PREDICTIVE MODELLING OF STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING APPROACHES: A CASE STUDY IN UNIVERSITI ISLAM PAHANG SULTAN AHMAD SHAH

NURUL HABIBAH BINTI ABDUL RAHMAN

MASTER OF SCIENCE

UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	NURUL HABIBAH BINTI ABDUL RAHMAN
Date of Birth	:	15TH FEBRUARY 1985
Title	:	PREDICTIVE MODELLING OF STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING APPROACHES:A CASE STUDY IN UNIVERSITI ISLAM PAHANG SULTAN AHMAD SHAH
Academic Session	:	SEMESTER II 2023/2024

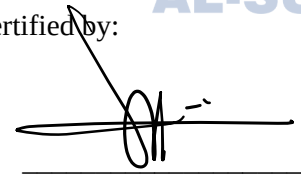
I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☒	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang Al-Sultan Abdullah reserves the following rights:

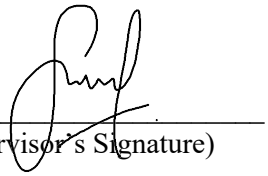
1. The Thesis is the Property of Universiti Malaysia Pahang Al-Sultan Abdullah
2. The Library of Universiti Malaysia Pahang Al-Sultan Abdullah has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:



(Student's Signature)

Date: 3rd July 2024



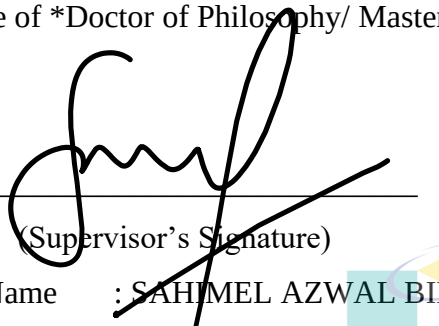
(Supervisor's Signature)

Sahimel Azwal bin Sulaiman
Name of Supervisor
Date: 03/07/2024

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

I/We* hereby declare that I/We* have checked this thesis/project* and in my/our* opinion, this thesis/project* is adequate in terms of scope and quality for the award of the degree of *Doctor of Philosophy/ Master of Science.

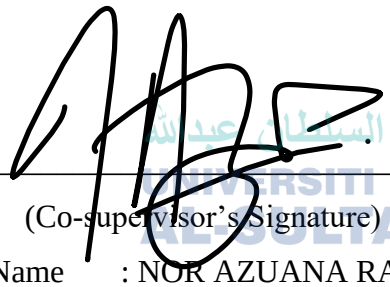


(Supervisor's Signature)

Full Name : SAHIMEL AZWAL BIN SULAIMAN

Position : SENIOR LECTURER

Date : 1/7/2024



(Co-supervisor's Signature)

Full Name : NOR AZUANA RAMLI

Position : SENIOR LECTURER

Date : 1/7/2024

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang Al-Sultan Abdullah or any other institutions.

(Student's Signature)

Full Name : NURUL HABIBAH BINTI ABDUL RAHMAN

ID Number : MSS21002

Date : 1/7/2024

PREDICTIVE MODELLING OF STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING APPROACHES:

A CASE STUDY IN UNIVERSITI ISLAM PAHANG SULTAN AHMAD SHAH

NURUL HABIBAH BINTI ABDUL RAHMAN

Thesis submitted in fulfillment of the requirements

for the award of the degree of

Master of Science

Centre for Mathematical Sciences

UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

JULY 2024

ACKNOWLEDGEMENTS

By the name of Allah, the Most Gracious and the Most Merciful. All praise to Allah for granting me the chance and fortitude to pursue and accomplish my research proposal for my master in Statistics.

First and foremost, I would like to thank my supervisor, Dr. Sahimel Azwal bin Sulaiman, and my co-supervisor, Dr. Nor Azuana binti Ramli, for providing guidelines and reviewing all of my inadequate submissions with patience. It was a privilege for me to work under such brilliant supervisors who assisted me throughout my studies.

Very special thanks to my beloved husband, Mohd Shahid bin Mohamed Fathil and my parents for their support and prayers during my studies. Without their will, it would not have been easy for me to complete my studies.

Last but not least, I would like to convey my appreciation to my son, Muhammad Ziyad Iman bin Mohd Shahid and all friends who have inspired and supported me during my studies and made it possible to complete this journey.

ABSTRAK

Sejak kebelakangan ini, analisis ramalan telah semakin popular dalam bidang pengajian tinggi kerana ianya dapat menyediakan maklumat yang sangat membantu kepada pendidik dan ia juga berpotensi membantu mereka menambahbaik prestasi pelajar. Berdasarkan kajian literatur, kajian mengenai mesin pembelajaran dan analisis ramalan untuk menambahbaik prestasi pelajar di peringkat pengajian tinggi di Malaysia adalah terhad. Di samping itu, peningkatan kadar keciciran di kalangan pelajar-pelajar adalah isu yang penting di Institusi-institusi Pengajian Tinggi. Dengan kadar keciciran yang tinggi ini, reputasi institusi pendidikan akan merundum. Tambahan pula, ia boleh menyebabkan kehilangan modal insan yang ketara kepada negara. Sasaran utama kajian ini ialah membangunkan model ramalan yang paling tepat bagi meramal tahap prestasi pelajar menggunakan teknik pembelajaran mesin seperti regresi logistik multinomial, pokok Keputusan, Hutan Rawak, jiran terdekat K, Naïve Bayes, dan mesin vektor sokongan. Kajian ini telah menggunakan kaedah Korelasi Cramer's V dan Pekali Korelasi Pangkat Spearman bagi menentukan faktor yang paling mempengaruhi tahap prestasi pelajar. Metrik penilaian prestasi merangkumi ketelitian, pengingatan, ketepatan, skor F1, dan luas di bawah keluk operasi penerima ciri. Berdasarkan set data yang merangkumi pelajar yang mendaftar dalam kursus Statistik Perniagaan di Universiti Islam Pahang Sultan Ahmad Shah dari 2013 hingga 2022, kajian ini telah menentukan bahawa markah terkumpul pelajar sebagai faktor yang paling berpengaruh dalam menentukan tahap prestasi pelajar. Secara khususnya, pokok keputusan dikenalpasti sebagai model ramalan yang paling tepat, yang mempunyai nilai ketepatan 0.60. Model tersebut juga turut mendapat skor yang tinggi bagi pengingatan dan skor F1 berbanding model-model yang lain. Akhirnya, empat model telah mendapat skor sempurna, 1.00 bagi luas di bawah keluk operasi penerima ciri untuk membezakan gred pelajar gagal. Pada hujung kajian ini, dicadangkan supaya kajian akan datang dapat menilai semula model ramalan ini dengan menumpukan penambahan pembolehubah atau teknik yang dapat membantu menimkatkan ketepatan ramalan. Algoritma ramalan tersebut juga boleh dibangunkan melalui Sistem Pengurusan Pembelajaran bersama-sama papan pemuka supaya dapat memudahkan analisis pada masa akan datang.

ABSTRACT

Recently, predictive analytics research has grown in popularity in higher education because it provides helpful information to educators and potentially assists them in enhancing student achievement. Based on the literature review, studies on machine learning and predictive analytics to improve student performance are still scarce in Malaysian higher education. Besides that, the increment of dropout rates among students is crucial issue in Higher Education Institutions. With a huge number of students drop out, the higher education institution's reputation might be dropped. Furthermore, it may cause a significant loss of human capital for the country. The main goal of the study was to develop the most accurate predictive model for predicting students' performance levels using machine learning techniques such as multinomial logistic regression, decision trees, Random Forest, k-nearest neighbor, Naïve Bayes, and support vector machine. This study used Cramer's V correlation and Spearman's Rank Correlation Coefficient to determine the most correlated factor towards students' performance level. Evaluation metrics encompass precision, recall, accuracy, F1-score, and area under the receiver operating characteristics curve. Drawing from a dataset spanning students enrolled in the Business Statistics course at Universiti Islam Pahang Sultan Ahmad Shah from 2013 to 2022, this study identifies students' carry marks as the most correlated factor in determining performance levels. Particularly, the decision tree is identified as the most accurate predictive model, having a 0.60 accuracy value. The model also has the highest value for recall and F1-score compared to other models. Finally, four models, namely multinomial logistic regression, decision tree, Random Forest, and Naïve Bayes, have perfect scores, 1.00 of area under the receiver operating characteristics curve to distinguish fail grade students. At the end of this study, it is recommended that future research might reassess the model by considering additional variables or techniques that may help improve the predictive accuracy. The predictive algorithm can also be added to the Learning Management System along with a dashboard so that it is easier to do analyses in the future.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

θ	Theta
n	Total Number
D	Outcome Variable
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
X_k	Predictor Factors
ρ	Rho
φ	Phi Coefficient
χ^2	Chi Square
k	Number of Column
r	Number of Row
d_i	Distance between Two Ranks
P_i	Probability of i^{th} Class
t	Nodes
K	Number of Classes
I_G	Gini Index
H	Entropy
z_K	Logit for K
e	The Base of Euler's Number

LIST OF ABBREVIATIONS

DM	Data Mining
UnIPSAS	Universiti Islam Pahang Sultan Ahmad Shah
LMS	Learning Management System
ML	Machine Learning
DT	Decision Tree
RF	Random Forest
LR	Logistic Regression
MLR	Multinomial Logistic Regression
k-NN	k-Nearest Neighbor
SVM	Support Vector Machine
NB	Naïve Bayes
HEIs	Higher Education Institutions
MOE	Ministry of Education
VLEs	Virtual Learning Environments
BS	Business Statistics
BM	Business Mathematics
CGPA	Cumulative Grade Point Average
GPA	Grade Point Average
CM	Carry Marks
IQR	Interquartile Range
Q1	First Quartile
Q3	Third Quartile
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
AUC	Area Under Curve
ROC	Receiver Operating Characteristics
SRCC	Spearman's Rank Correlation Coefficient
MSE	Mean Squared Error
MAE	Mean Absolute Error

RMSE	Root Mean Squared Error

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Research Background

Predictive analytics is a part of data mining (DM) that is able to predict outcomes or possibilities based on datasets. This analysis recommends the prediction of future events by developing predictive models using statistical modeling, DM and Machine Learning (ML) approaches. DM is a field which perceives knowledge from large historical data. This field addresses the inquiry of how good to use the historical data to determine common routines and enhance the decision-making process (Mitchell et al., 1999). On the other hand, ML is known as a scientific method which involve the practice of how machines acquire knowledge or learn from the occurrence (Kavakiotis et al., 2017). Utilizing ML and DM approaches in predictive analytics is crucial to provide tools to build accurate predictive models. The predictive model is formed based on the data used as well as the objective of the predictive analysis which is either regression or classification techniques (Mishra et al., 2012).

Predictive analytics research has gained popularity in higher education in recent years for its capacity to give educators valuable insights and potentially aid in bolstering students' academic success. Educators have the opportunity to utilize predictive analytics to establish effective measures aimed at enhancing student performance, averting dropout instances, and ensuring student retention (Abdul Bujang et al., 2021). In higher education environments, two key application purposes of DM and ML may be identified which are predictors and early warning systems. A predictor seeks to forecast the outcome of a course given a specific set of input data, whereas an early warning system performs the same tasks as a predictor and reports its findings to teachers and students at an early stage. Therefore, efforts to avoid or minimize possibly unfavorable consequences can be done. Student's performance, risk of failing a course, dropout risk, grade prediction, and graduation rate are all common prediction targets in higher education sector (Tatar et al., 2020).

Each university has its database system that allows student's data to be accessed by lecturers and academic administrators of the university. Therefore, this institution should utilize the database as well as possible to guide them in making decisions on various aspects, especially those related to students' academic improvement. There is no doubt that the trend in the use of ML methods to predict student performance among public and private universities in Malaysia is increasing, however, as per a literature review, there remains no evidence to suggest that this technique has been implemented in any Islamic universities. Students from Islamic universities and public universities commonly have slightly different information on their cultural and educational backgrounds. Most of the students from Islamic universities do not excel in subjects involving science and mathematics. Therefore, developing a predictive model to identify the failure at an early stage is very important.

Generally, a student's performance for a certain course is determined by the student's grade for the course or by classifying whether they pass or fail the course. Preliminary predictions of students' performance, relying on their accumulated marks, past achievements, and various other factors, proved beneficial for students and lecturers in early actions to enhance their achievement before the examination. It enabled students to prepare and strengthen their carry marks to improve their final grades. Moreover, lecturers could focus more on students predicted to have lower grades. Therefore, the goal of this study is to develop a predictive model for students' performance levels which according to students' final grades for a selected course and details will be discussed further in the rest of this thesis.

1.2 Problem Statement

Dropout rates among students is a big issue in Higher Education Institutions (HEIs). According to the New Straits Time reports on 2nd August 2022, the Dewan Rakyat was presented that 17,613 undergraduate students failed to complete their studies in 2021. The figure shows an increment of more than 4,000 dropouts from year 2020. Dropouts have a detrimental impact on both educational institutions and stakeholders. Moreover, with the prevalence of virtual learning methods in today's education system, dropout rates continue to rise (Kabathova et al., 2021). Numerous factors, such as academic performance, health, family, and personal reasons, can contribute to dropout rates, which vary depending on the field of study and the higher education institution. If

a substantial number of students drop out of their respective universities, it could tarnish the reputation of the higher education provider. Furthermore, dropping out of school would lead to a significant loss of human capital for the country, as public universities would produce fewer professionals and experts (Sani et al., 2020).

Generally, when students fail, they must retake the subject in the subsequent semester. It incurs additional university expenses and burdens various parties, including parents and lecturers. Moreover, the financial pressure on the family escalates as the college loan needs to be repaid despite the student not completing their studies. Household income in Malaysia is categorized into the groups of Bottom 40% (B40), Middle 40% (M40) and Top 20% (T20). Department of Statistics Malaysia have been reported that the B40 household median income in 2023 was less than RM6,338.00. This amount is very burdensome compared to the amount of fees that need to be paid if their children are studying in a private university which is not less than RM8,000 to RM10,000 per year. Consequently, students are less inclined to maintain motivation in their studies if they know their parents cannot afford the fees.

Through the grading system at Universiti Islam Pahang Sultan Ahmad Shah (UnIPSAS), the passing point for Grade Point Average (GPA) was 2.00. According to the record of student results for the subject of Business Statistics (BS), the percentage of students who got at least grade C (grade point 2.00) for the year 2022 has decreased by 12% compared to the previous year. In addition, students who failed (grade E) have increased from 5% in 2021 to 8% in 2022. The total number of students who took this subject for both years was an average of 25 students. BS is a core subject for all courses in management field under the Faculty of Management and Informatics at UnIPSAS. Failure to get a good pointer, which is a pointer of 2.00 and above, can affect the student's GPA for the semester and subsequently can affect their Cumulative Grade Point Average (CGPA).

Currently, in Malaysian higher education, research on predictive analytics through ML techniques to enhance students' achievement are still scarce (Abdul Bujang et al., 2021). Typically, in every university, students' grades and final marks are revealed following the final examination, indicating that students become aware of their achievements only after the faculty has announced their grades. At that time, students just have to accept if they fail and lecturers cannot offer any help to improve their results.

Whereas, there is a vast of students' data that can be used to identify trends and patterns of student achievement more earlier using various statistical methods. Based on that reason, student's performance predictive analytics is required to identify students with a high likelihood of dropping out. This predictive analytics needs to be done to prevent student failure so that students can avoid the burden of paying higher fees to repeat the subject. At the same time can improve student's GPA and CGPA so that they can graduate with excellence.

One of the most effective techniques in predicting students' academic performance, which has an impact on early intervention, personalized learning, and educational policy decisions is machine learning models. Although machine learning models are an effective tool for forecasting students' performance, attaining high accuracy rates is still very difficult. A number of factors can affect how accurate predictive models are in this area, including feature selection, data quality, model complexity, and result interpretability. This study will also look at how accurate the model's prediction rate is against real data where the number is also limited. Further research is also necessary to determine whether these models can be applied to diverse student populations and how to balance model complexity and forecast accuracy.

1.3 Research Questions

This study is conducted to answer the following research questions:

1. What are the factors that affect the students' performance level in the Business Statistics course?
2. What is the most accurate predictive model for students' performance level in the Business Statistics course at UniPSAS?
3. How to evaluate the performance of the proposed models?

1.4 Research Objectives

This study focuses on the following objectives:

1. To determine the factors that affect the students' performance level in the Business Statistics course by using Cramer's V and Spearman's Rank Correlation Coefficient.
2. To propose the most accurate predictive model for students' performance level in the Business Statistics course at UniPSAS by using the machine learning techniques.
3. To evaluate the performance for the proposed models using the value of precision, recall, accuracy, F1-score, and area under receiver operating characteristics curve (AUC) as well as new dataset.

1.5 Research Scope

The case study focuses on students enrolled in the BS course at UnIPSAS. It was conducted at UnIPSAS using a database from the UnIPSAS LMS called eCampus UnIPSAS, as well as lecturer's records for the BS course. The course's records from the years 2013 to 2022 were analysed to develop a predictive model for the students' performance levels using ML techniques. The dataset was analysed using Python software and libraries such as Numpy, Scikit-learn, Pandas, and Matplotlib. The BS course is taken by students from diploma programmes in Business Study, Finance and Banking, Marketing Management, and Accounting.

1.6 Research Significance

With the increasing accessibility of student data, there is a need to understand and utilize it to gain insights into the educational landscape. This study is very important to various parties such as the government, specifically to the Malaysian Ministry of Education (MOE), the industry, the higher education provider, particularly UniIPSAS, parents, as well as other researchers in the same field. This research has the potential to inspire proactive measures by identifying failed students earlier. Consequently, it can diminish the likelihood of these students dropping out and discontinuing their studies.

The MOE in Malaysia is always committed to ensuring the academic success of graduates. With the development of technologies that continue to play an important role in the field of education, the integration of predictive analytics studies can enable the ministry to continuously refine education strategies and policies in Malaysia. Therefore, the educational system in Malaysia will remain responsive to the needs of students, faculties, and also various industries. In addition, MOE has reported the yearly outcome of Graduate Employability survey and always strives to increase the employability rate of graduates in Malaysia. By conducting the predictive analytics research for students' performance during their studies in order to prevent failure of certain courses, it will improve students' retention rate and next will ensure students' graduation on time. As the consequences, students who have graduated with good results will be able to continue their careers in various fields and it may help to increase the Graduate Employability rates.

Predictive analytics is an effective method to improve the way higher education providers deal with student achievement issues. Typically, in the context of HEIs in Malaysia, educators are required to analyse the academic performance of students reviewed via final examination directly from a large database at the end of each semester. Nevertheless, this database is insufficient for determining analytics, insights, and trends about student success or failure. Therefore, in order to overcome the weakness, it is necessary to obtain an effective decision that benefits both the institution and the students. This study will assist UniPSAS as a higher education provider in estimating students' performance, and next will help them to gain insight and trends of students' performance at an early stage.

In addition, this study will also benefit UnIPSAS's students in making their initial preparations before the final examination. With the predictive model developed in this study, not only their grades will be predicted, but they can also predict the GPA and CGPA for the current semester. Therefore, it is clear that this study is very important for students to improve their achievements for each course taken. It can reduce the rate of students who have to repeat subjects and get a poor GPA.

1.7 Thesis Organization

All chapters included in this thesis are illustrated in Figure 1.1.

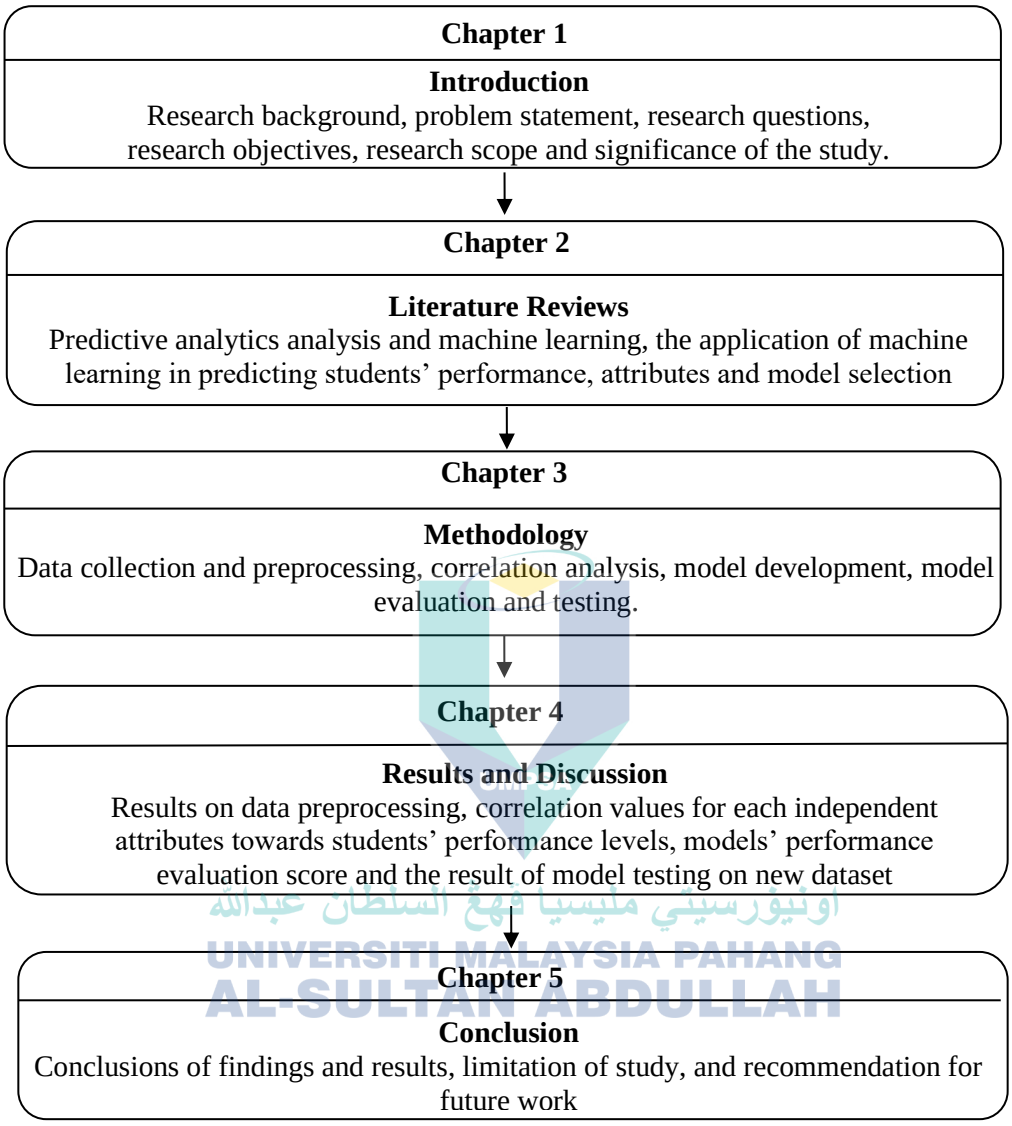


Figure 1.1 Thesis Organization

CHAPTER 2

LITERATURE REVIEWS

2.1 Introduction

This chapter begins with an overview of predictive analytics research and a general description of the purpose of predictive analytics in several sectors, including the higher education sector. Focusing on the related techniques that have been used in this study, this chapter also discusses the application of machine learning (ML) methods in predicting student performance based on previous studies, as well as the discussion of studies related to attribute selection to build the predictive models for students' grades, using ML techniques.

2.2 Predictive Analytics Research

Today's world is filled with data, much like the air. With data capture and consumption growth in the digital age, organizations have turned to predictive analytics more than ever before. This powerful tool enables them to enhance performance, streamline operations, mitigate risks, and identify fraudulent activities. Furthermore, predictive analytics for making decisions has garnered significant attention recently. Poornima and Pushpalatha (2018) explained predictive analytics as a set of statistical and analytical tools for generating unique approaches for predicting future outcomes. Furthermore, predictive analytics also defined as a sort of big data analytics that involves collecting information from data in order to forecast trends and behavior patterns. Predictive analytics determines the likelihood of a condition occurring or the likely outcome of an occurrence in the future. Based on Figure 2.1, Poornima and Pushpalatha (2018) also described those predictive analytics involves six steps in its process. Following this, Mishra et al. (2012) outlined a predictive analytics process comprising four stages: gathering and pre-processing raw data, converting pre-processed data into a format compatible with the chosen ML technique, building models with the transformed data, and delivering forecasts to users based on the developed learning model.

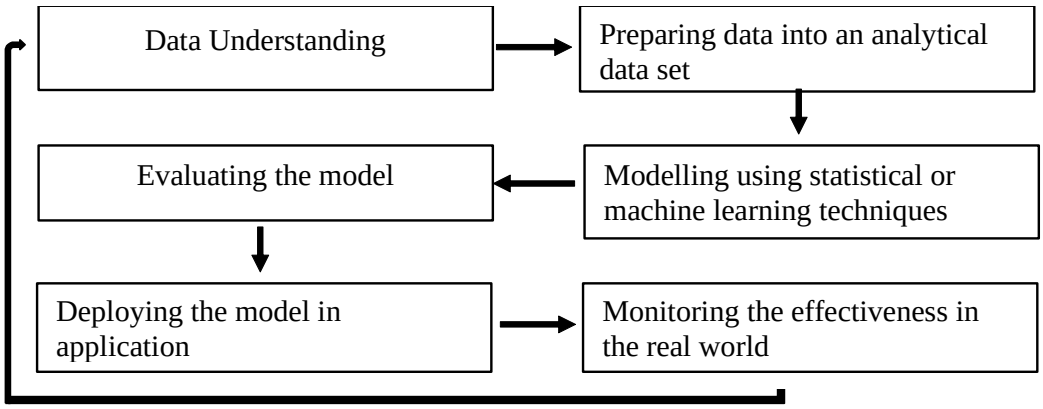


Figure 2.1 Predictive Analytics Process

Source: Poornima and Pushpalatha (2018)

According to Shmueli and Koppius (2011), predictive analytics involves statistical models and other empirical methods to achieve empirical forecasts, as well as methods for evaluating the accuracy of those predictions in reality. They also summarized explanatory statistical models (models are trained on the fundamental effect - cause relationships between theoretical constructs) and predictive analytics modelling (models are focused on correlations between measurable variables) into five different functions: analysis goal, variables of interest, model building optimized function, model building constraints, and model evaluation. The different functions and contexts in which explanatory modeling and predictive analytics are built and operate lead to many differences in the model building process, which translate into different final models and different power evaluations. Table 2.1 summarizes the differences between explanatory statistical modelling and predictive analytics as described by Shmueli and Koppius (2011) in their research.

Table 2.1 The Differences of Explanatory Statistical Modeling and Predictive Analytics

Function	**Explanatory Modeling**	**Predictive Analytics**
Analysis goal	Causal models were tested using explanatory statistical models.	Predictive model is used to evaluate predictability levels and forecast new data.
Variables of interest	Operationalized variables are simply employed as tools to investigate the initial theoretical entities and their relationships.	The focus is on the observable and measurable variables.
Model building optimized function	The goal of explanatory modelling is to reduce model bias. Type I and II errors are the riskiest.	The goal of predictive modelling is to reduce the combined bias and variance. Over-fitting is the biggest threat.
Model building constraints	The empirical model must be understandable, support statistical testing of the hypotheses of interest, and be consistent with the theoretical model.	Variables that are available at the time of model deployment should be used.
Model evaluation	Explanatory power evaluated by the strength-off it analyzes and tests	The accuracy of out-of-sample predictions is often used to evaluate predictive power.

Due to the fact that predictive analytics has been around for decades, it is a technology that has now reached its peak. One of the reasons is the growing volume and variety of data and the interest in leveraging data to generate variable insights. Additionally, predictive analysis is also an interactive method and easy to use with computer software. Therefore, predictive analytics is more than just the domain of mathematicians and statisticians; business analysts and professionals from many fields also use these technologies.

2.3 Predictive Analytics using Machine Learning Techniques in Higher Education Institutions.

Throughout human's evolution, they have utilized various tools to simplify specific tasks. The ingenuity of the human brain has spurred the development of multiple technologies, which in turn have facilitated numerous aspects of life, including travel, industry, and computing. Machine learning (ML) is among these transformative technologies. Large volumes of data were included in developing the best prediction analytics using ML techniques, which avoids many of the errors and limitations of traditional modelling techniques (Kendale et al., 2018).

As per Mahesh (2018), ML is a field of study that empowers computers to learn without explicitly programming patterns. ML serves as a method for instructing machines to manage data effectively. Logistic regression (LR), support vector machine (SVM), Random Forest (RF), Naïve Bayes (NB), decision trees (DT), and k-nearest neighbor (k-NN) are examples of algorithms used in machine learning. ML has been used to develop predictive models in various sectors, including health, business, education, and many other fields. As indicated by a study conducted by the Center for Digital Technology and Management (2015), the rise in the volume of education data due to digitalization has led to a heightened utilization of machine learning in education (Mduma et al., 2019).

There is no doubt that ML is making its way into a wide range of industries, including the education industry, and institutions of higher learning are no exception. In education, ML approaches have been used in hundreds of studies on anything from student enrollment to graduation predictions, failure rates, retention, and performance. ML also greatly aids Higher Education Institutions (HEIs) in decision making, particularly at the level of stakeholders, management, deans, and department heads.

The number of academic offers that prospective students turn down is rising every year. Based on previous students' intake data, research by Basheer et al. (2019) attempts to forecast whether a student would accept or reject an academic offer from a university. The experiment was to identify the best model to predict whether students would accept or reject the offer using DT and k-NN algorithms. Both algorithms have shown the best accuracy of 66% with fifteen selected attributes, which include: applicants' gender, applicants' SPM stream, university campuses, applicants' hometowns, disabilities,

campus visits, and the order in which courses were chosen in the application form, as well as orphan and acceptance status. Besides that, Esquivel et al. (2021) evaluated of the many factors impacting the admission status of freshman applicants at the Philippine University. The decision support system was built with the LR as the preference algorithm. As a result, an accuracy rating of 80.5 percent was achieved by using Weka's variable set to forecast the enrollment of applicants. There are many intelligent ways in which this kind of research can be used to improve the university's academic admissions process.

The decisions made in strategic planning affect the policies, strategies, and actions of HEIs (Nieto et al., 2019). An analysis of South American undergraduate engineering graduation rates using three supervised classification methods namely DT, LR and RF was presented in Nieto et al. (2019) research. The area under the Receiver Operating Characteristic (ROC) curve of RF was the best result, with 84.11% score. The early detection of students who are unlikely to graduate using these studies is highly effective. However, from this prediction, other aspects such as students' academic performance and dropout rates might be examined.

Nowadays, Virtual Learning Environments (VLEs) are sophisticated online education platforms that enable teachers to deliver courses that include well-managed resources and a plethora of engaging activities. Nowadays, the needs of the VLEs are particularly relevant and the trend is likely to continue in the future. However, dropout rates are a severe issue in contemporary e-learning systems as Kabathova et al. (2021) stated in their research which aimed to predict student dropouts in VLEs. The prediction accuracy of RF and LR achieved the highest score of 0.93 compared to NB, Neural Network, SVM, and DT. The information from this research may be used to develop a course recommendation. In other words, predicting students' performance should be the ultimate focus of all educational institutions worldwide (Ofori et al., 2020).

2.4 The Application of Machine Learning Models in Predicting Students' Performance

Predicting students' performance should be the ultimate focus of all educational institutions worldwide (Ofori et al., 2020). Khan et al. (2015) developed a predictive model for the final grade of secondary school using the J48 DT algorithm. As a result, the accuracy rate obtained was 84.53%, which is considered high accuracy. However, it is essential to evaluate the accuracy of multiple machine learning models to determine which model is the best to predict students' performance (Ofori et al., 2020). Some of the ML techniques that have been used by most researchers who make predictions for student performance include DT, RF, NB, SVM, LR, and k-NN.

2.4.1 Decision Tree

DT, known for its simplicity and ease of understanding, stands out as one of the most commonly employed prediction techniques. Researchers frequently rely on this approach to analyze small and large datasets and forecast values (Shahiri et al., 2015). Research conducted by Hamsa et al. (2016) revealed that, when predicting students' academic performance, DT exhibited a higher level of accuracy than the Fuzzy Genetic Algorithm. This conclusion was drawn from a study involving 120 students from bachelor's degree programs and 48 students from master's degree programs. Abana et al. (2019) created a classification model to forecast students' grades using DT method namely RepTree, Random Tree, and J48. The prediction model was tested on 133 samples with five attributes. Random Tree achieved the optimum accuracy of 75.188 percent, and considered as the best model compared to other DT models.

2.4.2 Random Forest

The RF is an algorithm for learning in groups. It is a method for supervised classification and made up of a large number of decision trees that have been constructed at random. RF produces and mixes many decision trees to improve forecast accuracy and consistency (Ünal, 2020). A finding from Zabriskie et al. (2019) has pointed out that RF performed better than logistic models by mixing in-class factors with institutional variables . Research by Xu et al. (2017) compared RF with LR, and k-NN method and found that RF was the best algorithm with the lowest Mean Squared Error (MSE) value.

2.4.3 Naïve Bayes

The NB method is a straightforward probabilistic classifier based on the Bayes theorem with strong as well as Naive assumptions of independence. NB is an inductive learning technique in ML and DM. This algorithm utilizes the theory of Bayesian probability to predict future likelihood by leveraging prior experience (Ofori et al., 2020). Pojon (2017) research sought to identify the most efficient comparison-based prediction algorithm for forecasting students' performance based on GPA and final grade. The outputs were set as the final mark and performance category. The results indicated that the NB classification algorithm performed the best in his first dataset, achieving an accuracy of 98 percent, while the DT algorithm was most effective in his second dataset, with an accuracy of 78 percent.

2.4.4 Support Vector Machine

The SVMs are a group of supervised learning algorithms that can be used for classification and regression (Ofori et al., 2020). The SVM algorithm aims to find a hyperplane in a N-number of variables (N-dimensional space) that distinctly classifies the data points. Hyperplanes are decision boundaries that assist in the classification of data items. Data points that lie on either side of the hyperplane may be assigned to distinct classes. Additionally, the size of the hyperplane depends on the number of attributes. Support vectors are data points that are closer to the hyperplane and influence its position and orientation. To build an SVM, the margin of the classifier needed to be maximized using these support vectors. The position of the hyperplane would change by adding or removing support vectors. Figure 2.2 shows the possible hyperplanes and the support vectors.

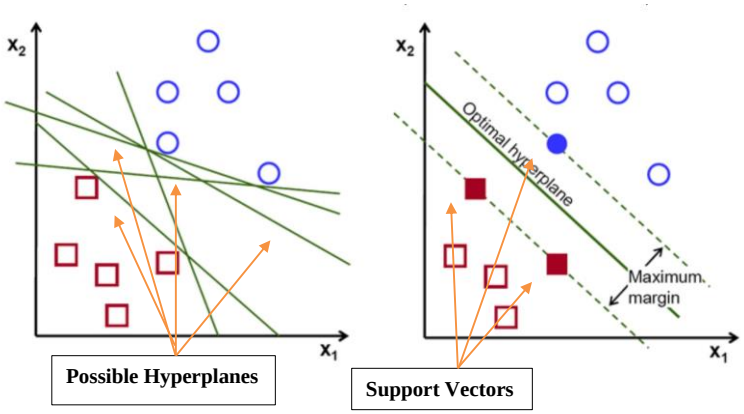


Figure 2.2 Possible Hyperplanes and Support Vectors

Source: https://www.analyticsvidhya.com/blog/2021/10/Support-vector-machinessvm-a-complete-guide-for-beginners/

Several research conducted by Abu Zohair et al. (2021), Venkat et al. (2018), Barnabas et al. (2018), and Anderson et al. (2017) established that this algorithm was the best model for predicting students' performance when compared to other algorithms such as NB, k-NN and DT.

2.4.5 Logistic Regression

The LR is a machine learning method that is used for classification problems. It is a predictive analytic technique and relies on the theory of probability. There are three primary LR types, namely binary, multinomial, and ordinal. They vary in theory and implementation. In binary regression, there are simply two potential values: yes or no. Multinomial logistic regression involves at least three values, i.e., the cat's food preference (wet food, dry food, or human food). In the case of ordinal regression, there is an association between the levels (Hosmer and Lemeshow, 2000).

Research by Zabriskie et al. (2019) employed RF and LR models to construct a predictive model of students' performances in Physics 1 and Physics 2 courses at a large eastern land-grant university. By combining variables such as homework grades and Cumulative Grade Point Average (CGPA), the study came up with an LR model to predict whether a student will receive a grade lower than "B" in the course, with 73% accuracy in Physics 1 (915 datasets) and 81% accuracy in Physics 2 (805 datasets).

Yildiz et al. (2020) studied ML algorithms to predict academic achievement of 421 students. This research aimed to estimate course success (yes or no) using the Neural

Network, k-NN, LR, SVM, DT, RF, and NB. When the prediction accuracy of machine learning algorithms was compared, the findings indicated that LR achieved a 78.4% accuracy value. In addition, LR was identified as the best algorithm to predict students' performance by Dhilipan et al. (2021), with a 97.05% accuracy value compared to DT entropy and k-NN.

2.4.6 k-Nearest Neighbor

The k-NN method stands out as one of the foundation and significant classification algorithm in ML. According to Dhilipan et al. (2021), the k-NN method is non-parametric and does not make assumptions regarding data distribution. Their research demonstrated a high accuracy rate of 93.7 percentfor student grade prediction using previous semester marks as predictor attribtes. However, in a study by Xu et al. (2017), the k-NN method exhibited the lowest accuracy compared to RF, LR, and Linear Regression models.

2.5 Attributes Selection and Data Sizes to Predict Students' Performance using Machine Learning

By studying previous data for future betterment, predictive analytics can improve and increase the quality of students' academic performance. Abdul Bujang et al. (2021) stated that various predictive analytics studies have been conducted utilizing ML to forecast student academic performance for the institution in order to improve the quality of decision-making. This research used demographic variables such as student's identification number, year intake, cohort, gender, and continuous marks as the independent variable to predict students' final grades for 489 students. As a result, the DT (J48) algorithm returned the highest accuracy compared to another method, namely RF, SVM, and LR. Next, Altabrawee et al. (2019) research found that DT also performed good accuracy after the ANN method with 76.93% accuracy in predicting only 161.

Previous studies have explored the prediction of students' final grades using their demographic attributes, including gender, student ID, class, year intake, and religion, as documented by Abdul Bujang et al. (2020), Karlos et al. (2020), Aman et al. (2019), and other researchers. One of the reasons why they used gender as one of the attributes is because students have different study or learning styles between males and females. Most female students exhibit a variety of positive learning styles and behaviors when compared

to male students (Shahiri et al., 2015). Other variables used to predict student's grades were continuous assessment marks such as tests and quizzes as stated in Altabrawee et al. (2019), Eman et al. (2016), and Khan et al. (2015) studies.

Table 2.2 summarizes the study on students' grade predictions using various ML techniques, data sizes, variables and the best method found with the highest accuracy percentage. According to Table 2.2, Altabrawee et al. (2019) used English course grade data as one of their independent variables to predict Computer Science course grades. The English subject was chosen as an attribute because of its relevance to computer science and most computer educational materials are learned and delivered in English (Altabrawee et al., 2019). However, there are other courses that seem more relevant to predicting computer science courses where those courses are basic courses in the field of computers such as basic computing as well as mathematics courses. Choosing a course in the same field is better because it provides the same level of knowledge for the students. For example, the level of knowledge in mathematics subjects is relevant in predicting student achievement in statistics, computer science, and accounting courses.

Table 2.2 also shows some studies that use a small dataset size, which is less than 500 data. Barnabas et al. (2018), Venkat et al. (2018), and Abu Zohair et al. (2019) have conducted a study using dataset sizes of 247, 197 and 50 samples respectively and the results of their study found that the SVM model is the best model for predicting students' performance. Ahmad N et al. (2021) conducted a predictive analysis research for student GPA using only a dataset of 59 first semester students from Universiti Teknologi Mara, Terengganu. The research successfully achieved 93.2% accuracy with the Artificial Neural Network model. Besides that, Razali M et al. (2022) used 141 datasets of Universiti Teknologi Sarawak students, to develop predictive models for students' grading using the Bayes Network, NB, Simple Logistic, JRip Rule Classifier, and RF. As a result, the JRip Classifier was the model that produced 92% model accuracy.

Table 2.2 Related Studies on Students' Performance Prediction using ML Techniques

Title / Author	Data size	Variable	Method	Evaluation Matrices	Best Method
A Predictive Analytics Model for Students Grade Prediction by Supervised ML (Abdul Bujang et al., 2021)	489	Student number, class, year intake, cohort, gender, CGPA, Continuous assessment mark, final mark, total mark, final grade	DT (J48), RF, SVM, LR	Accuracy, MAE, RMSE, RAE	J48
Predicting Students' Performance Using ML Techniques (Altabrawee et al., 2019)	161	Personal and life style, studying style, family related, educational environment satisfaction, 1st sem English course marks,Grade Point Average (GPA) and computer science outcome (Good, weak)	ANN, DT, LR, NB	Accuracy, precision, recall, F-measure, classification error, ROC index	ANN
The Design of Predictive Model for the Academic Performance of Students at University Based on ML (Barnabas et al., 2018)	247	Average matriculation results, self-study, lecturer's competency, attendance and first semester's GPA, performance classes (Good, average, poor)	Bayesian networks, SVM and DT	Accuracy, RMSE	SVM

Table 2.3 Continued

Title / Author	Data size	Variable	Method	Evaluation Matrices	Best Method
Grade Prediction Using Supervised ML Techniques (Eman et al., 2016)	2500	Student's information, test, quizzes and all assessments mark, class participation, grades	DT Classifier, k-NN, NB, Rule Induction	Accuracy, precision, recall	Rule Induction
Final Grade Prediction of Secondary School Student Using Decision Tree (Khan et al., 2015)	1500	Student performance, attendance, test and quiz, final mark, final grade	J48 DT algorithm	Accuracy	J48
Predicting Student Grades Using ML (Venkat N., 2018)	197	ID, year, attendance, gender, CGPA, continuous assessment and final marks, Grade	DT, NB, SVM, k-NN	Mean accuracy of 5-folds cross-validation	SVM
A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs (Xu et al., 2017)	1169	GPA, course information, Grades	Linear regression, LR, RF and k-NN	MSE	RF

Table 2.4 Continued

Title / Author	Data size	Variable	Method	Evaluation Matrices	Best Method
Prediction Of Student's Performance by Modelling Small Dataset Size. (Abu Zohair et al., 2019)	50	Student number, age, name, grade, course name, grade, Instructors name	Multilayer Perceptron Artificial Neural Network, NB, SVM, k-NN, Linear Discrimination Analysis (LDA	Accuracy, Cohen's kappa	SVM
A Decision Tree Approach for Predicting Student Grades in Research Project Using Weka (Abana et al., 2019)	133	Research Method grade, Research Project grade, gender, backlog, programming proficiency	Random Tree, RepTree and J48	Accuracy	Random Tree
Prediction of Students' Performance using ML (Dhilipan et al., 2021)	1173	previous semester marks, final grade	LR, DT, entropy and k-NN.	Accuracy, precision, recall, f1-score	LR

Table 2.5 Continued

Title / Author	Data size	Variable	Method	Evaluation Matrices	Best Method
Students' Performance Prediction using Artificial Neural Network (Ahmad N et al., 2021)	59	Interest in Electrical Engineering course, GPA, SPM results for Mathematics, Additional Mathematics, and Physics (binary class)	ANN	Confusion matrix ROC, cross entropy performance, and error	ANN
Predictive Model for Undergraduate Students Grading using Machine Learning for Learning Analysis (Razali M et al., 2022)	141	Demographic, study preparation, study behaviour and environment, student's motivation, and CGPA classification	Bayes Network, NB, Simple Logistic, JRip, RF, etc.	Accuracy	JRip

Based on observations from previous studies, some researchers use error measurements to predict classification output where it is not appropriate to use. Error measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are suitable for regression analysis. The achievement of a model is based on its accuracy value, because the higher the accuracy value, the better the model is in predicting the output. However, to add insight from the results of the analysis, the researcher can also make an AUC analysis using the One-vs-All setting, where the difference in prediction for each class can be determined.

Multinomial class prediction models must be evaluated using the One-vs-All AUC to determine how well one class is differentiated from the others, identify specific flaws in the model, and guarantee balanced and comprehensible performance. This method offers a thorough, detailed, and trustworthy evaluation, which makes it a vital tool for building and implementing durable machine learning models for complicated classification problems. Therefore, this study also measured the model's achievement from the AUC value using the One-vs-All technique to see which grade (according to performance level) can be predicted better by each model.

2.6 Summary

This chapter discusses predictive analytics research for academic performance. The initial part of this chapter begins with a general description of predictive analytics and relevant literature review on the application of predictive analytics using ML techniques. The last three subtopics were more focused on discussing the application of ML techniques to predict students' final grades and the variables or attributes determination as the factors that affected the final grade. As a summary of this chapter, from the literature review on related studies, there is no study conducted among Islamic universities in this field. Islamic university students have different backgrounds and characteristics from general university students. Their students are not science stream students, so the factors that affect their science stream subjects, such as statistics, may be different. Therefore, it is crucial to assist educators in Islamic universities to prevent failure in science stream subjects. This study will provide the best ML model to predict students' final grades and improve the failure rate before the final examination.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter outlines the machine learning (ML) methodology employed to develop predictive models for students' performance in the Business Statistics (BS) subject. As stated in the initial chapter, the primary aim of this research is to identify the optimal model capable of achieving higher accuracy in predicting students' performance. This study primarily focuses on supervised ML methods since it aims to predict students' performance through classification techniques. The dataset comprises student's demographic information such as gender, program and intake as well as student's achievement records (including CGPA before undertaking statistics courses, grades in Business Maths courses, carry marks, and BS performance levels) retrieved from the UnIPSAS Learning Management System (LMS), known as eCampus.

3.2 Machine Learning Process

The process of ML consists of developing self-learning algorithms and enabling the system to learn new things from the input. This study employed a few phases of the ML technique, as illustrated in the ML Pipeline (Figure 3.1). The data was collected from the LMS, and the variables were analyzed due to their perceived relevancy in the literature. The data analysis and preprocessing of the 450 student datasets began with an analysis of the variables' significance. Next, this research proposed to employ supervised models, which are multinomial logistic regression (MLR), support vector machine (SVM), Random Forest (RF), decision tree (DT), k-nearest neighbor (k-NN) and Naïve Bayes (NB) since the aim of this study is to obtain the best model to predict students' performance level.

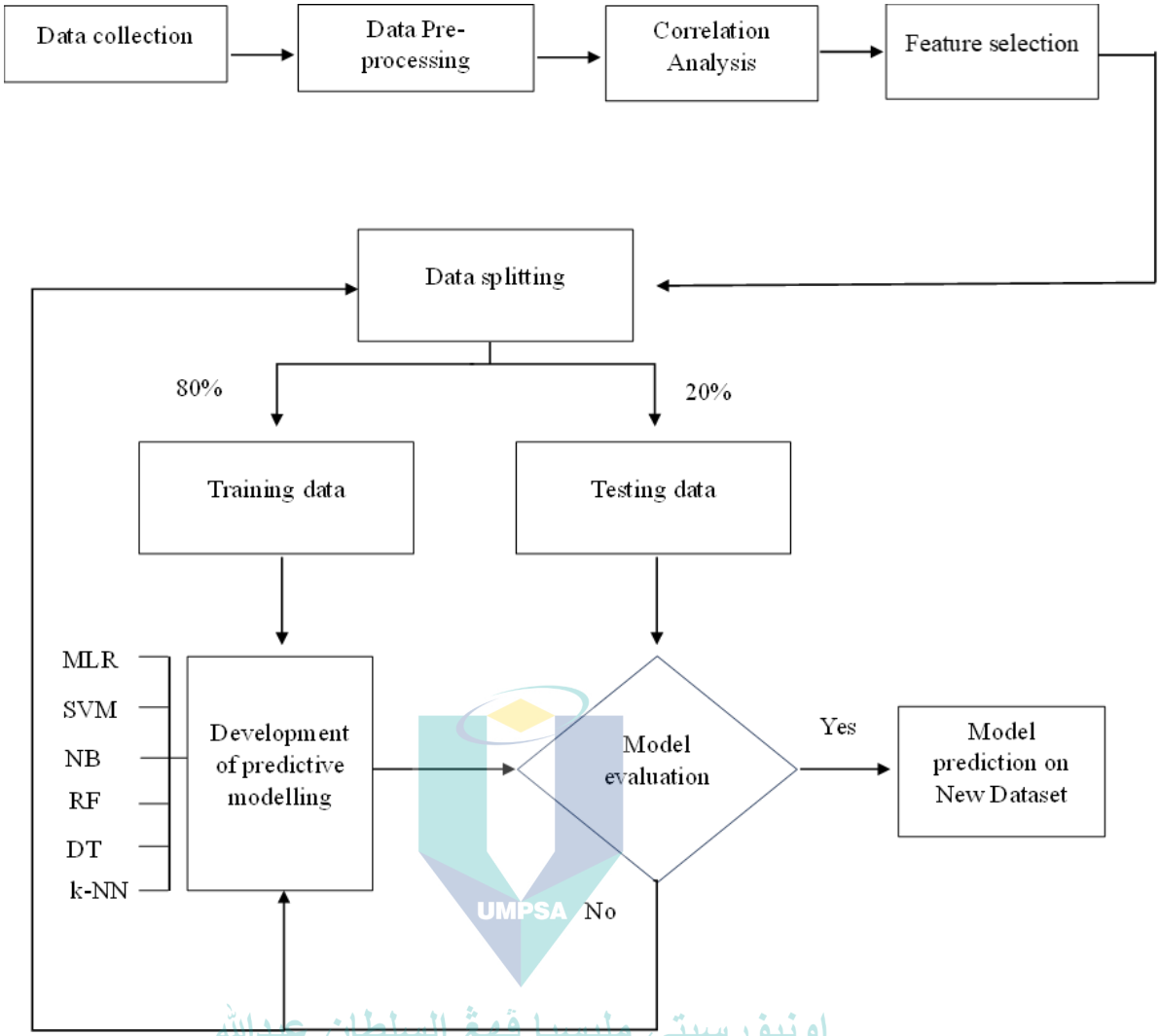


Figure 3.1 Machine Learning Pipeline

3.2.1 Data Collection

The ML techniques cannot be run without data. The data preparation phase involves two steps: data collection and preprocessing. To start with the process, we collected the required data and loaded it into the Python software. During the data collection phase, the entirety of the information was obtained from the Learning Management System (LMS), which includes students' personal details, course outcomes, and grade points. Initially, the dataset was compiled and stored as an Excel file, which was later imported into Python as a CSV file. The dataset utilized in the study encompasses real data from students enrolled in the Diploma programs in Business Studies, Accounting, Marketing Management, Finance and Banking at UnIPSAS from

June 2013 to June 2022. Specifically, the dataset focused on 450 students attended the BS course during the fourth semester. As outlined in Table 3.1, the attributes encompass gender, course, student intake, Cumulative Grade Point Average (CGPA) from the semester preceding the BS course, Business Mathematics (BM) grades, total carry marks, and students' performance levels based on their final grades in the BS final examination.

Table 3.1 List of Attributes

No	Attributes	Description	Details	Encode value
1.	Gender	Student's gender	Male Female	1 0
2.	Intake	Student's intake	First intake (June) Second intake (December)	1 2
3.	Program	Students' program	Diploma in Accounting Diploma in Business Studies Diploma in Finance and Banking Diploma in Marketing Management	0 1 2 3
4.	CGPA	Cumulative Grade Point Average (CGPA) for semester before taking BS subject.	1.00 – 4.00	-
5.	CM	Carry Marks for Business Statistics.	0 - 50	-
6.	BM_grade	Business Mathematics grade	A A- B+ B B- C+ C C- D+ D E	1 2 3 4 5 6 7 8 9 10 11

Table 3.2 Continued

No	Attributes	Description	Details	Encode value
7.	BS_Performance level (grades)	Student's Performance Level According to Grades in Business Statistics Final Examination	Excellent (A, A-) Very good (B+,B) Good (B-,C+) Pass (C,C-) Weak (D+,D) Fail (E)	1 2 3 4 5 6

3.2.2 Data preprocessing

Within machine learning (ML), data preprocessing encompassed the actions to refine raw data, rendering it suitable for developing and training ML models. Data preprocessing is a data mining technique employed in ML to transform real-world data into a structured and understandable format. Before preprocessing, it is imperative to conduct data analysis to identify any missing values. Addressing missing values is crucial as they can potentially impact the outcomes of ML models, thereby reducing their accuracy. However, in this study, missing values were not encountered as the dataset was extracted from the LMS, where student information was systematically recorded. Data preprocessing in this research used Python and involved procedures such as data cleaning, encoding, and handling outliers.

After the data was successfully loaded, it needed to be transformed. For instance, CGPA had to be within a reasonable range, which was between 1.00 and 4.00, as well as the carry marks (CM) for the BS course, where the marks should have lie between 0 and 50 marks. Most ML techniques, which have been shown to be effective in dealing with limited dataset sizes, require numeric variables (Abu Zohair et al., 2019). In this study, the encoding process needed to be done on the string variables, which were the student's gender, student's programs, BM's grades, and students' performance levels according to BS's final grades, as illustrated in Table 3.1. In Python, the encoding function used was LabelEncoder().

Outliers were data points that were significantly distant from other data points and also known as extreme values that could lead to skewness in data distribution. They could adversely affect the performance of a developed model by impacting its accuracy and

causing inaccurate predictions. Outliers were typically identified using boxplot diagrams. In this study, outliers were treated using the Winsorize method, which involved setting minimum and maximum limits. This process began with determining the interquartile range (IQR), defined as the difference between the first quartile (Q1) and the third quartile (Q3). The new minimum limit was then calculated by subtracting 1.5 times the IQR from the Q1 value, while the new maximum limit was determined by adding 1.5 times the IQR to the Q3 value. Subsequently, a new variable with updated minimum and maximum limits was created and used in subsequent analyses.

3.2.3 Feature Selection

In ML modelling, feature selection is a process of selecting the relevant attributes from the original list of attributes to be used in developing the predictive models. This process aims to increase the performance of ML models. Typically, for supervised techniques, the relevance of attributes is evaluated by their correlation values with the target variable, which can be either categorical or numerical (Li J. et al., 2017).

In machine learning, correlation analysis is a crucial step aimed at determining the relationship between independent attributes and the output. The correlation coefficient refers to the results of correlation analysis in which the values are in the range between $-1.0 to +1.0$. According to Taylor (1990), the closer the correlation value to ± 1.0 , the stronger the correlation between the two attributes. In this study, the variables with excellent or high correlation values (± 0.7 to ± 1.0) were identified as the factors that affect students' performance levels.

The method of identifying the correlation coefficient is based on the types of attributes used to develop the ML models. As stated in Table 3.1, the independent attributes (predictors) for this study are categorical and numerical types, while the target attribute is categorical (ordinal). Table 3.2 shows the different types of attributes used in this study and the appropriate correlation coefficient required to conduct correlation analysis between the predictors and the target attribute.

Table 3.3 Types of Attributes and Correlation Coefficients

Independent Attributes (Predictor)	**Types**	**Target Attribute**	**Types**	**Correlation Coefficient**
Gender	Categorical			Cramer's V
Intake	Categorical	Student's	Categorical	Cramer's V
Students' program	Categorical	Performance		Cramer's V
CGPA	Numerical	Level		Spearman's Rank
CM	Numerical			Spearman's Rank
BM's grades	Categorical			Cramer's V

Cramer's V Correlation

The association between two categorical attributes, like nominal or ordinal, could be obtained using the Pearson Chi-Square test. However, a significant p-value ($p < 0.05$) obtained from the Chi-square test could not determine the association strength. Cramer's V correlation is more appropriate to determine the correlation strength between two categorical attributes. Cramer's V correlation is very useful in checking the correlation strength when the two categorical attributes are significant, as determined by the Chi-square test. Therefore, the Chi-square value is also used to obtain the Cramer's V correlation and the equation is given by:

$$\sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \quad 3.1$$

Where φ is also known as phi coefficient, the value of χ^2 is obtained by the Chi-square test, n is the total number of observations, and the value of k and r represent the number of column and rows for the contingency table. Table 3.3 provides the strength interpretation suggested by Sapra, R et al., (2021).

Table 3.4 The Strength Interpretation of Cramer's V Correlation

Effect Size	**Interpretation**
0 and under 0.1	Negligible Association
0.1 and under 0.2	Weak Association
0.2 and under 0.4	Moderate Association
0.4 and under 0.6	Relatively Strong Association
0.6 and under 0.8	Strong Association
0.8 to 1.0	Very Strong Association

Spearman's Rank

Spearman's rank correlation coefficient (SRCC), denoted by ρ (rho), is known as a non-parametric measure used to determine the strength and direction of monotonic association between two attributes. SRCC is suitable to use for evaluating the relationship between attributes when one or both are ordinal scales. According to Sapra R. et al. (2021), SRCC is an appropriate and commonly used correlation coefficient for ordinal and numerical data. These attributes are ranked in preference order, with ranks assigned based on the quantitative order of continuous values from the dataset. If the assigned ranks are different integers, ρ is computed using the following equation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad 3.2$$

Where d_i is the difference between the two ranks and n represent the sum of all observations. In Python, the Spearmanr from Scipy module was imported to obtain the SRCC value. Yan et al., (2019) have provided the grading table to interpret the ρ values as follows:

Table 3.5 Grading Table of Spearman's Correlation Coefficient

Grading Standards	**Correlation Degree**
$\boldsymbol{\rho = 0}$	no correlation
$\boldsymbol{0 < \lvert\rho\rvert \leq 0.19}$	very weak
$\boldsymbol{0.20 \leq \lvert\rho\rvert \leq 0.39}$	weak
$\boldsymbol{0.40 \leq \lvert\rho\rvert \leq 0.59}$	moderate
$\boldsymbol{0.60 \leq \lvert\rho\rvert \leq 0.79}$	strong
$\boldsymbol{0.80 \leq \lvert\rho\rvert \leq 1.00}$	very strong

3.2.4 Data splitting

The main goal of developing an ML predictive model is to obtain the best accuracy of prediction on new datasets, as well as unseen datasets. The foundation of all validation strategies is data splitting, where the dataset is divided into training and testing sets. The model is trained using the training dataset, and the evaluation part is done on the test dataset. This study used the Train Test Split method, which is known for its ease and effectiveness in ML model development. The training data comprised the initial 80% of the dataset, while the remaining 20% was allocated for testing purposes. However, the study also compared other divisions, such as 70% of training data and 30% of testing data, in order to find the most relevant division to obtain the best performance of the model. The figure illustrates the partitioning procedure using the Train Test Split method.

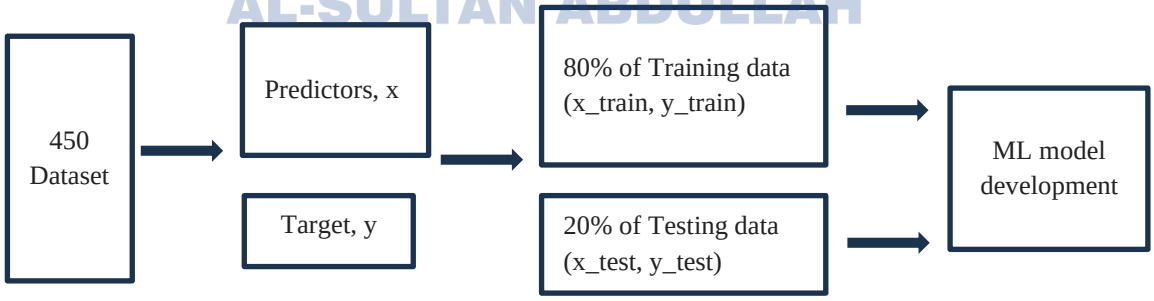


Figure 3.2 Train Test Split Procedure

3.2.5 Model Development

This research focused on implementing supervised ML techniques in predicting students' performance level. Supervised learning algorithms are taught by looking at examples that had been labelled, like a set of inputs (i.e., students' carry marks, CGPA, and pre-requisite subject's grade) for which the desired output is known. The learning algorithm received a set of inputs along with the proper outputs, and it learned by comparing its actual outcome with the correct outputs to identify the model's performance. The model is then modified appropriately until the desired performance is achieved.

Classification is a process of finding a method that helps divide a dataset into classes depending on a variety of factors. A classification problem has a discrete value as its output. The objective of the classification method is to identify the mapping function that corresponded to the input to the discrete output. Types of algorithms in classification model development included decision tree (DT), Random Forest (RF), Naïve Bayes (NB), support vector machine (SVM), multinomial logistic regression (MLR), and k-nearest neighbor (k-NN).

Decision Tree (DT)

A DT is a diagram similar to a flowchart that depicts the various consequences of a set of choices. It can be represented as a tree consisting of nodes from root to leaf, with inspections on characteristics set in internal nodes and class variables shown in leaf nodes (Eman et al., 2016). In order to predict the outcome, DT makes sequential, hierarchical decisions about the outcome variable based on the predictor data (Altabrawee et al., 2019). Changing parameters such as the quality measure, splitting criteria, minimum number of records per node, and pruning procedure may enhance the accuracy of a decision tree model's predictions. In Python, the DecisionTreeClassifier was used to develop the DT model. The hyperparameters that were tuned were the Gini index and entropy. Gini Index and entropy are calculated as:

$$\text{Gini Index: } I_G(t) = 1 - \sum_{i=1}^{K} P_i{}^2 \qquad 3.3$$

$$\text{Entropy: } H(t) = -\sum_{i=1}^{K} P_i \log_2(P_i) \qquad 3.4$$

where P_i the probability of ith class in node, t and K is number of classes.

Random Forest (RF)

The RF algorithm is an extension of the decision tree method in which a large number of trees are created for the model instead of a single tree. This 'forest' of decision trees is used to analyze the data, and the ensemble then votes for the most probable result. Each decision tree categorizes each participant, and the classification that occurs most often in the forest is chosen (Zabriskie et al., 2019). RF is a bagging approach that creates an ensemble of trees by generating multiple training sets with replacement. The dataset is divided into N samples using randomized sampling. A model is built on all the samples, and then, using a single learning algorithm, the relevant predictions are combined using parallel voting. In summary, steps include in RF modelling were:

1. Take *n* number of random records from the data set having *k* number of records.
2. Individual decision trees are built for each sample.
3. Each decision tree will produce an outcome.
4. Final outcome is based on majority voting for classification.

In an RF model, the number of decision trees to be built in the ensemble is specified by the n_estimators hyperparameter. By combining predictions from more giant trees, increasing n_estimators improves the model's resilience and accuracy by lowering variance and enhancing generalization. It is expected to begin fine-tuning n_estimators with fewer trees, say, 100 or 200, and progressively increase them while tracking performance improvements until a point at which the benefits of adding more trees become less. The ideal value for n_estimators can be found using automated techniques such as Grid Search with cross-validation. This way, the model's performance can be maximized without needlessly lengthening the training period.

Naïve Bayes (NB)

The NB model was among the most popular supervised ML methods. The NB classifier is a straightforward probabilistic classifier based on Bayes' theorem. NB's effectiveness stems from the assumption of attribute independence, which may only hold in some real-world datasets (Aileen et al., 2023). Several approaches have been taken to address this assumption, with attribute selection being one of the most important. However, conventional methods of attribute selection in NB have significant computational costs (Aileen et al., 2023). The classifier used to develop the NB model in Python was Gaussian Naive Bayes (GaussianNB()). It is a classification method in ML techniques relying on the probability approach as well as Gaussian distribution. This classifier considers that all attributes are able to contribute independently to predict the outcome. The formula for NB classifier is expresses by:

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(y) \times P(x_1 \mid y) \times P(x_2 \mid y) \times \ldots \times P(x_n \mid y)}{P(x_1) \times P(x_2) \times \ldots \times P(x_n)} \quad 3.5$$

where:

$P(y \mid x_1, x_2, \ldots, x_n)$ is the rear probability of class y with $x_1, x_2, \ldots, x_n$ as attributes.

$P(y)$ is the previous probability of class y.

$P(x_i \mid y)$ is the probability of attribute x_i with class y.

$P(x_i)$ is the previous probability of attribute x_i.

Support Vector Machine (SVM)

The SVM is trained using the cost function. To ensure the accuracy of SVM, the value of theta (θ) have to be minimized. In the equation below, the functions cost_1 and cost_0 refer to the cost for an example where $y = 1$ and $y = 0$ respectively. The cost function determined by kernel (similarity) functions.

$$\theta = C \sum_{i=1}^{m} \left[y^{(i)} \text{cos}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)} \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \quad 3.6$$

The kernel function is essential in the SVM algorithm as it is critical in determining the decision limit and converting the input into a higher-dimensional space where the data might be more distinguishable. Three standard kernel functions in SVM are the linear kernel, polynomial kernel, and Radial Basis Function (RBF) kernel. The linear kernel is a default kernel in many SVM executions and is suitable for data that could be separated linearly. The polynomial kernel is another kernel that allowed the SVM model to solve nonlinear associations between attributes and reflect on interactions of high-level attributes. Additionally, the RBF kernel is a popular option for SVM, and it captures the nonlinear association of data. It is essential to tune different kernels to obtain the best performance values.

Besides the kernel function, other parameters that should have been considered in SVM's hyperparameter tuning are regularization (C) and Gamma. C controls the trade-off between enhancing the margin and reducing the error. In SVM, the margin is defined by the nearest support vectors on one of the hyperplanes. The C parameter helps to implement soft-margin SVM, where small C values increase the margin, while large C values indicate a hard margin for SVM.

Gamma (γ) determines the effect of individual training data on the decision boundaries. A small value of gamma indicates a more significant radius effect for each support vector, resulting in a smoother decision boundary. The parameters in SVM are commonly tuned using the Grid Search technique. This technique involves assessing the model's performance using different values of gamma and choosing the value that obtained the best performance.

Multinomial Logistic Regression (MLR)

Logistic regression is a mathematical modeling approach that explaines the correlation between several predictor factors X_1 *to* X_k a and an outcome variable, D. MLR is one of the classification models applicable to predict more than two classes of outputs. In MLR, the probability for each class is modeled as an independent variable function and the Softmax function is utilized for probability calculation. Number of classes is denoted by *K*, where the value of *K* is more than 2 and the probability is $P(y = K \mid x)$, where x the independent variables. The equation of softmax function for k number of classes is given by:

$$P(y = K \mid x) = \frac{e^{z_K}}{\sum_{i=1}^{K} e^{z_i}} \quad 3.7$$

where the logit (raw score) for K is denoted as z_K , e is known as the base of Euler's number, and the divisor is the summation of e^{z_i} where i is equals to 1 until K.

Next, hyperparameter tuning in MLR involved the hyperparameter solver, such as Limited memory Broyden Fletcher Goldfarb Shanno (lbfgs), Libliner, Stochastic Average Gradient (SAG), Newton-cg, and so on, for multinomial classification. The Solver is the optimization algorithm utilized by the MLR algorithm to determine the best coefficient to predict the likelihood of each class

k-Nearest Neighbor (k-NN)

The k-NN technique is a fundamental and straightforward supervised machine learning approach that can handle classification and regression problems. The k-NN classifiers identified a data item as belonging to the training dataset class to which it is geometrically closest (Anderson et al., 2017). Additionally, the k-NN algorithm assumes that similar things are near each other. For classification cases, k-NN operated by calculating the distance between an inquiry and each example in the data, picking the k closest to the query, and then voting for the most frequent label. For the distance, this study used the default setting, Euclidean distance. To choose the right *k* for the data, the k-NN algorithm needed to be executed several times with various *k* values. The final value of *k* was chosen based on the minimum number of errors while retaining the system's capacity to generate correct predictions when testing with new data. The primary issue with the k-NN method was that excessive or irrelevant information could severely impact its accuracy. Similarly, its accuracy was reduced if the variable weights were inconsistent with their significance (Ofori et al., 2020).

3.2.6 Evaluation of Model Performance

The model's performance was determined by four performance metrics: accuracy, precision, recall, and F1-score. All of these values can be obtained in Python through the classification reports. The classification report is commonly used for ML tasks, especially for classification. The values of these four metrics were obtained by the values of True

Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) through the confusion matrix.

Since this research considered multi-class classification (6 levels of performance according to 11 possible grades for the BS subject), the TP, TN, FP, and FN needed to be measured for each class. The value of TP referred to the frequency of predictions where the classifier correctly predicted the positive class to be positive. In contrast, the value of TN referred to the frequency of predictions where the classifier or model correctly predicted the negative class as negative. On the other hand, FP referred to the frequency of predictions made by the model, which incorrectly predicted the negative class as positive. Meanwhile, FN is the number of predictions in which the model incorrectly predicted the positive class as negative.

The accuracy of a system could be defined as the degree of similarity between the expected and actual values of a quantity (Eman et al., 2016). The higher the accuracy value, the better the performance of a model. It is most prevalent when all classes are of equal importance. The following equation can measure the accuracy:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{(TP+FP+TN+FN)}} \qquad 3.8$$

The precision value is the proportion of predictions as a positive class are actually positive. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{(TP+FP)}} \qquad 3.9$$

The recall value indicated what proportion of all positive samples the classifier accurately identified as positive. It is also referred to as the TP rate, the sensitivity, and the probability of detection. To calculate recall, the following equation should be used:

$$\text{Recall} = \frac{\text{TP}}{\text{(TP+FN)}} \qquad 3.10$$

F1-score gives a method for combining accuracy and recall into a single metric that includes both characteristics. Traditionally, F1-score is calculated as follows:

$$\text{F1-score} = \frac{(2 \times \text{Precision} \cdot \text{Recall})}{(\text{Precision}+\text{Recall})} \quad 3.11$$

Besides that, another evaluation method that could be used to validate model performance is the value of the area under the receiver operating characteristic (ROC) curve (AUC). The One-vs-All setting was employed to achive the AUC for multi-class models. This technique was used to create an overall AUC by classifying all classes as positive. This method trained a group of independent classifiers, with one class being positive and the others being negative. Each class's AUC was averaged to provide a final area under the ROC curve for each model.

The values of accuracy, precision, recall, and F1-score, as well as the AUC, could be interpreted as follows: a score less than 0.5 indicated poor value, between 0.5 to 0.7 indicated moderate to good performance, and more than 0.8 indicated the best. In contrast, 1.0 indicated the perfect performance value.

3.2.7 Model Prediction on New Dataset

Data testing is vital to evaluate the performance and generalization abilities of the best model. The initial step was to ensure the new dataset was correctly prepared and pre-processed using the same procedures as the training datasets. It includes encoding categorical variables and handling outliers. The complete data of 22 students who took the BS course in the most recent semester was analyzed, and predictions were formed based on their level of performance in the final exam results using the most accurate predictive model obtained in model development.

3.3 Python Libraries

Python has emerged as a leading programming language in scientific computing, gaining popularity in academic settings and the corporate landscape. Using this platform, Scikit-learn provides cutting-edge implementations of numerous well-established ML techniques while maintaining a user-friendly interface seamlessly integrated with the Python programming language. It satisfied the rising need for statistical data analysis by

non-specialists in software and online businesses and non-computer science subjects such as biology and physics (Pedregosa et al., 2011). Python libraries enabled users to access, analyze, and alter data for machine learning, which required regular data processing (Kumar, 2021).

These are some of the most comprehensive libraries accessible for ML methods, according to Kumar (2021).

1. Scikit-learn can manage fundamental machine learning methods such as clustering, logistic and linear regression, regression, and classification.
2. Pandas are used for sophisticated data and structural analysis. It enables the merging and filtering of data, as well as the collection of data from other sources (such as Excel).

As shown in Figure 3.1 in this study, all processes involved in the proposed ML pipeline were conducted using Python software. It started with completing data analysis processes such as selecting related attributes, encoding, and partitioning. The development of predictive models involving ML models, the measurement of the performance of proposed classifiers, and ending with the testing of the best model selection were all carried out.

3.4 Summary

This chapter elaborated on the methodology used in this study based on the ML pipeline. ML techniques and Python coding were utilised to achieve all the objectives in this study, starting from importing datasets, data preparation, variable selection, model development and evaluation, and model testing. To obtain the best model, the hyperparameters for each model were tuned before the best model was selected to predict new data. The proposed models to be developed and evaluated were DT, MLR, k-NN, RF, NB, and SVM. The results of all analyses will be discussed in the next chapter.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter discusses each result obtained from the model development process that has been carried out. It begins with a discussion of the findings from the process involved in the data preparation by providing overall details about the dataset used in the study and the data preprocessing steps used. Then, the correlation analysis results have also been obtained to identify the variables or factors that were most related to students' performance in the final exam. Next, the data partitioning process for training and testing data is also discussed, as well as the finding of model performance, in which the best model is identified. Finally, the chapter discusses the results when the new dataset is applied to the best model.

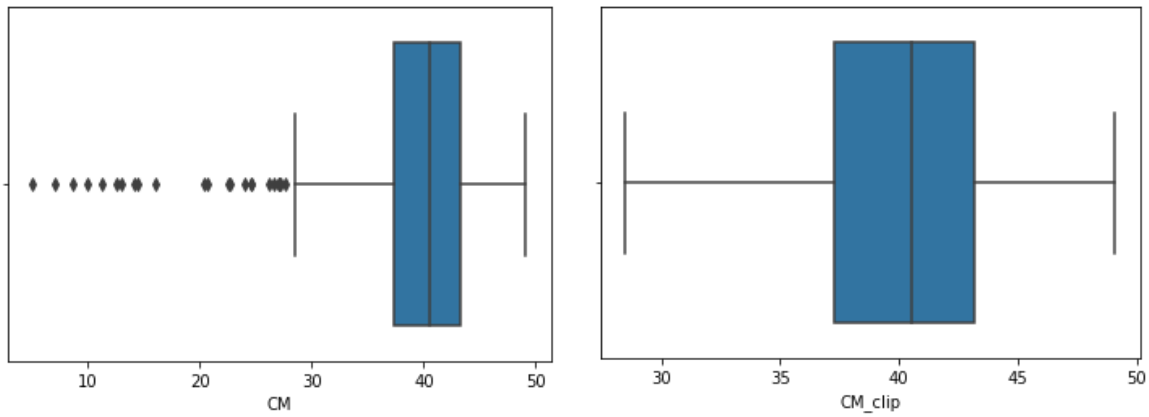
4.2 Data Preprocessing

The data of 450 students was successfully retrieved from the LMS and selected as a raw dataset. As stated in Table 3.1, six variables were involved: gender, student programmes, student CGPA before taking the business statistics subject, total carry marks (CM), grade for the prerequisite subject (BM), and the performance level according to the final grades in business statistics (BS).

As stated in Chapter 3, this dataset had no missing values or duplicate data. Therefore, the first step in data preprocessing was encoding, where all non-numeric variables were encoded into numerical data. The LabelEncoder() function in Python was used to encode data related to gender, course, and performance level.

To mitigate extreme values in the data distribution, identified outliers were handled using the Winsorize method, and new minimum and maximum limits for the data were set. Outlier values were identified from boxplot diagrams. Figure 4.1 shows the

boxplot for the CM with the outliers, and the boxplot after the outliers were treated where the attribute CM was replaced by the new attribute, CM_clip.



Before After

Figure 4.1 The boxplot before and after treating the outliers

4.3 Feature Selection

The feature selection process aims to identify all the relevant attributes for predicting students' performance and exclude unrelated attributes for ML modelling. This study employed correlation coefficient analysis to finalize the list of attributes used to develop the model. Besides measuring the strength of the relationship between all predictors (independent variables) towards the target variable, correlation tests were conducted among independent variables to identify multicollinearity. Multicollinearity is a situation where independent variables have a very strong relationship among them. There are various adverse effects of multicollinearity. Some of the consequences that may occur are the reduction of ML model performance, and multicollinearity can also lead to overfitting of the model.

As tabulated in Table 3.2, two correlation analysis methods were used, namely Cramer's V Correlation Coefficient to measure the relationship between the two categorical variables, and Spearman's Rank Correlation Coefficient (SRCC) to measure the relationship between numerical and categorical variables. It should also be noted that when identifying multicollinearity, there were two numerical variables also have been considered, namely CGPA and carrymarks. Therefore, to measure the relationship

between these two numerically scaled independent variables, Pearson's Correlation Coefficient method was used. The overall results of the correlation analysis have been presented in Table 4.1 and Table 4.2.

Table 4.1 Correlation Coefficient Between Predictors and Target Variable

Predictors	**Method**	**Results**
gender	Cramer's V	0.249
course	Cramer's V	0.216
intake	Cramer's V	0.0998
math_grade	Cramer's V	0.239
CGPA	SRCC	−0.543
carrymarks	SRCC	−0.74

Table 4.2 Correlation Coefficient Among Predictors

Predictors	**gender**	**course**	**intake**	**math_grade**	**CGPA**	**carrymarks**
gender	-	0.123	0.079	0.21	−0.20	−0.21
course	-	-	0.04	0.23	−0.23	−0.17
intake	-	-	-	0.149	0.029	0.48
math_grade	-	-	-	-	−0.59	−0.35
CGPA	-	-	-	-	-	0.37
carrymarks	-	-	-	-	-	-

Cramer's V results, as shown in Figure 4.1, show that there are three variables moderately associated with students' performance levels, which are gender, course and math_grade. According to SRCC results, carry marks have a strong negative correlation, showing that higher carry marks will give a lower code for performance level, which represents the higher level as explained in the previous chapter. At the same time, CGPA shows a moderate negative correlation with students' performance levels. In addition, one independent variable has no correlation towards students' performance level, which is students' intake. Therefore, students' intake was excluded from the list of attributes to predict students' performance levels in the model process.

Based on Table 4.2, all independent variables obtained weak to moderate relationships with each other. It is shown that no multicollinearity occurred among independent attributes.

4.4 Model Development and Performance Evaluation

In this phase, ML models were developed using six algorithms: MLR, DT RF, SVM, NB, and k- NN. The results of this phase are significant to achieve the second and third objectives of the study, which are to determine the most accurate predictive model for students' performance level and to evaluate the performance for the proposed models using the value of precision, recall, accuracy, F1-score, and area under receiver operating characteristics curve (AUC). Since these models were trained to predict six classes (six performance levels), the AUC performance was determined by calculating the average AUC value of all classes. As a reference, Hameed et al. (2022) used the average AUC score for multiclass classification of breast cancer histopathology images to determine the AUC's performance. Using the One-vs-All setting, this method trains a group of independent classifiers, with one class being positive and the others being negative. The curves' areas are averaged after calculating the AUC for each classifier to provide a final area under the ROC curve.

The first ML model is the MLR. The model achieved a moderate score for both precision and recall, which are 0.50 and 0.48, respectively, with a balanced F1-score of 0.49. The accuracy score 0.56 suggested that the model's ability to predict correct performance levels is moderate. Meanwhile, the AUC average score of 0.865 indicated that the MLR can discriminate between different performance levels. For the hyperparameter, the chosen solver was Conjugate Gradient with Newton's method (cg-newton). Figure 4.1 shows the classification report for the MLR model and the ROC curve with the AUC for each class.

	precision	recall	f1-score	support
1	0.85	0.79	0.81	42
2	0.38	0.45	0.41	20
3	0.23	0.30	0.26	10
4	0.22	0.20	0.21	10
5	0.33	0.17	0.22	6
6	1.00	1.00	1.00	2
accuracy			0.56	90
macro avg	0.50	0.48	0.49	90
weighted avg	0.57	0.56	0.56	90

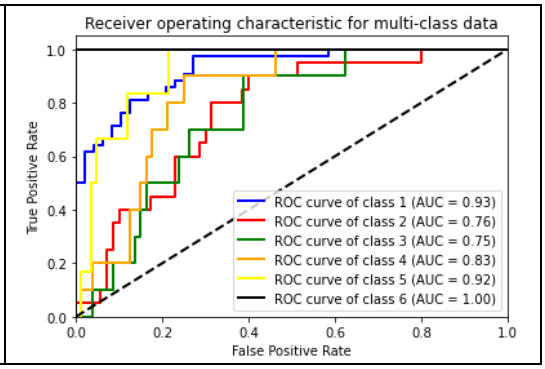


Figure 4.2 Classification Report and The ROC Curve for MLR Model.

Next, the DT model achieves the highest accuracy score compared to other models, which is 0.60. It indicates that the model correctly predicted the performance level for 60% of students from the testing dataset. The precision and recall obtained the exact value of 0.52, and F1-score is 0.52. As shown in Figure 4.3, the maximum depth graph shows that the entropy performed much better than the Gini index to achieve high accuracy for the DT model. However, DT has the lowest average of 0.66 for the AUC score. It means that DT can moderately distinguish the classes of students' performance. Figure 4.4 shows the DT model's classification report and the ROC curve.

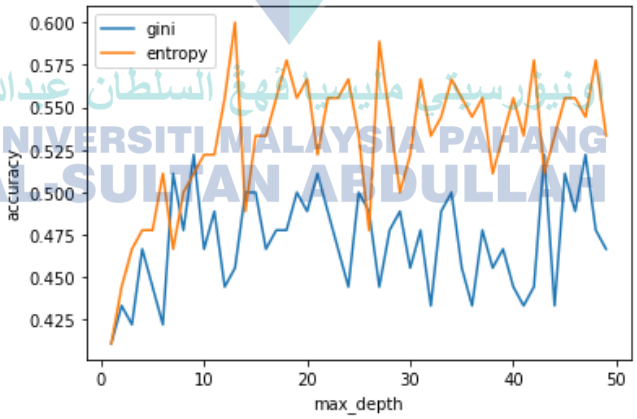


Figure 4.3 Maximum Depth for DT Model

	precision	recall	f1-score	support
1	0.72	0.82	0.77	34
2	0.52	0.55	0.53	22
3	0.50	0.41	0.45	17
4	0.57	0.40	0.47	10
5	0.33	0.25	0.29	4
6	0.50	0.67	0.57	3
accuracy			0.60	90
macro avg	0.52	0.52	0.51	90
weighted avg	0.59	0.60	0.59	90

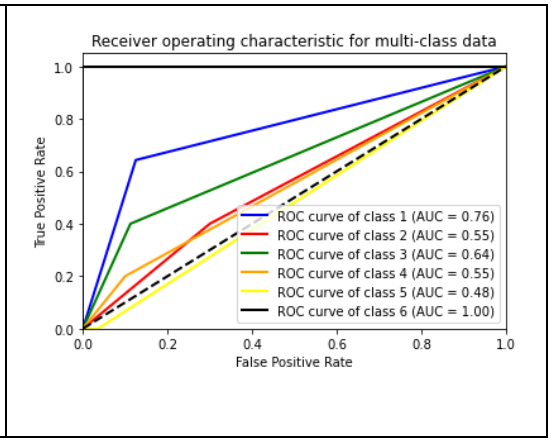


Figure 4.4 Classification Report and The ROC Curve for DT Model.

The SVM model scores a prediction accuracy of 0.52 while the precision is 0.51. However, the recall and F1-score metrics only achieve 0.45 and 0.39, respectively, as illustrated in Figure 4.5. From the ROC curve, the average AUC obtained from the area scores for each class is 0.83. All results for the SVM model indicate that the accuracy is 52%, and the model is very good at distinguishing six classes of students' performance levels. From GridSearchCV, the most appropriate kernel for the hyperparameter was the Radial Basis Function (RBF); the selected regularization value was 1000, and the gamma was 0.001.

	precision	recall	f1-score	support
1	0.78	0.74	0.76	39
2	0.32	0.57	0.41	21
3	0.33	0.08	0.13	12
4	0.25	0.10	0.14	10
5	1.00	0.20	0.33	5
6	0.38	1.00	0.55	3
accuracy			0.52	90
macro avg	0.51	0.45	0.39	90
weighted avg	0.56	0.52	0.50	90

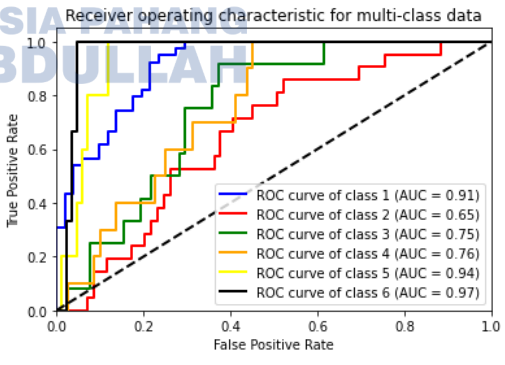


Figure 4.5 Classification Report and The ROC Curve for SVM Model.

Figure 4.6 represents the classification report and the ROC curve for the RF model. The model achieves a moderate score for accuracy which is 0.56. The precision score obtained the highest value among other evaluation metrics for RF, which is 0.63. Besides that, recall and F1-score obtained 0.48 and 0.47 scores, while the average AUC score is 0.863. The number of trees, also known as the n_estimator, used to develop the model with the most appropriate performance was 250.

	precision	recall	f1-score	support
1	0.89	0.74	0.81	42
2	0.39	0.55	0.46	20
3	0.21	0.30	0.25	10
4	0.27	0.30	0.29	10
5	1.00	0.00	0.00	6
6	1.00	1.00	1.00	2
accuracy			0.56	90
macro avg	0.63	0.48	0.47	90
weighted avg	0.64	0.56	0.56	90

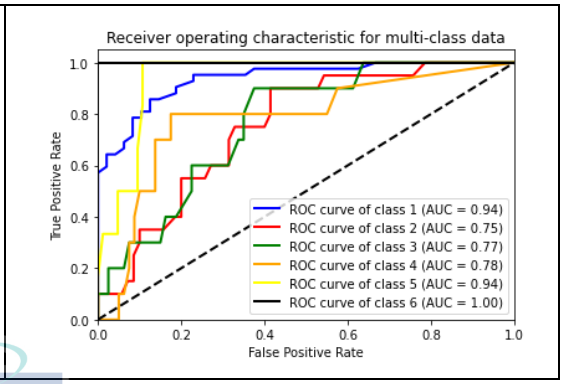


Figure 4.6 Classification Report and The ROC Curve for RF Model.

The evaluation metrics for the NB model are presented in Figure 4.7. The classification report obtained the precision, recall, and F1-scores at 0.39, 0.46, and 0.40, respectively. Meanwhile, the accuracy value achieves 0.51, meaning the model has a 51% accuracy rate in predicting students' performance levels correctly. In the ROC curve, the calculated average AUC is 0.833.

	precision	recall	f1-score	support
1	0.89	0.74	0.81	42
2	0.33	0.40	0.36	20
3	0.13	0.20	0.16	10
4	0.17	0.10	0.12	10
5	0.40	0.33	0.36	6
6	0.40	1.00	0.57	2
accuracy			0.51	90
macro avg	0.39	0.46	0.40	90
weighted avg	0.56	0.51	0.53	90

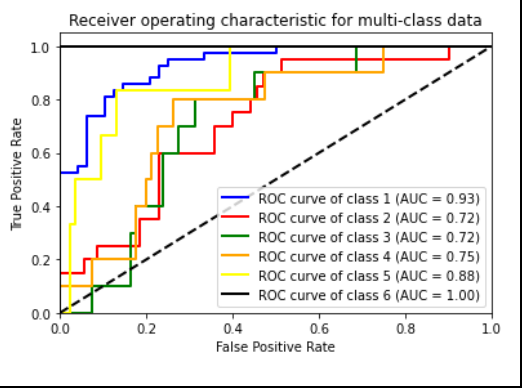


Figure 4.7 Classification Report and The ROC Curve for NB Model.

The last model that has been developed was the k-NN. Like other models, the k-NN also used the same metrics in measuring the model's performance. The precision value is 0.48, while the values of recall and F1-score are 0.36. Model accuracy for k-NN is 0.54. Based on the graph of error rates *k* values, the selected number of neighbours was 3, as illustrated in Figure 4.8, and the distance used was Euclidean. In addition, the values of AUC for each class were averaged to 0.763. Figure 4.9 shows the classification report and the ROC curve for the k-NN model.

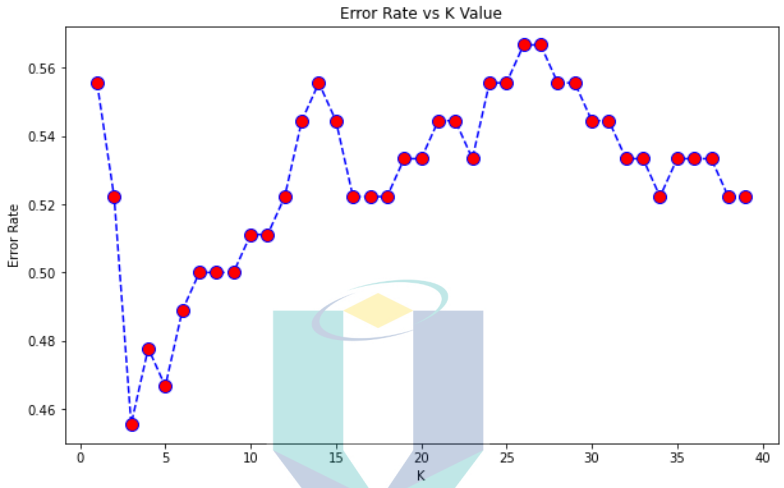


Figure 4.8 The Error Rate VS k Value Graph.

	precision	recall	f1-score	support
1	0.76	0.84	0.80	38
2	0.31	0.53	0.39	17
3	0.33	0.36	0.35	11
4	1.00	0.15	0.27	13
5	0.50	0.29	0.36	7
6	0.00	0.00	0.00	4
accuracy			0.54	90
macro avg	0.48	0.36	0.36	90
weighted avg	0.60	0.54	0.52	90

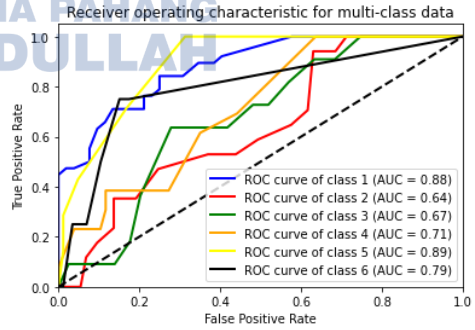


Figure 4.9 Classification Report and The ROC Curve for k-NN Model.

4.4.1 Comparison of Model Performance

Figure 4.10 shows the difference in performance between each model, and it is found that the average area under the ROC curve give better results compared to other performance values.

	Precision	Recall	F1-score	Accuracy	Average AUC
MLR	0.5	0.48	0.49	0.56	0.865
DT	0.52	0.52	0.51	0.6	0.66
RF	0.63	0.48	0.47	0.56	0.863
NB	0.39	0.46	0.4	0.51	0.833
SVM	0.51	0.45	0.39	0.52	0.83
kNN	0.48	0.36	0.36	0.54	0.763

Figure 4.10 Comparison of all classification models' performances

The above figure shows that the RF is the best model in precision performance and the average AUC. Meanwhile, the DT has the highest value for recall, F1-score and accuracy compared to other model. Therefore, DT is considered the best model to predict students' performance level for BS subjects because of the high accuracy compared to other models. In addition, DT has also achieved consistent scores, which are considered moderate to good models according to other performance metrics.

On the other hand, by comparing the AUC scores for each class, insights into the relative predictive performance of the model for different classes could be gained. Based on the ROC curves, the AUC values for all six classes are obtained according to each model. For the "excellent" level, the RF model has the highest AUC, which is 0.94, followed by MLR and NB, which is 0.93. It shows that the RF is the best model in distinguishing instances of excellent students who get grades A- and A. Besides that, the MLR has the highest AUC to distinguish the "very good" level of performance, which is

0.76. The lowest AUC score for the "very good" level is obtained by the DT model, which is 0.55. It indicates that the DT model distinguishes "very good" performance levels poorly.

The "good" performance level represents the B- and C+ grades. According to the ROC, the RF model is considered the best model to differentiate instances of "good" performance levels of students. The RF achieves the highest AUC value, 0.77, while the DT model obtains the lowest AUC. In distinguishing instances of students with "pass" levels, grades C and C-, the MLR was the best model with an AUC value of 0.83. The RF, NB, and SVM also obtained good AUC values of 0.78, 0.75, and 0.76, respectively.

Finally, four models, namely MLR, DT, RF, and NB, have perfect scores (1.00) of AUC to distinguish class 6 (fail). It shows that the models are excellent in distinguishing instances of failing students. The summary of AUC values for each level of student's performance is clearly illustrated in Figure 4.11.

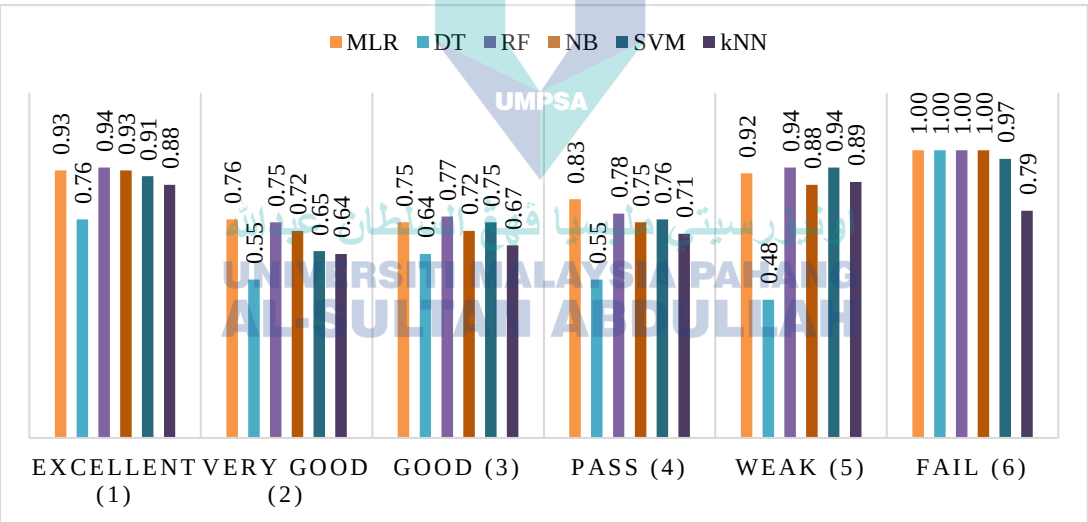


Figure 4.11 The AUC of ROC Curve for Each Performance Level

4.5 Model Performance on The New Dataset

Table 4.3 shows a list of actual and predicted data of 22 students taking BS subject for the June 2023 session. The actual performance level is according to the actual BS's grade obtained by students in the final examination. Meanwhile, the predicted performance level is the level of performance obtained by the prediction process before the final examination using the best model, the DT.

Table 4.3 The List of Actual VS Predicted Performance Level of 22 Students

No.	Student's ID	Actual Performance Level	Predicted Performance Level
1	PPD21002	2	2
2	PPD21015	2	2
3	PPD21021	3	4
4	PPD21026	3	3
5	PPD21028	4	4
6	PPD21034	4	3
7	PPD21025	1	1
8	PPD21033	5	4
9	PPD21004	5	2
10	PPD21005	3	3
11	PPD21006	3	3
12	PPD21017	6	4
13	PPD21019	4	4
14	PPD21022	5	2
15	PPD21023	2	1

Table 4.4 Continued

No.	Student's ID	Actual Performance Level	Predicted Performance Level
16	PPD21047	4	4
17	PPD21038	3	4
18	PPD20019	6	6
19	PPD21013	6	6
20	PPD21039	4	3
21	PPD21041	6	4
22	PPD20006	4	6

Visual inspection in Figure 4.7 reveals that the direction from the predicted and actual values are generally comparable. The predicted line, despite variances, resembles the general trend of the actual line. From the graph, eleven students got the final grade as predicted, and from the analysis, the accuracy score was 0.5.

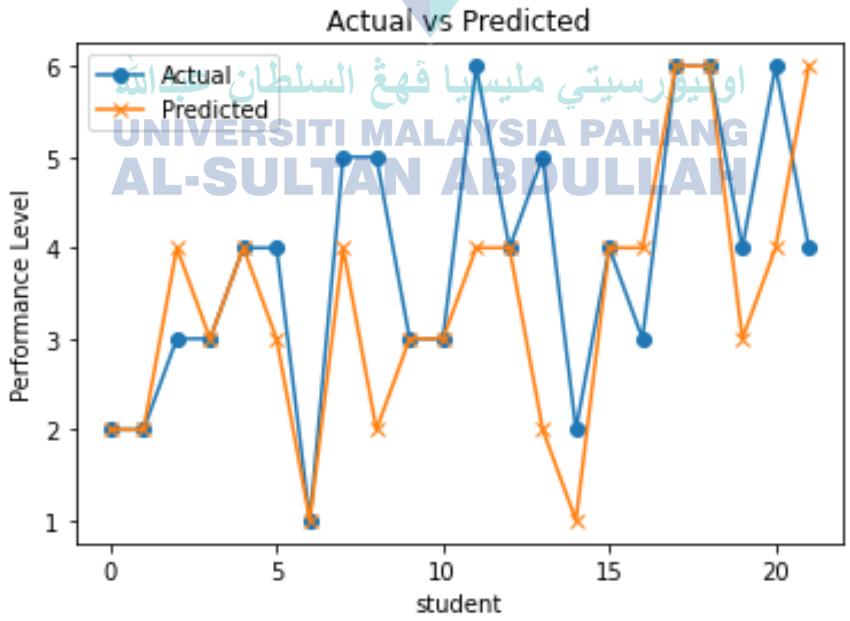


Figure 4.12 The Line Graph of Actual Versus Predicted Data for Student's Performance Levels

Each predictive model that has been developed must be able to benefit the relevant department. Accordingly, the predictive model for students' final grades in BS subject can be used by lecturers at UnIPSAS, particularly at the Faculty of Management and Informatics. This is because the fields of statistics and mathematics are from this faculty. Students who are predicted to get an E grade must be isolated and given intensive attention. Based on the findings, the student's cumulative score is most related to student failure for the BS subject. These students can retake the progress tests or do additional coursework to improve their cumulative scores.

4.6 Summary

In this chapter, the findings for each step stated in the ML pipeline have been reported to answer the study's objectives. The original data of students taking the Business Statistics subject was used to develop a predictive model to predict students' performance in the final examination. As a result of the performance evaluation, the AUC score for the ROC curve is better than the performance of accuracy, recall, F1-score, and precision. From the results obtained, the accuracy of the DT model is 0.60, making this model the most accurate model for predicting students' performance in the BS course. For testing and validation purposes, new data from the students' latest records was used to verify the effectiveness of the best-selected predictive model.

CHAPTER 5

CONCLUSION

5.1 Introduction

The main purpose of this research is to develop the most accurate predictive model for students' performance levels using machine learning methods. The background, problem statement, research objectives, scope, and significance of the study were covered in Chapter 1. In Chapter 2, the literature review was conducted under the research's keywords, namely predictive analytics research, predictive analysis employing machine learning techniques, the use of machine learning techniques in Higher Education Institutions, and the application of machine learning to predict students' performance and final grades. The machine learning method was described in depth in Chapter 3, including data collection, data preprocessing, feature selection, data splitting, predictive modelling, models' performance evaluation and model testing. This research produced important outcomes which have been described in Chapter 4. The Python software has been utilized to process the data and finish the study in accordance with the proposed machine learning pipeline.

In conclusion, this study has successfully achieved its objectives. The first objective is to determine the factors that affect the students' final grades in the Business Statistics course. Students' carry marks was the most correlated factor and strongly influenced students' performance levels. Carry marks have an inverse correlation towards performance level because the grades are represented by the level of performance (excellent (1), very good (2), good (3), pass (4), weak (5), and fail (6)). In other words, the higher level of performance is represented by the lowest encode value. The higher marks carried by students during the semester provided a lower class of performance level, which represents the higher grades.

The second objective is to develop the most accurate predictive model for students' performance levels in the Business Statistics course at UniPSAS by using the ML method. Based on the result, the decision tree (DT) was the most accurate model to predict students' performance levels for the Business Statistics subject at Universiti Islam Pahang Sultan Ahmad Shah. The model achieved the highest accuracy score compared to other models, which is 0.60. This model can be considered the best model since this field of study was for the development of learning and is not a critical field such as medical, science studies, and so on. In addition, the data that has been used is real data and there were also constraints in terms of data sources and attributes that limited the achievement of the model. Furthermore, in specific scenarios, students' performance in the final examination may differ from their performance in assessment marks. Certain students might have been unable to study for their final exams due to illness or an emergency, resulting in lower grades. Conversely, some students may have concentrated more and performed well in their final exams despite performing poorly in their assessments before the final exam. As the model cannot account for such factors, achieving a 100 percent accurate final grade prediction is impractical (Eman et al., 2016).

The third objective of the study is to evaluate the performance for the proposed models using the value of precision, recall, accuracy, F1-score, and area under receiver operating characteristics curve (AUC) as well as new dataset. The accuracy value could not give a high value when it only recorded a moderate to good score (0.51 to 0.60). Meanwhile, from the AUC results, the level of performance that was perfectly distinguished among other levels was level 6 which is grade E (fail). Therefore, it is highly beneficial for identifying students at risk of failure, enabling lecturers to take early interventions to enhance student performance. If lecturers can anticipate which students are likely to receive an E or fail before the final exam, these students can work on improving their carry marks to avoid failing in the end.

The DT model was employed to determine the predicted values of students' performance levels in the Business Statistics course for a new dataset. A total of 22 students' data from the previous academic sessions were used to test the DT model's performance. As a result, from the predicted values, DT model successfully predicted 50% students' performance levels correctly. This achievement is good enough to help the

lecturers of this subject use the DT model to predict student achievement for the next semester.

5.2 Limitation of Study

Research limitations are something that cannot be avoided and resolved when the research is done. There are several limitations in conducting this study, namely from the aspect of data size and also the attributes. The size of the data for this study is not too large considering that students who registered for management programs that were taking business statistics course were not as many as students who take other programs at UnIPSAS. In addition, since this study has collected data from the UnIPSAS learning system, there is a constraint to obtaining additional information from the students themselves as an attribute of this study. This is because the students have completed their studies at UnIPSAS. The data that can be accessed from the system was also quite limited.

5.3 Suggestion and Recommendation for Future Work

In the future, adding more data on students enrolled in this course is recommended to enhance the accuracy of the model. Further analysis and improvements are necessary to enhance the model's accuracy and align the predicted values more closely with the actual values. It is crucial to reassess the model, evaluate alternative approaches, and consider additional variables or techniques that may help improve the predictive capability and accuracy of the model. Furthermore, at the end of this study, it is recommended that academicians be able to utilize the predictive model to forecast students' grades leading up to the final examination. The algorithm can also be added to the Learning Management System along with a dashboard so that it is easier to do analyses in the future.

REFERENCES

Abana, E. C. (2019). A Decision Tree Approach for Predicting Student Grades in Research Project using Weka. In IJACSA) International Journal of Advanced Computer Science and Applications (Vol. 10, Issue 7). www.ijacsa.thesai.org

Abdul Bujang, S. D., Selamat, A., & Krejcar, O. (2021). A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning. IOP Conference Series: Materials Science and Engineering, 1051(1), 012005. https://doi.org/10.1088/1757-899x/1051/1/012005

Abu Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education, 16(1). https://doi.org/10.1186/s41239-019-0160-3

Ahmad, N., Hassan, N., Jaafar, H., & Enzai, N. I. M. (2021). Students' Performance Prediction using Artificial Neural Network. IOP Conference Series: Materials Science and Engineering, 1176(1), 012020. https://doi.org/10.1088/1757-899x/1176/1/012020

Aileen Chun Yueng Hong, KHAW, K. W., XINYING CHEW, & WAI CHUNG YEONG. (2023). Prediction of US airline passenger satisfaction using machine learning algorithms. Data Analytics and Applied Mathematics (DAAM), 8–24. https://doi.org/10.15282/daam.v4i1.9071

Akoglu, H. (2018). User's guide to correlation coefficients. In Turkish Journal of Emergency Medicine (Vol. 18, Issue 3, pp. 91–93). Emergency Medicine Association of Turkey. https://doi.org/10.1016/j.tjem.2018.08.001

Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting Students' Performance Using Machine Learning Techniques. JOURNAL OF UNIVERSITY OF BABYLON for Pure and Applied Sciences, 27(1), 194–205. https://doi.org/10.29196/jubpas.v27i1.2108

Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M. (2019, July 1). A Predictive Model for Predicting Students Academic Performance. 10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019. https://doi.org/10.1109/IISA.2019.8900760

Anderson, T. and Anderson, R. (2017). 'Applications of Machine Learning To Student Grade Prediction in Quantitative Business Courses', Global Journal of Business Pedagogy, 1(3), pp. 13–22.

Anshul Saini (2024, January 23). Guide on Support Vector Machine (SVM) Algorithm. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/Support-vector-machinessvm-a-complete-guide-for-beginners/

Barnabas Ndlovu Gatsheni, & Olga Ngala Katambwa. (2018). The Design of Predictive Model for the Academic Performance of Students at University Based on Machine Learning. J. of Electrical Engineering, 6(4). https://doi.org/10.17265/2328-2223/2018.04.006

Basheer, M. Y. I., Mutalib, S., Hamid, N. H. A., Abdul-Rahman, S., & Malik, A. M. A. (2019). Predictive analytics of university student intake using supervised methods. IAES

International Journal of Artificial Intelligence, 8(4), 367–374. https://doi.org/10.11591/ijai.v8.i4.pp367-374

Data normalization in Python. (2022). https://www.educative.io/answers/data-normalization-in-python

Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of Students Performance using Machine learning. IOP Conference Series: Materials Science and Engineering, 1055(1), 012122. https://doi.org/10.1088/1757-899x/1055/1/012122

Eman, E., Majeed, E. A., & Junejo, K. N. (2016). Grade Prediction Using Supervised Machine Learning Techniques. https://www.researchgate.net/publication/304689292

Esquivel, J. A., & Esquivel, J. A. (2021). A Machine Learning Based DSS in Predicting Undergraduate Freshmen Enrolment in a Philippine University. International Journal of Computer Trends and Technology, 69(5), 50–54. https://doi.org/10.14445/22312803/ijctt-v69i5p107

Hameed, Z., Garcia-Zapirain, B., Aguirre, J. J., & Isaza-Ruget, M. A. (2022). Multiclass classification of breast cancer histopathology images using multilevel variables of deep convolutional neural network. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-19278-2

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. Procedia Technology, 25, 326–332. https://doi.org/10.1016/j.protcy.2016.08.114

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). Machine Learning Based Student Grade Prediction: A Case Study. http://arxiv.org/abs/1708.08744

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences (Switzerland), 11(7). https://doi.org/10.3390/app11073130

Karlos, S., Kostopoulos, G., & Kotsiantis, S. (2020). Predicting and interpreting students' grades in distance higher education through a semi-regression method. Applied Sciences (Switzerland), 10(23), 1–19. https://doi.org/10.3390/app10238413

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. In *Computational and Structural Biotechnology Journal* (Vol. 15, pp. 104–116). Elsevier B.V. https://doi.org/10.1016/j.csbj.2016.12.005

Kendale, S., Kulkarni, P., Rosenberg, A. D., & Wang, J. (2018). Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology, 129(4), 675–688. https://doi.org/10.1097/ALN.0000000000002374

Khan, B., Sikandar, M., Khiyal, H., & Khattak, M. D. (2015). Final Grade Prediction of Secondary School Student using Decision Tree. In International Journal of Computer Applications (Vol. 115, Issue 21)

Kumar, S. (2021, June 23). Why Python is Best for AI, ML, and Deep Learning. https://www.rtinsights.com/why-python-is-best-for-ai-ml-and-deep-learning/

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Variable selection: A data perspective. In ACM Computing Surveys (Vol. 50, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3136625

Mahesh, B. (2018). Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Self Flowing Generator View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review. International Journal of Science and Research. https://doi.org/10.21275/ART20203995

Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. In Data Science Journal (Vol. 18, Issue1) Ubiquity Press. https://doi.org/10.5334/dsj-2019-014

Michael Galarnyk (2022, July 28). Understanding Train Test Split. builtin. https://builtin.com/data-science/train-test-split

Mishra, N., Silakari, D., Proudyogiki Vishwavidyalaya, G., Sc, C., & Gandhi Proudyogiki Vishwavidyalaya, R. (2012). Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

Nieto, Y., Gacia-Diaz, V., Montenegro, C., Gonzalez, C. C., & Gonzalez Crespo, R. (2019). Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions. IEEE Access, 7, 75007–75017. https://doi.org/10.1109/ACCESS.2019.2919343

Ofori, F., & Gitonga, R. (2020). Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review Cloud security View project Cloud securuty View project. https://www.researchgate.net/publication/340209478

Oyedeji, A. O., Salami, A. M., Folorunsho, O., & Abolade, O. R. (2020). Analysis and Prediction of Student Academic Performance Using Machine Learning. JITCE (Journal of Information Technology and Computer Engineering), 4(01), 10–15. https://doi.org/10.25077/jitce.4.01.10-15.2020

Pedregosa Fabianpedregosa, F., Michel, V., Grisel Oliviergrisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot Andédouardand, M., Duchesnay, Andédouard, & Duchesnay Edouardduchesnay, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In Journal of Machine Learning Research (Vol. 12). http://scikit-learn.sourceforge.net

Pojon, M. (2017). Using Machine Learning to Predict Student Performance. M.Sc. thesis, Faculty of Natural Sciences Software Development, University of Tampere, Tampere, Finland

Poornima, S., & Pushpalatha, M. (2018). A survey of predictive analytics using big data with data mining. International Journal of Bioinformatics Research and Applications, 14(3),

269–282. https://doi.org/10.1504/IJBRA.2018.092697

Razali, M. N., Zakariah, H., Hanapi, R., & Rahim, E. A. (2022). Predictive Model of Undergraduate Student Grading Using Machine Learning for Learning Analytics. Proceedings - 2022 4th International Conference on Computer Science and Technologies in Education, CSTE 2022, 260–264. https://doi.org/10.1109/CSTE55932.2022.00055

Roslan, N., Mohd Jamil, J., Nizal, I., & Shaharanee, M. (2021). Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application. In Turkish Journal of Computer and Mathematics Education (Vol. 12, Issue 3).

Sani, N. S., Nafuri, A. F. M., Othman, Z. A., Nazri, M. Z. A., & Nadiyah Mohamad, K. (2020). Drop-Out Prediction in Higher Education Among B40 Students. International Journal of Advanced Computer Science and Applications, 11(11), 550–559. https://doi.org/10.14569/IJACSA.2020.0111169

Sapra, R., & Saluja, S. (2021). Understanding statistical association and correlation. Current Medicine Research and Practice, 11(1), 31. https://doi.org/10.4103/cmrp.cmrp_62_20

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shmueli, G., & Koppius, O.R. (2011). Predictive Analytics in Information Systems Research. Economics of Networks eJournal.

Tatar, A. E., & Düştegör, D. (2020). Prediction of academic performance at undergraduate graduation: Course grades or grade point average? Applied Sciences (Switzerland), 10(14). https://doi.org/10.3390/app10144967

Tekin, A. (2014). Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach. In Eurasian Journal of Educational Research (Vol. 54).

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. Journal of Diagnostic Medical Sonography (JDMS), 35–39. https://doi.org/10.1177/875647939000600106

Ünal, F. (2020). Data Mining for Student Performance Prediction in Education. www.intechopen.com

Venkat, N. (2018). Predicting Student Grades using Machine Learning.

Xu, J., Ho Moon, K., & van der Schaar, M. (2017.). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742-753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560

Yan, Zhihong & Wang, Shuqian & Ma, Ding & Liu, Bin & Lin, Hong & Li, Su. (2019). Meteorological Factors Affecting Pan Evaporation in the Haihe River Basin and China. Water. 11. 317. 10.3390/w11020317.

Yildiz, M., & Börekci, C. (2020). Predicting Academic Achievement with Machine Learning

Algorithms. Journal of Educational Technology and Online Learning. https://doi.org/10.31681/jetol.773206

Zabriskie, C., Yang, J., Devore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. Physical Review Physics Education Research, 15(2). https://doi.org/10.1103/PhysRevPhysEducRes.15.020120

APPENDICES

Appendix A: The Summary of Literature Reviews.

Author	Year	Target Variable					Algorithm															
		CGPA	GPA	Final Grade	Final Mark/Score	Performance Category	Decision Tree (DT)	Random Forest (RF)	Support Vector Machine (SVM)	Naïve Bayes (NB)	k-Nearest Neighbor (k-NN)	Linear Regression	Semi-Regression	Artificial Neural Networks (ANN)	Bayesian Network (BN)	Rule Induction	Logistic Regression	Fuzzy Genetic Algorithm (FGA)	Latent Dirichlet Allocation (LDA)	Multilayer Perceptron (MLP)	Ensemble Model
Hassan Zeineddine et al.	2021					X															X
J. Dhilipan et al.	2021			X			X				X						X				
Siti Dianah Abdul Bujang et al.	2020			X			**X**	X	X			X									
Muhammed Berke YILDIZ et al.	2020					X	X	X	X	X	X	X		**X**							
Ajibola O. Oyedeji et al.	2020				X							**X**		X							
Phauk Sokkhey et al.	2020				X		X	X	X	X											
Taiwo Olaleye et al.	2020		X				**X**			X											
Stamatis Karlos et al.	2020			X									X								
Cabot Zabriskie et al.	2019			X				**X**									**X**				
Ferda Ünal et al.	2019			X			X	X		X											
Fazal Aman et al.	2019			X																	
Abu Zohair et al.	2019			X					**X**	X	X								X	X	
B. Prasanalakshmi et al.	2019				X		X	X	X								X			X	
Alaa Khalaf Hamoud et al.	2019					X								X							

Ammar Almasri et al.	2019					X																X
E. T. Lau et al.	2019	X													X							
Abana E et al.	2019			X				X														
Diego Buenaño-Fernández et al.	2019					X		X														
Hussein Altabrawee et al.	2018			X				X			X				**X**			X				
Naveen Venkat et al.	2018			X				X		**X**	X	X										
Barnabas Ndlovu Gatsheni et al.	2018		X					X		**X**						X						
Anderson et al.	2017			X						**X**	X	X										
Jie Xu et al.	2017			X					**X**			X	X									
Murat Pojon et al.	2017			X				**X**			**X**		X						X			
Emaan Abdul Majeed et al.	2016			X				X			X	X					**X**					
Agoritsa Polyzou et al.	2016			X									X									
Hashmia Hamsa et al.	2016				X			**X**												X		
Syed Tanveer Jishan et al.	2015			X				X			X				X							
Bashir Khan et al.	2015			X				X														
Anal Acharya et al.	2014			X				**X**		X					X	X						

The bold X's (**X**) is representing the best model in particular studies.

Appendix B: Descriptive Statistics

	gender	course	intake	CGPA	math_grade	level	CM_clip
count	449.000000	449.000000	449.000000	449.000000	449.000000	449.000000	449.000000
mean	0.325167	1.111359	1.091314	2.978575	6.859688	1.167038	39.727372
std	0.468960	0.957154	0.288376	0.515075	3.228003	1.328325	4.941438
min	0.000000	0.000000	1.000000	1.640000	1.000000	0.000000	28.450000
25%	0.000000	0.000000	1.000000	2.600000	4.000000	0.000000	37.300000
50%	0.000000	1.000000	1.000000	3.010000	7.000000	1.000000	40.500000
75%	1.000000	2.000000	1.000000	3.410000	10.000000	2.000000	43.200000
max	1.000000	3.000000	2.000000	3.960000	11.000000	5.000000	49.100000

Appendix C: Treating Outliers using Winsorize Method

TREAT THE OUTLIERS

```
In [19]: Q1=data_raw['CM'].quantile(0.25)
         Q3=data_raw['CM'].quantile(0.75)
         IQR=Q3-Q1

In [20]: Q1
Out[20]: 37.3

In [21]: Q3
Out[21]: 43.2

In [22]: IQR
Out[22]: 5.900000000000006

In [23]: lower_lim=Q1-IQR*1.5
         higher_lim=Q3+IQR*1.5

In [24]: lower_lim
Out[24]: 28.44999999999999

In [25]: higher_lim
Out[25]: 52.05000000000001
```


#Treating the Outliers with Winsorize method: limit outliers with an upper and lower limit. set the limit.

```
In [26]: data_raw['CM_clip']=data_raw['CM'].clip(lower=lower_lim, upper=higher_lim)

In [27]: data_raw['CM_clip'].describe()
Out[27]: count    449.000000
         mean      39.727372
         std        4.941438
         min       28.450000
         25%       37.300000
         50%       40.500000
         75%       43.200000
         max       49.100000
         Name: CM_clip, dtype: float64
```

Appendix D: List of Predictors (x) and Target Variable (y)

DECLARE FEATURES X, Y

In [27]:
```
x = data_raw.drop('level',axis=1)
y = data_raw['level']
```

In [53]:
```
y
```

Out[53]:
```
0      2
1      1
2      1
3      2
4      1
      ..
445    4
446    1
447    4
448    6
449    6
Name: level, Length: 449, dtype: int64
```

In [54]:
```
x
```

Out[54]:

	gender	course	CGPA	math_encode	CM_clip
0	1	1	2.88	10	44.10
1	0	2	3.15	2	42.00
2	0	2	3.33	2	42.40
3	0	1	2.17	5	41.40
4	0	2	3.41	1	44.60
...	...	...	...	...	...
445	0	1	2.76	8	38.17
446	0	0	3.43	4	43.90
447	1	2	3.38	5	38.40
448	0	1	2.92	3	28.45
449	0	3	2.52	11	28.45

449 rows × 5 columns

Appendix E: The process of Testing DT model on New Dataset in Python

```
In [199]: testing_data=pd.read_csv('Real_data.csv')
```

```
In [200]: testing_data.head()
```

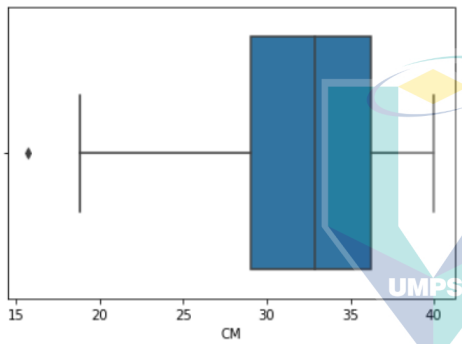
Out[200]:

	name	ID	gender	course	intake	CGPA	CM	math_encode	final_grade	Per_level
0	AMY LYANA MAISARA BINTI MUSLY ZAMBRIE	PPD21002	0	0	1	3.30	35.2	5	B	2
1	NURUL HIDAYAH NINTI YUSSOF	PPD21015	0	0	1	3.41	39.6	3	B	2
2	ISMAT IMRAN BIN SUHAIRULAZHAR	PPD21021	1	0	1	2.82	35.3	5	C+	3
3	NOR AFIZA AMIRA BINTI MOHAMMAD AFIZU	PPD21026	0	0	1	3.46	36.5	6	B-	3
4	SYAZWAN IQBAL BIN SHAHRUL MUNIR	PPD21028	1	0	1	2.27	30.9	7	C-	4

```
In [191]: sns.boxplot(testing_data.CM)
```

```
C:\Users\User\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variab
arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
t keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[191]: <AxesSubplot:xlabel='CM'>




```
In [201]: Valid_Q1=testing_data['CM'].quantile(0.25)
          Valid_Q3=testing_data['CM'].quantile(0.75)
          Valid_IQR=Valid_Q3-Valid_Q1
```

```
In [202]: lwr_lim=Valid_Q1-Valid_IQR*1.5
          hgh_lim=Valid_Q3+Valid_IQR*1.5
```

```
In [203]: testing_data['CM_clip']=testing_data['CM'].clip(lower=lwr_lim, upper=hgh_lim)
```

```
In [204]: testing_data['CM_clip'].describe()
```

```
Out[204]: count    22.000000
          mean     31.899432
          std       6.222932
          min      18.187500
          25%      29.025000
          50%      32.900000
          75%      36.250000
          max      40.000000
          Name: CM_clip, dtype: float64
```

In [205]:
```
testing_data.drop(['name','ID','final_grade','intake'],axis=1,inplace=True)
testing_data.head()
```

Out[205]:

	gender	course	CGPA	CM	math_encode	Per_level	CM_clip
0	0	0	3.30	35.2	5	2	35.2
1	0	0	3.41	39.6	3	2	39.6
2	1	0	2.82	35.3	5	3	35.3
3	0	0	3.46	36.5	6	3	36.5
4	1	0	2.27	30.9	7	4	30.9

In [206]:
```
testing_data.drop(['CM','Per_level'],axis=1,inplace=True)
testing_data.head()
```

Out[206]:

	gender	course	CGPA	math_encode	CM_clip
0	0	0	3.30	5	35.2
1	0	0	3.41	3	39.6
2	1	0	2.82	5	35.3
3	0	0	3.46	6	36.5
4	1	0	2.27	7	30.9

In [207]:
```
prediction=dtree.predict(testing_data)
```

In [208]:
```
print(prediction)
```

```
[2 2 4 3 4 3 1 4 2 3 3 4 4 2 1 4 4 6 6 3 4 6]
```

Appendix F: The Process of Evaluating DT Model Accuracy on New Dataset in Python

In [209]:
```
Accuracy_score=pd.read_csv('DT.predict.csv')
Accuracy_score.head(22)
```

Out[209]:

	y_real	y_predict
0	2	2
1	2	2
2	3	4
3	3	3
4	4	4
5	4	3
6	1	1
7	5	4
8	5	2
9	3	3
10	3	3
11	6	4

In [210]:
```
true_label=Accuracy_score['y_real']
predicted_label=Accuracy_score['y_predict']
```


In [211]:
```
from sklearn.metrics import accuracy_score

# Calculate accuracy
accuracy = accuracy_score(true_label, predicted_label)

# Print the accuracy
print("Accuracy:", accuracy)
```

```
Accuracy: 0.5
```