulkarnaen, Deris Safaai, Mat Yaacob Nik Nurul Hafzan, Mohamad Mohd Saberi, and Safaai Siti Syuhaida. A review on predictive modeling technique for student academic performance monitoring. In *MATEC Web of Conferences*, volume 255, page 03004. EDP Sciences, 2019.

[Bernard *et al.* 2014] Robert M Bernard, Eugene Borokhovski, Richard F Schmid, Rana M Tamim, and Philip C Abrami. A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education*, 26(1):87–122, 2014.

[Bosch 2016] Nigel Bosch. Detecting student engagement: human versus machine. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 317–320, 2016.

[Bote-Lorenzo and Gómez-Sánchez 2017] Miguel L Bote-Lorenzo and Eduardo Gómez-Sánchez. Predicting the decrease of engagement indicators in a mooc. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 143–147, 2017.

[Bowers and Zhou 2019] Alex J Bowers and Xiaoliang Zhou. Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, 24(1):20–46, 2019.

[Brownlee 2016] Jason Brownlee. Machine learning mastery with python. *Machine Learning Mastery Pty Ltd*, pages 100–120, 2016.

[Castro *et al.* 2007] Félix Castro, Alfredo Vellido, Angela Nebot, and Francisco Mugica. Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment*, pages 183–221. Springer, 2007.

[Cavus and Zabadi 2014] Nadire Cavus and Teyang Zabadi. A comparison of open source learning management systems. *Procedia-Social and Behavioral Sciences*, 143:521–526, 2014.

[Chen and Cui 2020] Fu Chen and Ying Cui. Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, 7(2):1–17, 2020.

[Cocea and Weibelzahl 2009] Mihaela Cocea and Stephan Weibelzahl. Log file analysis for disengagement detection in e-learning environments. *User Modeling and User-Adapted Interaction*, 19(4):341–385, 2009.

[Cristianini *et al.* 2000] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[Deepika and Sathyanarayana 2019] Kongara Deepika and Nallamothu Sathyanarayana. Classification and prediction of student academic performance using gray wolf optimization based relief-f budget random forest. 2019.

[Dewan *et al.* 2019] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1, 2019.

[Domingos 2012] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[Durgabai and Bhushan 2014] RPL Durgabai and Y Ravi Bhushan. Feature selection using relieff algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(10):8215–8218, 2014.

[Dutt and Ismail 2019] Ashish Dutt and Maizatul Akmar Ismail. Can we predict student learning performance from lms data? a classification approach. In *3rd International Conference on Current Issues in Education (ICCIE 2018)*, pages 24–29. Atlantis Press, 2019.

[Gardner and Brooks 2018] Josh Gardner and Christopher Brooks. Student success prediction in moocs. *User Modeling and User-Adapted Interaction*, 28(2):127–203, 2018.

[Gómez-Aguilar *et al.* 2015] Diego Alonso Gómez-Aguilar, Ángel Hernández-García, Francisco J García-Peñalvo, and Roberto Therón. Tap into visual analysis of customization of grouping of activities in elearning. *Computers in Human Behavior*, 47:60–67, 2015.

[GopalaKrishnan and Sengottuvelan 2016] T GopalaKrishnan and P Sengottuvelan. A hybrid pso with naïve bayes classifier for disengagement detection in online learning. *Program*, 2016.

[Guo *et al.* 2008] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

[Haiyang *et al.* 2018] Liu Haiyang, Zhihai Wang, Phillip Benachour, and Philip Tubman. A time series classification method for behaviour-based dropout prediction. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)*, pages 191–195. IEEE, 2018.

[Hashim *et al.* 2020] Ali Salah Hashim, Wid Akeel Awadh, and Alaa Khalaf Hamoud. Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering*, volume 928, page 032019. IOP Publishing, 2020.

[Hew *et al.* 2020] Khe Foon Hew, Xiang Hu, Chen Qiao, and Ying Tang. What predicts student satisfaction with moocs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, 145:103724, 2020.

[Hu and Li 2017] Min Hu and Hao Li. Student engagement in online learning: A review. In *2017 International Symposium on Educational Technology (ISET)*, pages 39–43. IEEE, 2017.

[Hu *et al.* 2016] Min Hu, Hao Li, Wenping Deng, and Hua Guan. Student engagement: one of the necessary conditions for online learning. In *2016 International Conference on Educational Innovation through Technology (EITT)*, pages 122–126. IEEE, 2016.

[Hussain *et al.* 2018] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, and Syed Muhammad Raza Abidi. Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience*, 2018, 2018.

[Izenman 2013] Alan Julian Izenman. Linear discriminant analysis. In *Modern multivariate statistical techniques*, pages 237–280. Springer, 2013.

[Kamath *et al.* 2016] Aditya Kamath, Aradhya Biswas, and Vineeth Balasubramanian. A crowdsourced approach to student engagement recognition in e-learning environments. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[Kang *et al.* 2020] Tingting Kang, Zhengang Wei, Jianxiong Huang, and Zhaoliang Yao. Mooc student success prediction using knowledge distillation. In *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, pages 363–367. IEEE, 2020.

[Kaur *et al.* 2018] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018.

[Kuhn *et al.* 2013] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[Kumar *et al.* 2018] A Dinesh Kumar, R Pandi Selvam, and K Sathesh Kumar. Review on prediction algorithms in educational data mining. *International Journal of Pure and Applied Mathematics*, 118(8):531–537, 2018.

[Longadge and Dongre 2013] Rushi Longadge and Snehalata Dongre. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*, 2013.

[Macarini *et al.* 2019] Buschetto Macarini, Luiz Antonio, Cristian Cechinel, Matheus Francisco Batista Machado, Vinicius Faria Culmant Ramos, and Roberto Munoz. Predicting students success in blended learning—evaluating different interactions inside learning management systems. *Applied Sciences*, 9(24):5523, 2019.

[Margulieux 2015] Lauren Margulieux. Mixing in-class and online learning: Content meta-analysis of outcomes for hybrid, blended, and flipped courses. 2015.

[Mining 2012] Through Educational Data Mining. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *Proceedings of conference on advanced technology for education*, pages 1–64, 2012.

[Mongkhonvanit *et al.* 2019] Kritphong Mongkhonvanit, Klint Kanopka, and David Lang. Deep knowledge tracing and engagement with moocs. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 340–342, 2019.

[Moubayed *et al.* 2018] Abdallah Moubayed, MohammadNoor Injadat, Abdallah Shami, and Hanan Lutfiyya. Relationship between student engagement and performance in e-learning environment using association rules. In *2018 IEEE World Engineering Education Conference (EDUNINE)*, pages 1–6. IEEE, 2018.

[Osguthorpe and Graham 2003] Russell T Osguthorpe and Charles R Graham. Blended learning environments: Definitions and directions. *Quarterly review of distance education*, 4(3):227–33, 2003.

[Owston and York 2018] Ron Owston and Dennis N York. The nagging question when designing blended courses: Does the proportion of time devoted to online activities matter? *The Internet and Higher Education*, 36:22–32, 2018.

[Park and Kim 2020] Taejung Park and Chayoung Kim. Predicting the variables that determine university (re-) entrance as a career development using support vector machines with recursive feature elimination: The case of south korea. *Sustainability*, 12(18):7365, 2020.

[Park 2014] Yeonjeong Park. Analysis of online behavior and prediction of learning performance in blended learning environments. *Educational Technology International*, 15(2):71–88, 2014.

[Porter *et al.* 2014] Wendy W Porter, Charles R Graham, Kristian A Spring, and Kyle R Welch. Blended learning in higher education: Institutional adoption and implementation. *Computers & Education*, 75:185–195, 2014.

[Raj and Evangeline 2020] Pethuru Raj and Preetha Evangeline. *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*. Academic Press, 2020.

[Ramaswami and Bhaskaran 2009] M Ramaswami and R Bhaskaran. A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*, 2009.

[Romero and Ventura 2013] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[Romero *et al.* 2013] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472, 2013.

[Romero *et al.* 2014] Cristóbal Romero, José Raúl Romero, and Sebastián Ventura. A survey on pre-processing educational data. In *Educational data mining*, pages 29–64. Springer, 2014.

[Seiffert *et al.* 2008] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. A comparative study of data sampling and cost sensitive learning. In *2008 IEEE International Conference on Data Mining Workshops*, pages 46–52. IEEE, 2008.

[Sheshadri *et al.* 2019] Adithya Sheshadri, Niki Gitinabard, Collin F Lynch, Tiffany Barnes, and Sarah Heckman. Predicting student performance based on online study habits: a study of blended courses. *arXiv preprint arXiv:1904.07331*, 2019.

[Silva *et al.* 2016] João C Sedraz Silva, Jorge LC Ramos, Rodrigo Lins Rodrigues, Alex Sandro Gomes, Fernando da Fonseca de Souza, and Alexandre Magno Andrade Maciel. An edm approach to the analysis of students' engagement in online courses from constructs of the transactional distance. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pages 230–231. IEEE, 2016.

[Sinha *et al.* 2014] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv preprint arXiv:1407.7131*, 2014.

[Soffer and Cohen 2019] Tal Soffer and Anat Cohen. Students' engagement characteristics predict success and completion of online courses. *Journal of Computer Assisted Learning*, 35(3):378–389, 2019.

[Sperandei 2014] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1):12–18, 2014.

[Van Goidsenhoven *et al.* 2020] Steven Van Goidsenhoven, Daria Bogdanova, Galina Deeva, Seppe vanden Broucke, Jochen De Weerdt, and Monique Snoeck. Predicting student success in a blended learning environment. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 17–25, 2020.

[Vo *et al.* 2017] Hien M Vo, Chang Zhu, and Nguyet A Diep. The effect of blended learning on student performance at course-level in higher education: A meta-analysis. *Studies in Educational Evaluation*, 53:17–28, 2017.

[Waheed *et al.* 2020] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. Predicting academic performance of students from vle big data using deep learning models. *Computers in Human Behavior*, 104:106189, 2020.

[Webb *et al.* 2010] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15:713–714, 2010.

[Wefald and Downey 2009] Andrew J Wefald and Ronald G Downey. Construct dimensionality of engagement and its relation with satisfaction. *The Journal of Psychology*, 143(1):91–112, 2009.

[Wellman *et al.* 2014] Richard W Wellman, Kelly D Phillipps, and David B Gonzalez. *Machine Learning for Student Engagement*, July 24 2014. US Patent App. 13/749,618.

[Wells *et al.* 2016] Marc Wells, Alex Wollenschlaeger, David Lefevre, George D Magoulas, and Alexandra Poulovassilis. Analysing engagement in an online management programme and implications for course design. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 236–240, 2016.

[Widyahastuti and Tjhin 2017] Febrianti Widyahastuti and Viany Utami Tjhin. Predicting students performance in final examination using linear regression and multilayer perceptron. In *2017 10th International Conference on Human System Interactions (HSI)*, pages 188–192. IEEE, 2017.

[Yahaya *et al.* 2020] Che Akmal Che Yahaya, Che Yahaya Yaakub, Ahmad Firdaus Zainal Abidin, Mohd Faizal Ab Razak, Nuresa Fatin Hasbullah, and Mohamad Fadli Zolkipli. The prediction of undergraduate student performance in chemistry course using multilayer perceptron. In *IOP Conference Series: Materials Science and Engineering*, volume 769, page 012027. IOP Publishing, 2020.

[Zacharis 2016] Nick Z Zacharis. Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications*, 7(5):17–29, 2016.