PREDICTING STUDENTS' ACADEMIC PERFORMANCE USING A HYBRID OF MACHINE LEARNING ALGORITHMS

ROSELYNE BONARERI AYIENDA

MASTER OF SCIENCE

(Information Technology)

JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

2024

Predicting Students' Academic Performance Using a Hybrid of Machine Learning Algorithms

Roselyne Bonareri Ayienda

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Technology of the Jomo Kenyatta University of Agriculture and Technology

2024

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University

Signature……Date……

Roselyne Bonareri Ayienda

This thesis has been presented for examination with our approval as the University Supervisors

Signature……Date……

Dr. Richard Rimiru, PhD

JKUAT, Kenya

Signature……Date……

Prof. Wilson K. Cheruiyot, PhD

JKUAT, Kenya

DEDICATION

I dedicate this thesis to my Loreto family members for their financial support and my family members Charlse, Elius, Elijah and Jackline for their moral support and patience during my busy schedule and to my parents Raphael Ayienda and Christine Kwamboka for their constant encouragement in my academics. To my supervisors Dr.Richard Rimiru and Prof. Wilson Cheruiyot for their tireless guidance in ensuring that this work was accomplished accordingly. I salute you all and may our Almighty God grant you thy abundant blessings.

ACKNOWLEDGEMENT

To God is the glory for the strength and protection upon my life. I also take this moment to sincerely appreciate my supervisors Dr. Richard Rimiru and Prof. Wilson Cheruiyot for their dedicated efforts towards the completion of this thesis. Many thanks also go to Jomo Kenyatta University of Agriculture and Technology class work lecturers for giving me the basic knowledge towards the recognition and finally the development of this work. I also salute my master's student colleagues for their constant encouragement and advice towards the completion of my work. May God bless you all for the support that you gave to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ACRONYMS AND ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area under the Curve
CV	Cross Validation
DBN	Deep Belief Network
EDM	Educational Data Mining
FN	False Negative
FP	False positive
FS	Feature Selection
IDS	Intrusion Detection System
IoT	Internet of Things
K-NN	K-Nearest Neighbor
LR	Logistic Regression
LSSVM	Least Square Support Vector Machine
MI	Mutual Information
ML	Machine Learning
MLPNN	Multilayer Perceptron Neural Network
MLR	Multi Linear Regression
NB	Naïve Bayes
OASIS	Open Access Series of Imaging Studies
PCA	Principle Component Analysis
RMSE	Root Mean Square Error

SAP	Student Academic Performance
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WGI	Weighted Gini Index
WVC	Weighted Voting Classifier

ABSTRACT

The research highlights the significance of student academic performance as a crucial factor in determining the success of educational institutions. Educational data mining (EDM) is employed to analyze educational data, aiming to explore student academic performance. The study proposes a novel, feature-rich model for predicting student performance, integrating backlog information and student grades identified as key aspects through data mining techniques. It demonstrates the feasibility of developing a predictive model with satisfactory accuracy rates, even when trained on a limited dataset. Additionally, the research identifies critical parameters such as student behavior, family education, and subject grade averages essential for constructing the model and presenting data. To determine the optimal model, the study evaluates various algorithms using important attributes. A diverse set of classifiers, including decision trees, support vector machines (SVM), and k-nearest neighbor (KNN), is employed to assess the model's efficiency. Moreover, ensemble techniques such as bagging, boosting, stacking, and random forest are utilized to enhance classifier performance. The research concludes by establishing a clear correlation between students' attributes (such as social interactions and absenteeism), past exam performance (G2), family education (mothers' education), and their final grades (G3). With an accuracy rate of 91.5%, the findings validate the effectiveness of ensemble approaches in improving prediction models for student academic performance

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

According to Ajibade et al., (2018) prediction of students' academic performance has been an enormous concern for higher institutions everywhere. The massive usability of LMS has produced immense quantity of data on interactions and communications between teachers and students. Vora and Rajamani (2022) discovered that the data collected holds concealed insights that can enhance students' academic performance. Predicting students' educational outcomes has become more complex because of the large volume of data in the databases. Current prediction systems struggle to effectively analyze and track student progress, often overlooking potential strong performances. This is attributed to the selection of unsuitable methods and delayed investigations. Opara et al. (2020) found that over the years, large record of student data exists in institutions of higher learning because students graduate from these institutions yearly. This has necessitated the need to explore those data to discover some patterns and relationships existing in them and as well make strategic decisions for a better education system. Burman & Som (2019) focused on developing an enhanced hybrid data mining model to mine students' progress and performance for knowledge discovery and for decision making purposes. Education plays a vital role in a person's life. It helps in overall development of an individual.

According to Adejo and Connolly (2018), learning goes beyond the traditional student-teacher relationship and includes methods such as storytelling, meetings, mentorship, and analysis. These forms of education occur throughout a person's life and play a key role in shaping an individual's future. Students take various exams at different stages, and one major challenge facing higher education worldwide is the high attrition rate, especially during the first year. Educational institutions are seeking ways to boost retention and graduation rates by identifying at-risk students' early on using performance prediction.

However, most existing predictive models lack efficiency and accuracy due to limitations within the individual classifiers used and the inclusion of limited or irrelevant variables.

According to Zulfiker et al., (2020) the databases of the different universities store a large volume of data. This data includes the data of the students, teachers, and employees of the universities. By analyzing this data, different patterns can be derived which will be helpful to make decisions. Using diverse machine learning and data mining techniques on these data, many kinds of knowledge can be discovered and this knowledge can be used to predict the enrolment status of the students in a course, to detect illegal activities in the online examination and, to identify unusual marks in the result sheet Educational Data Mining (EDM) represents an innovative approach aimed at uncovering valuable patterns through data mining techniques. Vora & Rajamani (2019) stated that extracting valuable insights from academic information systems, such as those dealing with enrollment, admissions, and management across various educational levels, including schools, colleges, and universities, is crucial. Researchers in this field focus on improving students' learning outcomes and boosting performance efficiency.

Pedraza & Beruvides, (2016) found that in Educational Data Mining (EDM), the use of multiple data sources combined with diverse ensemble techniques is highly efficient and accurate for predicting student performance and identifying students at risk of dropping out. Through data mining, university managers can extract knowledge from student management system datasets, uncovering hidden information that allows them to quickly develop new strategies to enhance students' academic performance. Students can improve their final course grades and intermediate exam results by focusing on early performance and adjusting controllable academic behaviors such as class attendance. Excelling early in a course can encourage positive academic habits throughout the semester, while achieving intermediate goals like performing well on quizzes can contribute to success on exams.

The research that was done by Adejo & Connolly, (2018) revealed that the educational administrators and policymakers working within educational sector in the development of new policies and curriculum on higher education that are relevant to student retention. In addition, the general implications of this research is to practice its ability to accurately help in early identification of students at risk of dropping out of Higher Education from the combination of data sources so that necessary support and intervention can be provided. Bhogan et al., (2017) found that by predicting student performance, instructors can help to improve student performance in the examination and significantly reduce drop out ratio from college, which will enhance the performance of college students

1.2 Statement of the Problem

According to Solomon, (2018) education is a fundamental component of personal and societal development, yet students face numerous challenges that impact their academic journey Educational Data Mining (EDM) employs data mining tools within educational contexts, aiming to predict student performance based on key attributes in existing datasets. However, selecting the most effective technique and features for prediction models remains an overwhelming task.

Siddiai & Pak (2020) introduced a novel process flow for filter-based feature selection, incorporating normalization or transformation before classification to mitigate these challenges. By implementing and evaluating the effects of normalization before feature selection, their approach seeks to enhance the predictive accuracy of models. The efficiency and effectiveness of feature selection methods are often hindered by high data dimensionality.

The research by Chen et al., (2020) revealed that feature selection is crucial for optimizing data mining and analysis by removing irrelevant features from datasets. However, the diverse criteria employed by various feature selection algorithms pose challenges in determining the most suitable algorithm for specific datasets. To address this, ensemble

methods integrate multiple feature selection outcomes, overcoming the limitations of individual techniques.

Chen et al., (2020) found that despite the widespread categorization of feature selection algorithms into filter, wrapper, or embedded techniques, there has been limited exploration of tree-based feature selection methods like Decision Tree, XG Boost, and Random Forest in predicting student performance This research utilized tree-based feature importance techniques, exploring longitudinal and temporal features to enhance predictive accuracy and enable personalized interventions. By leveraging these techniques, the efficacy of educational interventions and support systems can be improved, ultimately enhancing student outcomes.

1.3 The Objectives

1.3.1 The General Objectives

The main objective of this research is to develop a student performance prediction model that leverages on longitudinal and temporal features using an ensemble model of machine learning algorithms.

1.3.2 Specific Objectives

The specific objectives of this research were to:

i. To investigate the primary attributes utilized in the prediction of student performance using ensemble Machine Learning (ML) approaches for students' academic performance prediction.
ii. To develop an ensemble model using machine learning algorithms to predict students' academic performance.
iii. To evaluate the developed hybrid model of predicting students' academic performance that used ensemble techniques and base classifiers to predict the student performance.

1.3.3 Research Questions

The following are questions that guided the research into achieving all its objectives:

i. What are the primary attributes used in predicting students' academic performance?
ii. Which ensemble techniques are robust and for predicting student's academic performance?
iii. What is the accuracy of the student prediction model?

1.4 Justification

Francis & Babu, (2019) found that prediction of students' academic performance can be done using MLP and SVM classification techniques. The achievement in academics of students is a huge concern for academic institutions all over the globe. The huge use of learning management system generates huge amount of data about learning and teaching interactions. These data comprise of hidden knowledge that could be used to develop the academic performance of students. Most education management systems lack away of mining deep hidden information from the continuously growing volumes of educational data. This prevents the institution from achieving its quality objectives since it lacks deeper knowledge about its increasing data. These systems do not provide lecturers with advanced information about students who might need more attention. Despite keeping students these systems do not provide predicting or warning tools that can assist university managers to extract knowledge on the performance of their students.

Therefore, it will be of great help for both the learner and the institution, if there is prior knowledge of the student final performance. The performance prediction depends on variables within the school as well as an outside school that affect the academic performance of the students. Kapur (2018) said that the results obtained after the prediction will be useful for instructor as well as students and it will help in taking appropriate decision to improve student's performance. Bhogan et al., (2017) showed that

“Higher education has long been rich in data but slow in converting that data into useful information”. However, through the Educational Data Mining models approach like the one in this research, these knowledge gaps can easily be bridged and aid institutions in achieving their goals.

1.5 Scope of the Research

The main focus of this research was on predicting students’ academic performance by first doing data prepossessing and getting the features with a higher contribution to students’ academic performance using tree-based feature selection techniques. Then to use the features that have been gotten with the traditional classifiers that is SVM, Decision Tree and KNN to get their accuracy with 10-fold cross validation. Then apply the ensemble techniques that is Bagging, Boosting, Random Forest and Stacking to improve their accuracy then do the testing and validation of the machine learning algorithms and get their corresponding accuracies.

1.6 Research Contribution

This research will explore novel feature selection techniques that leverage longitudinal and temporal features to improve predictive accuracy and facilitate personalized interventions. Educational datasets often comprise of longitudinal datasets, such as students' academic courses, course data, and temporal patterns of engagement. Feature selection techniques custom-made to handle longitudinal and temporal data could help recognize useful features that capture students' changing learning patterns and performance over time.

1.7 Limitation of the Research

There are certain limitations to this study that should be mentioned. The study relies on publicly available datasets rather than a student dataset. Furthermore, the dataset was limited, with only a few hundred records. More data-driven research may yield more conclusive results. The majority of EDM researchers are currently cagey to release their

study dataset for two reasons: The first is concerned with privacy, integrity, and legality; the second is with dataset collecting, which is a laborious, time-consuming, and costly operation. Based on a combination of privacy protection, economic impact, and scholarly ramifications. This study employed offline data, but an increasing amount of online data remains untapped, allowing us to train the model to predict offline student performance in real-time.

1.8 Structure of the Thesis

The organization of the thesis is as follows: Chapter 1: Give the background of the research, statement of the problem, the research objectives, and research questions, justification of the problem, the scope of the research and limitation of the research. Chapter 2: Discusses the related studies that includes Machine learning overview, the machine learning techniques used, the theory of the classifiers used in prediction, feature selection methods used in educational data, general applications of the classifiers used in prediction; educational data mining, SVM, Decision Tree and KNN as approaches to performance prediction then the research gaps. Chapter 3: Gives the student data followed by data pre-processing, Feature importance analysis, Data splitting, setting up of the environment required for the experiment, training of the model, testing of the traditional classifiers, applying of the ensemble techniques then validation and evaluation of the classifiers 4: Gives the result analysis and discussion, metrics measures and the results evaluation. Chapter 5: Gives the Conclusion and the future recommendations of the research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Chapter Summary

This chapter, gives the machine learning overview followed by theoretical background of the classifiers. The general application of machine learning in prediction and their related literature are discussed followed by machine learning application in education data mining and their related literature.

2.2 Machine Learning Overview

Machine learning techniques often involve a learning process that aims to complete a task based on "experience" gained from training data. In machine learning, data consists of instances, each described by a set of attributes known as features or variables. These features can be nominal (categories), binary (0 or 1), ordinal (e.g., A+ or B), or numeric (Sharma & Ranjan 2021). A performance metric that improves with experience is used to evaluate the performance of an ML model on a specific task. Various statistical and mathematical models have been utilized to calculate the performance of ML models and algorithms. After the learning process is completed, the trained model can be applied to classify, predict, or cluster new examples (testing data) based on the expertise gained during the training phase. (Liakos et al.,2018), Figure 2.1 depicts a typical machine learning process.

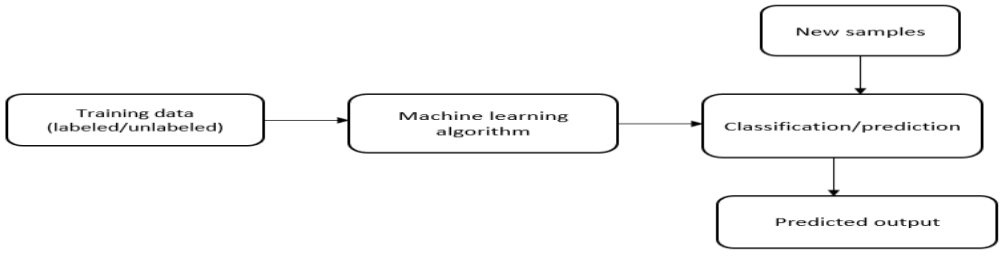


Figure 2.1: Machine Learning Process

Source: (Liakos et al., 2018).

2.2.1 Machine Learning Techniques

Machine learning is a form of artificial intelligence that enables computers to learn independently. The three types of machine learning are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning aims to construct a model from labeled training data that can predict future data. Two types of supervised learning include classification, which predicts categorical outcomes, and regression, which predicts continuous outcomes. The objective of reinforcement learning is to develop a system or agent that enhances its performance based on interactions with its environment. (Alloghani et al., 2020)

2.3 Theory of Classifiers used in Prediction

2.3.1 Naïve Bayes

Nave Bayes classifier is a type of probabilistic classifier that is based on the Bayes theorem and assumes high independence between features or characteristics (Murty & Devi 2011). The letters X and C in equation 2.1 stand for evidence and hypothesis, respectively. $P\ (C|X), given$ examples or evidence X and assuming C, equation 2.1 is used to compute the chance of C occurring given evidence X. This is known as Posterior probability.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{2.1}$$

Where $P(c)$ is the probability of the hypothesis C

Where $P(c|x)$ is the class posterior probability and $P(c)$ is the prior probability class: $P(x|c)$ is the likelihood that is the predictor given class probability.

P(x) is the predictor prior probability.

Evidence X is in the form which is shown in the equation 2.2

$$\{X = X_1, X_{2,} \dots \dots X_n\} \tag{2.2}$$

X is a set of features or attributes, and n is the total number of attributes in the set. The goal of Nave Bayes is to give a prediction on the hypothesis C with the highest posterior probability given an instance of evidence X.

$$P(C_1|X) > P(C_J|X) \tag{2.3}$$

The maximum posterior hypothesis is the class C_i with the highest $P(C_i|X)$. A previously unknown example is assigned to a class with the highest posterior. Since educational data is characterized by multiple attributes, X_i to X_n, obtaining all of the necessary probabilities to construct $P(X|C_i)$ is computationally intensive (likelihood). Therefore equation 2.4, assumes of class conditional independence, which says that the probabilities of distinct characteristics having specific values are conditionally independent of one another, and we get $P(X|C_i)$.

$$P(X|C_i)\prod_{k=1}^{n} P(X|C_i) = P(X_1|C_i) \times P(X_2|C_i) \times \ldots \times P(X_k|C_i) \tag{2.4}$$

Instead of knowing the class conditional probabilities for every combination of X, we can estimate the conditional probability of eachX_i, given C and the prior probabilities of X_i and C, using the conditional independence assumption. The Nave Bayes classifier computes the posterior probability for each class C using the formula in equation 2.5 to categorize an unseen student record.

$$P(C|X) = \frac{(C)\prod_{i=1}^{d} P(X_i|C)}{P(X)} \tag{2.5}$$

$P(X)$ is constant for each class since it is independent of class membership. As a result, selecting the class that maximizes the numerator term is adequate (Hasudungan, 2020). Extracting equation 2.5 from equations 2.4 and eliminating the denominator, where $P(C)$ indicates proportionate.

$$P(C|X) \propto P(C)\prod_{i=1}^{d} P(X_i|C) \tag{2.6}$$

2.3.2 Multilayer Perceptron

Multi-layer Perceptron (MLP) is a neural network capable of learning the relationship between linear and non-linear data. MLP is named for its resemblance to human perception and consists of at least three layers of neurons: input, hidden, and output layers. The network is interconnected internally but not directly linked to the external environment. Generally, the hidden layer is primarily responsible for processing information. In some instances, an MLP can include multiple hidden layers, especially when the input units are linear. However, it has been shown that a single hidden layer is sufficient to approximate any continuous non-linear function, as long as there are enough input units within the network (Altaf et al., 2019). The general structure of a fully connected MLP with input nodes, hidden neurons, and output neurons is shown in Figure 2.2

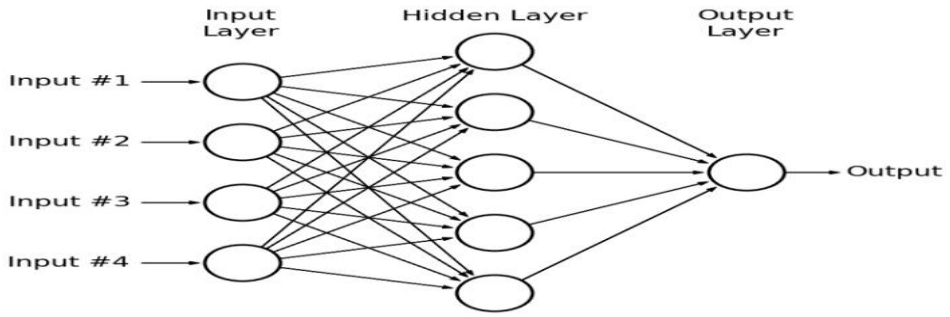


Figure 2.2: General Structure of Artificial Neural Network

Source: (Altaf et al., 2019)

MLP has the advantage of being applied to complex non-linear problems, works well with large input data. It provides quick predictions after training. The same accuracy ratio can be achieved even with smaller data. It has also the following draw back it is not known to what extent each independent variable is affected by the dependent variable. (Rezaei et al., 2022). Computations are difficult and time consuming. The proper functioning of the model depends on the quality of the training (Liu & Chao 2021). The error of the k^{th}

output node in the data point n can be represented by the equation below where d and c represent the actual and predicted values respectively.

$$e_k(n) = d_k(n) - c_k(n) \tag{2.7}$$

2.3.3 Support Vector Machine

SVM (Support Vector Machine) is a supervised learning technique for data classification. It separates the dataset into classes using a hyper plane, aiming to maximize the margin between classes as much as possible. This involves creating partitions by drawing parallel lines. The margin represents the maximum distance between the nearest data points of different classes. The method selects the largest margin to minimize generalization error (Burman & Som 2019). Figure 2.3 illustrates the SVM diagram, showing the optimal separating hyper plane, maximum margin, and support vectors.

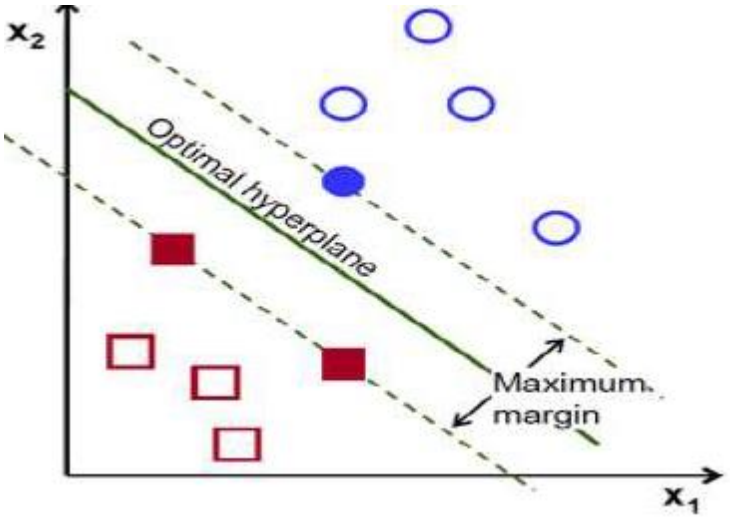


Figure 2.3: Hyper Plane Diagram

Adapted from (Oloruntoba, & Akinode, 2017).

$$w.x + b = 0 \tag{2.8}$$

Where

w is a vector normal to hyper plane

b is an offset.

If the value of $w.x + b > 0$ then it is a positive point otherwise it is a negative point.

The SVM algorithm works as follows:

Separable case is the one in which data can be perfectly linearly separated. Here, infinite numbers of boundaries are possible, and it selects the optimal hyper plane where in the boundary gives the maximum distance. Given a function.

$$f(y) = x.y + z \tag{2.9}$$

SVM divides the data points as

$f(y) > 0, iff\ y \in X$, and

$f(y) <= 0, iff\ y \in Z$

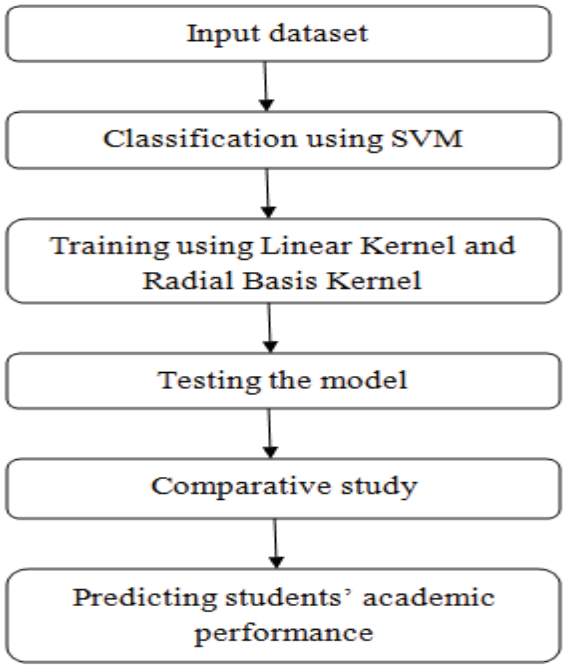


Figure 2.4: SVM in Predicting the Students' Performance

Source: (Burman, I., & Som, S. 2019).

When there is a clear margin of distinction between classes, SVM has the advantage of operating rather effectively. It works better in three-dimensional spaces. When the number of dimensions exceeds the number of samples, this method works well. It uses a small amount of memory. It is not ideal for huge data sets due to the following limitations. When the data set contains additional noise, such as overlapping target classes, it performs poorly. The SVM will underperform if the number of features for each data point exceeds the number of training data samples. There is no probabilistic justification for the classification because the support vector classifier works by placing data points above and below the classifying hyper plane (Ferjaoui et al., 2021).

2.3.4 K-Nearest Neighbour

K-Nearest Neighbor (K-NN) algorithm is a method for classifying objects based on learning data that is the closest distance to the object. K-NN is a supervised learning algorithm where the results of the new query instance are classified based on the majority of the categories in K-NN. The class that appears the most will be the class that results from the classification. The purpose of this algorithm is to classify new objects based on attributes and training samples. The K-Nearest Neighbor algorithm uses the neighbor classification as the predicted value of the new query instance. This algorithm is simple, works based on the shortest distance from the query instance to the training sample to determine its neighbors. (Noercholis & Zainuddin, 2020). Steps to calculate the K-Nearest Neighbor method with the closest distance (Euclidian) include:

i). *Determine the parameter k*

ii). Calculate the distance between data to be evaluated with all training

iii). Sort the distance formed

iv). Determine the closest distance to the order k

v). Pair the appropriate class

vi). Look for the number of classes from the nearest neighbor and sets the class as the data class to be evaluated. Equation 2.10 will be used to calculate the Euclidian distance between data to be evaluated with all training.

$$d_i = \sqrt{\sum_{i=1}^{p} x_{2i} - x_{1i})^2} \tag{2.10}$$

Where

x_{2i}= Sample data

x_{1i}= Test data or testing data

i = Data variable

d = Distance

Different distance functions like: Minkowski Distance, Manhattan Distance, Euclidean Distance are used in KNN algorithm. The Minkowski Distance for two points U $(u1, u2, \ldots un)$ and V $(v1;\ v2 \ldots\ vn)$ can be represented by the following equation, where q represents the order of the Minkowski Distance. (Zulfiker *et al*., 2020).

$$distance(U,V) = \sum_{i=1}^{n}(|u_i\ - v_i|\ )^q)^{1/q} \tag{2.11}$$

Manhattan Distance

With two N- dimensional points a= $(a1 \ldots . aN)$ and$b = [b1 \ldots . bN]$, the Manhattan distance d can be calculated as

$$d = |a1 - b1| + \cdots + |aN - bN| \tag{2.12}$$

KNN technique does not perform well with high-dimensional data because it becomes harder for the algorithm to calculate the distance in each dimension when the number of dimensions increases. Before applying the KNN algorithm to any dataset, feature scaling (standardization and normalization) is required. If we don't, KNN may make incorrect predictions. KNN is susceptible to noise in the dataset and is sensitive to missing values and outliers. We must manually fill in missing values and eliminate outliers. (Chong *et al*., 2019).

2.3.5 Logistic Regression

The classification algorithm logistic regression is used to determine the probability of event success and failure. When the dependent variable is binary (0/1, True/False, Yes/No), this method is utilized. It aids in the classification of data into discrete classes by examining the link between a set of labeled data. It takes the given dataset and learns a linear relationship before adding non-linearity in the form of the sigmoid function Yaacob et al., (2019) as shown in Figure2.3

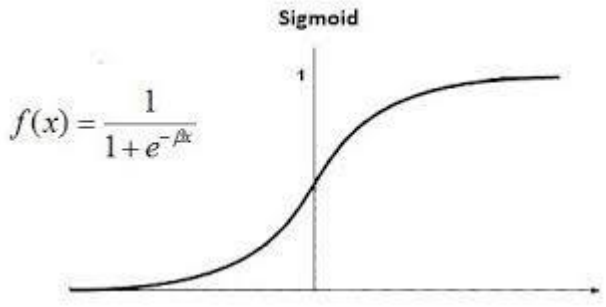


Figure 2.5: Sigmoid Function

Adapted from (Yaacob et al., 2019)

Binomial logistic regression is another name for logistic regression. It is based on the sigmoid function, with probability as the output and input ranging from -infinity to infinity. Logistic regression is more straightforward to apply, analyze, and train. It makes no assumptions about how classes are distributed in feature space. It's simple to extend to several classes (multinomial regression) and a probabilistic view of class predictions. It not only indicates the usefulness of a predictor (coefficient size), but also the direction of relationship (positive or negative). It classifies unfamiliar records fairly quickly. It performs well when the dataset is linearly separable and has good accuracy for many simple data sets. Model coefficients can be seen as indicators of feature relevance (Ghorbani & Ghousi, 2020)

Over-fitting is less likely with logistic regression, but it can happen in high-dimensional datasets. The following are its flaws: Logistic Regression should not be used if the number of observations is less than the number of features; otherwise, it may result in over-fitting. It establishes linear boundaries. It assumes that the dependent variable and the independent variables are linear. Only discrete functions may be predicted with it. As a result, the discrete number set is tied to the dependent variable of Logistic Regression. Because logistic regression has a linear decision surface, it cannot tackle non-linear issues.

The average or no multi-collinearity between independent variables is required for logistic regression. (Tillmanns & Krafft, 2021)

Revealed that complex associations are difficult to establish using logistic regression. This approach is readily outperformed by more powerful and compact algorithms such as Neural Networks. The independent and dependent variables are linearly connected in Linear Regression. However, in Logistic Regression, independent variables must be linearly connected to log chances($log\ (p/\ (1-p))$. When the outcome is a discrete variable, logistic regression is applied. For instance, determining who will win an election, whether a student will pass or fail an exam, whether a client will return, and whether an email is spam (Albreiki *et al*., 2021)

Equation 2.13 is used to calculate Logistic Regression.

$$f(x) = \frac{L}{1+e^{-k(x-x0)}} \tag{2.13}$$

Where

f(x)=output of the function
L=the curve's maximum value
k=logistic growth rate or steepness of the curve
x0=the x value of the sigmoid midpoint
x=real number

Regression analysis is used to determine the relationship between distinct variables. Linear Regression analysis can be used if the relationship is linear. However, this method cann't be used when the variables have a nonlinear connection. Then linear regression and logistic regression can be used. Linear Regression has been generalized into Logistic Regression (Zulfiker *et al*., 2020). Consider equation 2.14 for the Linear Regression:

$$y = a_0 + a_1 Z_1 + a_2 Z_2 + \cdots + a_n Z_n \tag{2.14}$$

The response variable is y, and the variables $Z_1;\ Z_2;\ Z_3$ and the predictor variables areZ_n. The logistic function can be obtained by applying the sigmoid function to the equation as shown in equation 2.15

$$l = \frac{1}{1+e^{-(a_0+a_1\,Z_1+a_2\,Z_2+\cdots+a_n\,Z_n)}} \tag{2.15}$$

2.3.6 K-Fold Cross Validation

Cross-validation is a statistical approach for assessing the efficacy of machine learning algorithms. There are several cross-validation methods, however the k-fold cross-validation method was chosen because it is popular and simple to grasp, and it also produces lower bias than the other cross-validation methods. (Sokkhey & Okazaki, 2020) The following is an overview of the k-fold cross validation process.

i) Shuffle the entire samples randomly
ii) Split samples into k sub folds
iii) In the split k sub folds:
iv) Take 1-fold as a holdout or test set
v) Take the remaining k -1folds as the training set
vi) Retain the evaluation score and discard the model
vii) Repeat the iteration until every single fold was treated as a testing set.
Finally, compute the average score of the recorded scores. In this research, 10-fold cross-validation was used to access the proposed algorithms.

2.3.7 Weighted Voting Classifier (WVC)

Weighted voting classifier is an approach for combining the outputs of different base classifiers as it is hard to identify a specific classification algorithm that gives the best accuracy on a certain data. Both homogeneous and heterogeneous models can be aggregated using the voting classifier. In the WVC, a weight or coefficient is assigned to

each base classifier which is proportional to the base classifier's individual accuracy (Zulfiker et al., 2020). Equation 2.15 can be used to calculate Weighted Voting Classier.

$$H = s1 * h1 + s2 * h2 + s3 * h3 \ldots sn * hn \tag{2.15}$$

Where

$h1$,$h2$,$h3$, --- *hn are the outputs of n-different classifiers respectively and*
$s1$,$s2$, $s3$--- *sn are the assigned weights to each classifier, respectively, then the final output H of the Weighted Voting Classifier can be represented by equation 2.15*

2.3.8 Ensemble Techniques

Ensemble classification operates on the concept that a group of experts can produce more accurate results than a single expert (Pandey &Taruna 2014). Ensemble modeling merges a group of classifiers to create a single composite model that yields improved accuracy. Research indicates that predictions from a composite model outperform those from a single model. In recent decades, the field of ensemble methods has gained significant attention. Numerous experimental studies by machine learning researchers have shown that combining the outputs of multiple classifiers can reduce generalization error.

Ensemble methods is the use of supervised learning algorithms that merge a set of classifiers into a meta-classifier by considering the voting or weighted voting of their predictions to create a final forecast. In essence, it involves averaging the outputs of several models to solve complex problems, aiming to achieve higher accuracy and greater generalization. Studies have shown that the accuracy can be improved by up to 30% with ensemble methods compared to using the best single model, which is why the approach is strongly recommended (Adejo & Connolly 2018).

2.3.8.1 Boosting

Boosting is a method to enhance classifier performance and reduce the error rate of weak models. Boosting focuses on instances in the dataset that were misclassified or generated errors in previous models, aiming to improve the accuracy of subsequent models. This technique often applies the same or related algorithms and employs majority voting to make decisions. Although boosting has demonstrated greater predictive accuracy compared to bagging, it faces a significant drawback in the form of over fitting. (Adejo & Connolly 2018)

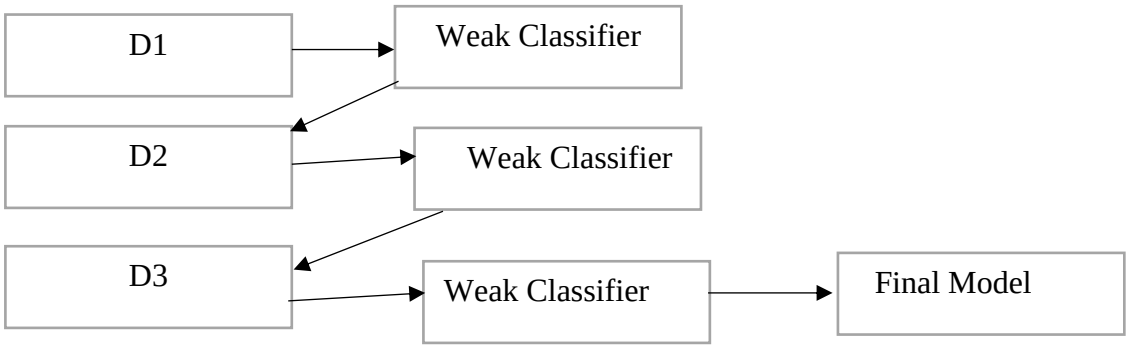


Figure 2.6: The General Boosting Procedure

Source: (Adejo & Connolly, 2018)

2.3.8.2 Bagging

Bagging is an independent ensemble-based method aimed at increasing the accuracy of unstable classifiers. The approach involves creating a composite classifier by combining the outputs of various learned classifiers into a single prediction. As shown in Figure 2.7, the bagging algorithm starts by resampling the original data into different training datasets, known as bootstraps (D1-Dn), where each bootstrap sample is the same size as the original training set. These bootstrap samples are then trained using different classifiers (C1-Cn). The results of individual classifiers are combined using a majority

voting process, where the class chosen by the highest number of classifiers becomes the ensemble's decision (Amrieh et al. 2016)

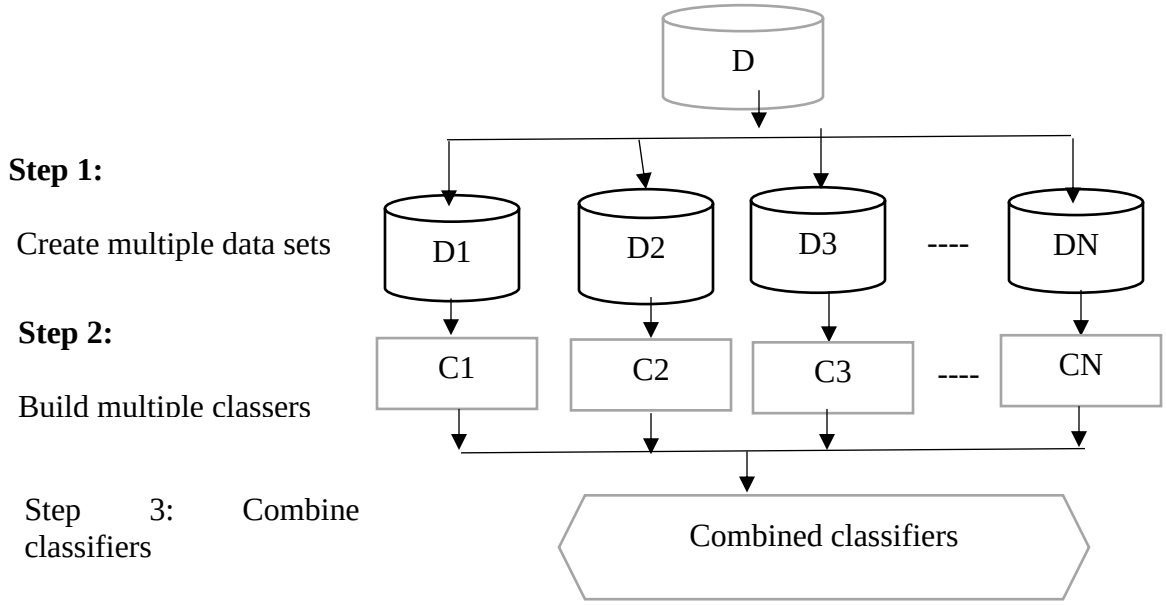


Figure 2.7: The General Bagging Procedure

Source: (Amrieh et al., 2016).

Stacking as an ensemble learning technique that combines multiple models in various ways using a meta-learner. Stacking works by constructing several distinct initial models that produce intermediate predictions, which then serve as inputs for the meta-classifier for the final prediction. This approach is more flexible and commonly used in ensembles than bagging or boosting because it can be applied to a wide variety of algorithms, making it a heterogeneous ensemble. Stacking helps reduce generalization error and improves performance accuracy. Figure 2.8 provides a simple diagrammatic representation of stacking.

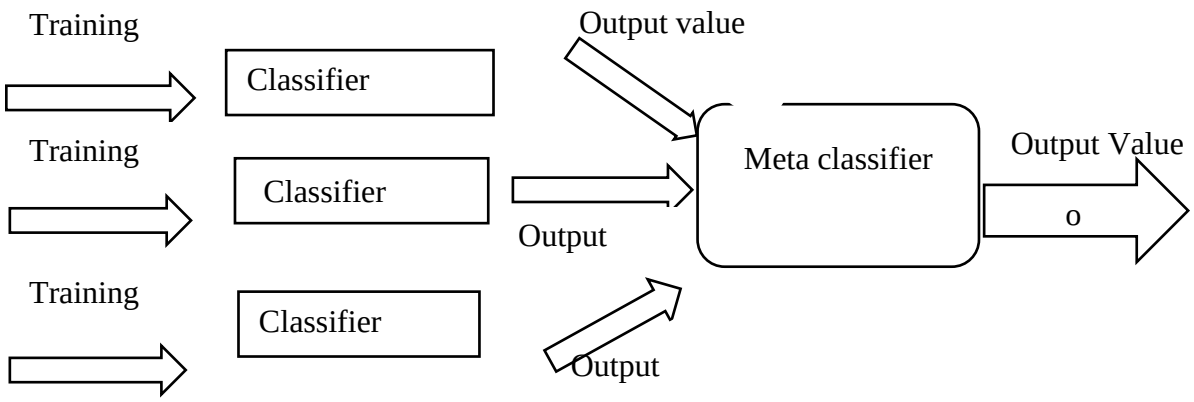


Figure 2.8: The General Stacking Procedure

Source: (Adejo & Connelly, 2018)

2.4 Feature Selection Methods in Education

Feature selection is a key pre-processing strategy in machine learning. This process is crucial in various research fields as it aids in making well-informed decisions (Adil et al. 2016)

2.4.1 Filter Based Methods

Feature selection is a method that uses a search approach to optimize an evaluation function, aiming to identify the best subset of features. The process consists of three stages: generating the feature set, measuring it, and testing it with a learning algorithm. Filter feature selection algorithms are efficient and quickly compute information from the features, basing their decisions on the measured information of the features. (Liu et al. 2017)

High data dimensionality can impact the efficiency and effectiveness of these methods. The researchers introduced a new process flow for filter-based feature selection utilizing a transformation technique. Normalization or transformation was applied before classification. The effects of normalization before feature selection were implemented and

evaluated. To provide a thorough analysis of power transformation effects, five different transformations were tested. Additionally, various feature selection methods were implemented and compared using the proposed process flow. The results indicated that, in comparison to existing process flows and feature selection methods. The ones that were proposed by Siddiai & Pak (2020) identified a more relevant set of features with greater efficiency and accuracy

Ren et al., (2020) conducted a study on three cases from the catchment attributes and meteorology for large-sample studies (CAMELS) data sets. Two termination criterion (TC) methods, the Hampel test and resampling, were comparatively analyzed. It was highlight that, there was no dominant FFS method that coupled with enELM or KNN., when resampling was applied to select a final model in the candidate combinations of the eight FFS methods and three regression models, PCI was the most favorable FFS method for the final model. Finally, the Hampel test TC was superior to the resampling TC in terms of stability and anti-over fitting. These findings have significant practical reference value for real-world monthly stream flow forecasting.

Zaffar et al. (2018) presented an analysis of the performance of filter feature selection algorithms and classification algorithms on two different student datasets. The results obtained from different Feature Selection (FS) algorithms and classifiers on two student datasets with different number of features helped researchers to find the best combinations of filter feature selection algorithms and classifiers. It is very necessary to put light on the relevancy of feature selection for student performance prediction, as the constructive educational strategies can be derived through the relevant set of features. The results of their study depicted that there was a 10% difference of prediction accuracies between the results of datasets with different number of features.

2.4.2 Wrapper Based Methods

Hui et al., (2017) found that feature selection plays a vital role in selecting the most representative feature subset for the machine learning algorithm. In contrast, the trade-off

relationship between capability when selecting the best feature subset and computational effort is inevitable in the wrapper-based feature selection (WFS) method. Improved WFS technique was used before integration with a support vector machine (SVM) model classifier as a complete fault diagnosis system for a rolling element bearing case study. The bearing vibration dataset made available by the Case Western Reserve University Bearing Data Centre was executed using the proposed WFS and its performance was analyzed and discussed. The results revealed that WFS secures the best feature subset with a lower computational effort by eliminating the redundancy of re-evaluation. WFS has therefore been found to be capable and efficient to carry out feature selection tasks

Babu & Vijayan (2016) explored a wrapper-based feature selection technique for semantic Information Retrieval (IR) and found that Semantics are the base for Information Retrieval's content description and query processing techniques. Semantic Similarity is about computing similarity between conceptually similar but lexically dissimilar terms. IR semantic features extraction based on word co-occurrence from web pages. Feature reduction was achieved through the use of wrapper-based feature selection technique comprising Latent Semantic Analysis (LSA) followed by Shuffled frog algorithm. The technique showed improved Precision and Recall when evaluated using Decision stump, Best First (BF) tree, and Random tree

Ma et al., (2017) revealed that new wrapper approach using polygon-based cross validation (CV) to overcome possible bias of object-based accuracy assessment for object-based classification was used. The new method is a two-step wrapper-based feature selection that involves the integration of: feature importance rank using gain ratio and feature subset evaluation using a polygon-based tenfold CV within a support vector machine (SVM) classifier. Several high-resolution images, including both unmanned aerial vehicle images and ISPRS (International Society for Photogrammetry and Remote Sensing) benchmark test data, were used to test the proposed method. Results show that, with the proposed polygon-based CV SVM wrapper, the mean overall accuracy is significantly higher than with an object-based CV SVM wrapper.

Shafiq et al., (2020) found that bijective soft set for effective feature selection to select effective features was used, and then a novel CorrACC feature selection metric approach. Afterward, a new feature selection algorithm named Corracc based on CorrACC was designed and developed which is based on wrapper technique to filter the features and select effective feature for a particular ML classifier by using ACC metric. For the evaluation four different ML classifiers were used on the BoT-IoT dataset. Experimental results obtained by the algorithms were promising and achieved more than 95% accuracy

According to Ghosh et al., (2020) a wrapper-filter combination of Ant colony optimization ACO, were introduced subset evaluation using a filter method instead of using a wrapper method to reduce computational complexity. A memory to keep the best ants and feature dimension-dependent pheromone update has also been used to perform FS in a multi-objective manner. Real-life datasets, taken from UCI Machine Learning repository and NIPS2003 FS challenge, using K-nearest neighbors and multi-layer perceptron classifiers were used to evaluate the model. The experimental outcomes were compared to some popular FS methods. The comparison of results clearly showed that the method outperforms most of the state-of-the-art algorithms used for FS. For measuring the robustness of the model, it has been additionally evaluated on facial emotion recognition and microarray datasets

2.4.3 Embedded Based Methods

In their study, Haoyue et al. (2019) compared different feature selection methods such as the weighted Gini index (WGI), Chi-squared (Chi2), F-statistic, and Gini index. The study found that F-statistic and Chi2 methods performed particularly well, especially when selecting a small number of features. Interestingly, as more features were selected, the likelihood of achieving optimal performance increased. The evaluation, which utilized metrics like the area under the receiver operating characteristic curve (ROC AUC) and F-measure, showed that ROC AUC remained consistently high even with a limited selection of features, with only minor variations when more features were included. However, for

the F-measure to achieve excellent performance, at least 20% of the features needed to be selected.

Maldonado & Lopez (2018) introduced an innovative feature selection technique aimed at addressing the challenges posed by class imbalance and high dimensionality in machine learning. The embedded method incorporates scaling factors to penalize the size of the feature set. It was implemented with two support vector machine (SVM) formulations: Cost-Sensitive SVM and Support Vector Data Description, both tailored to handle class imbalance effectively. The proposed concave formulations were solved using Quasi-Newton updates and Armijo line search. Experiments conducted on 12 highly imbalanced microarray datasets, utilizing both linear and Gaussian kernels, revealed that the proposed approach consistently outperformed established feature selection methods in terms of average predictive performance.

Feature selection aims to improve the effectiveness of data analysis by removing irrelevant features from datasets. However, selecting the most suitable algorithm for specific datasets can be challenging due to the varying criteria used by different algorithms. Chen et al., (2020) found that ensemble methods, which combine multiple feature selection outcomes, can overcome the limitations of individual methods. While existing literature categorizes feature selection algorithms into filter, wrapper, or embedded techniques, there has been limited exploration into combining these approaches to create ensemble methods. Experimental results suggest that combining filter techniques like principal component analysis with wrapper techniques such as genetic algorithms using the union method leads to superior outcomes, achieving high classification accuracy and significant reductions in the number of features

2.4.4 Hybrid Feature Selection

Ramaswami, et al., (2020) predicted students' academic performance with the LMS data from an online training course using various machine learning algorithms. The study further highlights that selection of features by using feature selection methods will enable

early prediction ofstudent academic performance. Early predictions can benefit those students who are considered to be at risk of failing a course with targeted early interventions to help improve their performance throughout the course.

Farissi and Dahlan (2019) addressed the challenge of high dimensional datasets impacting the accuracy of predicting student academic performance. They proposed the Genetic Algorithm based Feature Selection (GAFS) combined with a selected single classifier to enhance prediction accuracy. Using a Kaggle dataset, the study conducted two phases of experiments: one without GAFS and one with GAFS. The results demonstrated that the GAFS significantly improved the accuracy of student academic performance prediction compared to current techniques

Saifudin & Desyani (2020) found that management of academic performance is very essential to ensure that learners can complete their education on time. There have been many suggested applications of machine learning algorithms to forecast students' academic performance. Prediction is done by analyzing a dataset of historical academic of the student's grade. The dataset which analyzed has many variables (features), this can increase complexity and decrease model performance because maybe not all features are relevant. They proposed to implement the forward selection algorithm to select features that can improve model performance. The result showed that the performance of predictive models of students' academic scores can improve with the application of feature selection.

Saifudin & Desyani (2020) emphasized the importance of managing academic performance to ensure timely completion of education. They explored various applications of machine learning algorithms for predicting students' academic performance by analyzing historical academic grade datasets. Recognizing the complexity arising from numerous variables (features) in the dataset, they suggested implementing the forward selection algorithm to enhance model performance by selecting relevant features. The outcomes indicated that predictive models of students' academic scores could be enhanced through feature selection techniques.

Zaffar et al. (2017) conducted an evaluation and analysis of various feature selection algorithms. Their findings on student datasets revealed that existing feature selection algorithms in the Weka tool did not significantly alter performance. However, among these methods, principal components coupled with the Random Forest classifier yielded superior results. The study also noted that the MLP classifier outperformed other classifiers marginally on student datasets. It underscored the importance of subtle parameter tuning for enhancing the performance of feature selection methods. Additionally, the authors suggested exploring more feature selection techniques and their combinations, as well as using student datasets of varying sizes for future evaluations.

Satyanarayana & Nuckowski (2016) conducted research on data mining aimed at enhancing the prediction of student academic performance using ensemble classifiers. Their study revealed that filtering student data could substantially enhance predictive accuracy. By comparing single filters with ensemble filters, they demonstrated that ensemble filters were more effective in identifying and eliminating noisy instances. Furthermore, they highlighted that both majority and consensus voting methods resulted in improvements. The ensemble technique proved effective across two different settings: high school data and first-year college data. While decision trees, random forest, and naïve Bayes were utilized in their study, they suggested that other base classifier models could also be employed.

2.5 General Application of the Classifiers in Prediction

Machine learning techniques have been frequently used for classification and prediction in different fields (Nagahisarchoghaei et al., 2020)

2.5.1 Applications of Naive Bayes (NB) Classifier in prediction

2.5.1.1 Sentiment Analysis

A study by Adam et al. (2021) created a system that analyzed movie reviews for sentiment and visualizes the results. IMD movie reviews were used to create the dataset. The data

set used in this study had 50,000 instances in with two columns containing reviews and sentiments. The Bag of Words (BoW) and TF-IDF modeling with Nave Bayes classifiers were examined as two types of features. The authors reported an accuracy of 89% which was an indication that the Naïve Bayes could be used for sentiment analysis in reviewing movies.

According to a study by Kumar et al (2018) Naive Bayes algorithms are used for performing the sentimental analysis for differentiating the positive and negative reviews of the patients. This framework helps the medical institute for obtaining better knowledge for choosing which drug getting more benefit and give minimum side effects.

According to Tika et al. (2020) based on the results of the analysis conducted on the sentiment of online transportation services using the Naive Bayes method, researchers can draw some conclusions that the Naive Bayes Method is quite good in classifying data mining or text mining. This is because the algorithm can produce a fairly high accuracy value of 81.00%, which means that all the comments on the Instagram page with the NBC method can be accurately classified whether the comment is negative or positive. The results of this study can be used as recommendations to improve the performance of online transportation.

In their study, Bohra et al. (2017) described a platform-independent system that offers users medical guidance through an interface. This system utilizes the Naive Bayes algorithm to predict diseases based on symptoms and provides recommendations on daily hygiene, diet, and routines for healthy individuals to follow. Users can also connect with nearby specialist doctors for easy medical treatment and diagnosis. This system aims to offer instant health advice and simplify the process of accessing medical care.

Naïve Bayes algorithm and R tool have been used for prediction and visualization. The goal is to develop a cost-effective and easily accessible healthcare system that can benefit the medical practitioners to combat the prolonged procedures of diagnosis and faster retrieval of results (M & Sagar 2019)

Artificial intelligence has been used with Naive Bayes classification and random forest classification algorithm to classify many disease datasets like diabetes, heart disease, and cancer to check whether the patient is affected by that disease or not. A performance analysis of the disease data for both algorithms is calculated and compared. The results of the simulations show the effectiveness of the classification techniques on a dataset, as well as the nature and complexity of the dataset used (Jackins et al., 2021)

A study that was done by Amos Okutse (2019) it revealed that Naïve Bayesian probabilistic classifier was used for modeling the quality of patient care in a healthcare setting. Using secondary data, we assess the effectiveness of the Naïve Bayes machine learning classifier in modeling the probability of poor care. Exploratory data analytics are performed and visualized using bar graphs, density plots, and heat maps. The authors evaluated the performance of this classifier using confusion matrices, specificity, and sensitivity indices. R software is used for statistical programming. The Naïve Bayes classifier yielded an accuracy of 77%; 95%CI (0.5774, 0.9138). The classifier had sensitivity and specificity values of 0.80 and 0.71, respectively; denoting the chance of poor care being classed as poor care when it is poor care and the likelihood of poor care being reported as quality care, respectively. The proportion of poor care was 74%. The implementation of quality assessment systems in health is likely to drive efficiency in terms of patient care.

According to a study by Paas & Groot (2017) it was found that Naïve Bayesian (NB) classification as a method was used to allocate farms to types by using only a few variables, thus allowing the addition of new entries to a typology. We show for two example datasets that the performance of NB classification is already acceptable when 50% of the original survey dataset to construct the typology is used for training the NB classifier. For our datasets, the performance of Naïve Bayesian classification was improved when probabilities for observations to belong to multiple types were used, requiring a sample size of 30% of the survey dataset. Based on the results in this paper, we argue that NB classification is a powerful and promising statistical approach to increase the adaptability and usability of farm typologies.

From the research that was done by Fithri and Latifah (2018) it revealed that Naïve Bayes method was used to predict rice plants by using several criteria 731 used in determining the yield of rice production is the sex of a farmer, level of education, type of work, position, monthly income and the age level of farmers. But after the prediction is done using the Naïve Bayes method, the most dominant criteria in predicting the yield of rice crops is the level of education and occupation at work.

In the study by Kanchana and Sujatha (2016), Naive Bayes was used to classify real and discrete data by leveraging probability. Formal Concept Analysis was applied to map the results of the Naive Bayes algorithm to the given data. The study utilized the Naive Bayes algorithm to classify agricultural datasets and employed Formal Concept Analysis to map the results to farmer data. This approach provides recommendations for farmers based on the analysis.

According to Mohanapriya & Balasubramani (2019) Naïve Bayes classifier was used to detect unhealthy regions of plant leaves and also to classify them. Initially the leaf images are collected, color converted, segmented, feature extracted and finally the plant disease is classified. The accuracy of result obtained is about 97%.

According to the study that was done by Wang and Kim (2016) Naive Bayes (NB) classifier model was used for predicting congestion and incident in urban road networks. The study considered congestion or incident as a target variable and applies a NB model to classify its state (i.e., occur vs. not occur). Predictor variables or features considered in the study include network environment variables (time of day, day of week, and weather) and traffic condition variables (speed on bottlenecks). The study developed a data-driven approach for building and validating NB models. The models were trained and tested using actual traffic, incident, and weather data collected from Brisbane, Australia in 2014. The validation results showed that the proposed models can successfully predict congestion and incident occurrence with a desired level of accuracy.

Suh & Jeong (2022) conducted modeling and tests to show how Naïve Bayes classifiers learned in the form of supervised learning can help the route reorganization work. Results from a local governments' actual route reorganization study were used to train and test the proposed machine learning classification model. As the main contribution of the study, a prediction model was developed to support shortening decision-making for each route, using machine learning algorithms and actual route reorganization research case data. Results verified that such an automatic classifier, or initial route decision proposal software, can provide intuitive support in actual route reorganization research.

2.5.2 Applications of Support Vector Machine (SVM) Classifier in Prediction

According to Kok et al. (2021), SVM (Support Vector Machine) was compared with other models in precision agriculture (PA) to assess its interactions with variables and model performance, as well as its strengths and weaknesses. The review considered six machine learning (ML) applications in PA and confirmed features that can enhance the model in general (e.g., feature selection) or for specific applications (e.g., phenology). SVM was found to outperform most models, although its comparison with Random Forest (RF) was inconclusive, and it was found to be less effective than Deep Learning (DL). The review highlights ongoing efforts to improve SVM performance in PA through its integration with DL, suggesting that this approach is likely to become a future trend in ML model development for modern precision agriculture.

Jithender et al. (2019) used SVM (Support Vector Machine) learning techniques to classify agricultural produce based on various factors such as size, texture, shape, variety, color, and quality. For wheat, SVM classified grains based on quality aspects like color and texture, resulting in an overall accuracy of 94.45%, while the Naive Bayes classifier achieved an accuracy of 92.60%. Wheat was also classified using SVM and Artificial Neural Network (ANN), where SVM demonstrated higher accuracy compared to ANN SVM can be easily adapted for general inspection and segmentation of images in grains, fruits, vegetables, and other agricultural products. The study concluded that SVM offers superior classification performance over other machine learning tools.

According to Kumar et al., (2019) the efficacy of SVM in predicting the yield of rice crop was demonstrated. SVM-based classification models have been developed for the forecasting of rice yield in India. SVM classification models have been tested using 3-fold, 4-fold and 5-fold cross validation methods, one against-one multi classification method. The dataset encompasses the rice yield in India from the year 1950 to 2014. The best prediction accuracy for 4-year relative average increase has been obtained as 75.06% using 4-fold cross validation method. This is not a very high accuracy; however, it is believed that this can be improved by redefining training patterns, considering other KF. The learning parameters can also be optimized using particle swarm optimization or other related techniques. This work can also be explored further for yield prediction of other crops.

According to Battineni et al. (2019) the usage of support vector machine (SVM) in the prediction of dementia and validate its performance through statistical analysis. Data was obtained from the Open Access Series of Imaging Studies (OASIS-2) longitudinal collection of 150 subjects of 373 MRI data. Results provide evidence that better performance values for dementia prediction are achieved by low gamma (1.0E-4) and high regularized (C = 100) values. The proposed approach is shown to achieve accuracy and precision of 68.75% and 64.18%.

From the study that was done by Augusstine & Samy (2018) an innovative health monitoring system using the internet of things for accessing the patient's medical parameters in the local and remote area was done. The goal of this study was to transmit an emergency message to caretaker when the health condition goes critical. A cloud server records the data from the temperature sensor, and heartbeat sensor which was connected to the patient; the data was analyzed using support vector machine learning algorithms to detect the abnormal conditions, issues an emergency message to the caretaker of the patient through a mobile application, and sends an alert message to the nearest hospital.

Junyou et al. (2018) applied a support vector machine for predicting bus travel time to distribute many random factors such as weather, traffic congestion, and passenger flows.

Moreover, suggested that SVM can be used to analyze predictive objects and predict unknown data or new phenomena.

2.5.3 Applications of K- Nearest Neighbour (KNN) Classifier in Prediction

Xin & Chen (2016) developed a dynamic model to predict bus dwell time at downstream stops. The research also intends to test the proposed model using real-world data. This model is based on k-Nearest Neighbour (KNN) algorithm using history and current data collected by GPS (Global Positon System) fixed on buses. In the research, the data of buses of No.B1 line of Changzhou in China is used. In the test with real-world data, the proposed bus dwell time prediction model performed effectively both on accuracy and calculating speed.

Kumar et al. (2019) conducted a study to explore the use of Intelligent Transportation Systems (ITS) to make public transportation systems more attractive by providing timely and accurate travel time information of transit vehicles. However, for such systems to be successful, the prediction should be accurate, which ultimately depends on the prediction method as well as the input data used. In the present study, to identify significant inputs, a data mining technique, namely k-NN classifying algorithm is used. It is based on the similarity in pattern between the input and historic data. These identified inputs are then used for predicting the travel time using a model-based recursive estimation scheme, based on Kalman filtering. The performance is evaluated and compared with methods based on static inputs, to highlight the improved prediction accuracy.

In their study, Pavithra & Vadivel (2021) used the m-KNN (Modified k Nearest Neighbour) algorithm as a classification model and applied PCA (Principal Component Analysis) for feature extraction. This approach focuses on road traffic prediction, a crucial aspect of modern smart transportation systems. The proposed method enhances the performance of the m-KNN algorithm by preprocessing the training data and introducing a new attribute called "Validity" to the training samples. This attribute provides additional

information about the stability and robustness of the training data samples in the feature space

By incorporating the weighted KNN method, which uses validity as a multiplication factor, the classification becomes more robust compared to the simple KNN method. The traffic detection system allows for real-time monitoring of various road network areas, enabling the detection of traffic events almost instantly, often ahead of online traffic news websites.

Karthikeya et al. (2020) implemented a system to learn about crops and agriculture and find an efficient way of harvesting. The study focused on the agricultural datasets obtained from various portals belonging to some districts of Karnataka State. Datasets ordered in well-structured manner. K-NN algorithm was used for the prediction model and crop yield prediction and its accuracy is obtained.

2.5.4 Applications of Multi-Layer Perceptron (MLP) Classifier in Prediction

El Bilali (2019) applied the multi-level perspective (MLP) to examine agro-food sustainability transitions, particularly focusing on the understanding, conceptualization, and operationalization of niches, regimes, and landscapes. The study found that transition pathways within the MLP often fall short in addressing the unique aspects of the agro-food sector. Additionally, the impacts of transitions and the sustainability of niches are rarely considered, which means transitions are not well-analyzed in this field. Research on agro-food transitions benefits from the MLP's generalizability but suffers from inadequate empirical operationalization of niche, regime, and landscape concepts. To address this, an integrative approach is needed for conceptualizing and operationalizing MLP elements to better manage the complexities of sustainability transition processes and the specific needs of the agro-food system.

Wang et al. (2021) developed an improved Multilayer Perceptron (MLP) approach to predict sugar yield production in IoT agriculture. Their experimental results demonstrated

that the proposed MLP algorithm achieved a maximum accuracy of 99%, precision of 95%, and recall of 96%. Additionally, it had a minimum Mean Absolute Error (MAE) of 0.04% and a Root Mean Square Error (RMSE) of 0.006% for sugarcane yield detection in IoT agriculture.

Utku and Kaya (2022) created a real-world forecasting model based on the demand for transfer passengers in Istanbul, Turkey's largest and most developed city. They forecasted the number of transfer passengers using well-known machine learning methods such as k-Nearest Neighbors (kNN), Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Multilayer Perceptron (MLP). The dataset consisted of hourly passenger transfer counts collected at two public transportation transfer stations in Istanbul during January 2020. They rigorously evaluated each model's experimental data using parameters such as MSE, RMSE, MAE, and R2. The experimental results indicated that MLP outperformed other machine learning algorithms on most transportation lines.

A study by Bikku (2020) focused on supervised learning methods and their capability to find hidden patterns in the real historical medical data. The objective is to predict future risk with a certain probability using Multi-layer perceptron (MLP) method. In the proposed work, MLP based on data classification technique is used for accurate classification and risk analysis of medical data. The proposed method is compared with traditional classification methods and the results show that the proposed method is better than the traditional methods.

Amin & Ali (2017) research explained the utilization of Multilayer Perceptron (MLP) with back propagation (a supervised learning algorithm) in the determination of medical operation methods. They provided this with accumulating 80 pregnant women information. The results showed that Multilayer Perceptron (MLP) designed for this case study generates correct predictions for 95% test cases.

2.5.5 Applications of Logistic Regression (LR) Classifier in Prediction

In the study by Borucka, (2020) logistic regression was applied to analyze a distribution and trade company that supplies automotive spare parts, with a focus on local car repair shops as the most profitable group of customers. The quality of service was assessed based on delivery time, using a dichotomous predictor that classified deliveries as either late or on-time. Regressors that had statistically significant influence and could be modified were chosen from the available options. The research identified which factors impacted the dependent variable and to what extent, enabling the modification of strategy and the implementation of new solutions to increase customer satisfaction.

Phiophuead &Kunsuwan (2019) developed three models using various analytical factors based on survey data from a sample group of people in the Mae Pong watershed, Laplae district, Uttaradit province, Thailand. They found that factors influencing travel mode choice in all three models included gender, household size, families with young children, education level, car ownership, disaster experience, awareness of shelter locations, safety during evacuation, speed in reaching the destination, convenience of vehicle access, proportional family management for evacuation, ease of evacuation procedures, and the difference between travel time and walking time to the assembly point. The models achieved prediction accuracies of 78.40%, 73.46%, and 75.30% respectively.

Educational Data Mining (EDM) is a growing field focused on creating methods for examining the distinct types of data that originate from educational environments. These methods aim to enhance the understanding of students and their learning contexts. EDM encompasses techniques and tools that automatically extract insights from large collections of data produced by people's learning activities in educational settings. These educational databases hold valuable hidden information with numerous significant factors related to students' learning experiences. (Kumari et al. 2014)

Manjarres et al. (2018) conducted research on data mining techniques applied in educational settings and recognized that Educational Data Mining (EDM) is an emerging field focused on developing methods to analyze large volumes of data from educational

environments. This helps in gaining a better understanding of students' behavior, interests, and academic outcomes. Over recent years, there has been a growing body of work in this area, utilizing various data mining techniques to address diverse educational challenges. The field of data mining is also closely related to other disciplines, as illustrated in Figure 2.9.

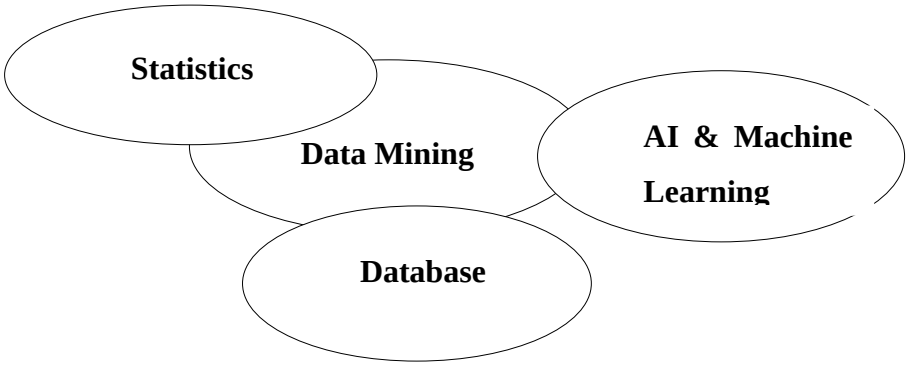


Figure 2.9: Relationship between Data Mining and Other Disciplines

Source: Manjarres et al. (2018)

2.6.1 Factors Influencing Students' Performance

Al Husaini & Shukor (2022) found that several factors significantly affect students' academic performance, including low entry grades, family support, accommodation, student gender, previous assessment grade, student internal assessment grade, GPA, and e-learning activity. Similarly, Mushtaq and Khan (2012) suggest that positive influences such as communication, adequate learning facilities, and proper guidance contribute positively to student performance, while family stress has a detrimental effect.

According to the study conducted by Qureshi et al. (2023), their findings, analyzed through structural equation modeling (SEM), indicate that social factors such as interaction with peers and teachers, social presence, and the use of social media have a

positive effect on active collaborative learning and student involvement, thereby enhancing their learning performance. The research also demonstrates the utilization of double mediation. With the increasing prevalence of online learning in education, it has been concluded that fostering collaborative learning and engagement through social factors enhances students' learning activities. Consequently, the integration of these factors should be encouraged in teaching and learning within higher educational institutions to positively influence students' academic development.

The extensive investigation carried out by Batool et al. (2023) highlighted that students' academic histories and demographic variables are the most dependable indicators of their performance. The study emphasizes that irrelevant data points in the dataset not only reduce predictive accuracy but also prolong the time it takes to process models. As a result, close to half of the studies utilize methods for selecting relevant features before developing prediction models

2.7 Student Performance Prediction

Students' academic achievement has been accurately predicted using EDM. The most prevalent machine learning algorithms used in EDM are decision tree (DT) and random forest (RF). Hussain et al., (2018) employed DT to predict student academic success, while Heuer & Breiter, (2018) used RF. In their study, Wasif et al., (2019) support vector machines (SVMs) were used to detect the student success rate by considering demographics and social factors. Machine learning models such as DT, RF, LR, and SVM were employed to predict student academic achievement using daily activities as a feature. Students' academic achievement has been accurately predicted using EDM. The most prevalent machine learning algorithms used in EDM are DT and random forest (RF).

For student performance prediction, White hill et al., (2017) introduced a technique called the combinational incremental ensemble of classifiers. Three classifiers are merged in the proposed technique, with each classifier calculating the prediction output. The total final prediction is chosen using a voting mechanism. When a fresh sample arrives, each

classifier predicts the outcome, which is useful for continuously created data. The voting system is used to choose the final prediction. Hellenic Open University provided the training data for this study. The dataset contains 1347 instances of writing assignment marks, each with four attributes and four features for written assignment scores.

Strecht et al. (2015) conducted a study to predict student outcomes, specifically identifying which students would pass and which would fail. They used a classification model to categorize students and a regression model to predict student grades. The classifiers utilized included KNN, classification and regression trees, AdaBoost, RF, NB, and SVM to classify 5,779 records containing varied attributes such as age, sex, scholarships, and student status. For the regression tasks, they used methods such as RF, AdaBoost, SVM, classification and regression trees, and plain least squares. The researchers compared their models using F1 scores for classification and RMSE for regression. Various combinations of machine learning techniques have been employed in educational data mining (EDM) research.

Haiyang et al. (2018) investigated the use of time-dependent variables to predict student performance in online learning. They presented an early warning system for predicting students' risky online learning performance. They emphasized the importance of time-dependent variables in determining student performance in Learning Management Systems (LMS). Their goals were to i) investigate data mining techniques for early warning, (ii) determine the effects of time-dependent factors, and (iii) choose a data mining technique with greater predictive potential. They examined the effectiveness of three machine learning classification models, namely "C4.5 Classification and Regression Tree (CART), Logistic Regression (LGR), and Adaptive Boosting (AdaBoost)," using data from 330 students in online courses from the LMS. Each instance in the dataset had ten features, and the classifiers' performance was measured in terms of accuracy, type I, and type II errors. The CART algorithm surpasses the competition, achieving accuracy of over 95%.

Hussain et al. (2018) looked at prediction models to identify students at risk in a course that uses standard-based marking. Furthermore, they used feature selection methods with data from the first-year engineering course at a Midwestern US university in 2013 and 2014 to reduce the size of the feature space. Class attendance grades, quizzes grades, assignments, team involvement, project milestones, mathematical modeling activity test, and examination scores were all included in the student performance dataset. LR, SVM, DT, MLP, NB, and KNN were among the six machine learning classifiers examined. Different accuracy measures were used to evaluate these classifiers, including overall accuracy, accuracy for pass students, and accuracy for failed students.

2.7.1 SVM Approach to Performance Prediction

Kadambande et al. (2017) researched on predicting student's performance system using SVM and concluded that Support Vector Machines are supervised learning models with associated learning algorithms that analyze and survey data used for classification and regression. It was simply a co-ordinate of individual observation. It was very crucial for cases where very high predictive power was required. Such algorithms are harder to visualize because of the more complexity in the formulation.

Smith et al. (2018) conducted a study on predicting student academic performance using SVM algorithms. They collected data on various student attributes, such as socioeconomic status, previous academic performance, and demographic information. The SVM model achieved high accuracy in predicting student outcomes, demonstrating the potential of SVM algorithms in this context. However, the study did not focus on feature selection or explore the impact of different SVM parameters on prediction accuracy.

Jones and Brown (2019) investigated the use of SVM algorithms to predict student success in online learning environments. They utilized a range of features, including student engagement, interaction patterns, and course performance data. The SVM model achieved good accuracy in predicting student success, but the study did not delve into the specific

feature selection process or explore the potential impact of different SVM kernels on prediction performance.

2.7.2 MLP Approach to Performance Prediction

Awad & Ewais (2018) performed a prediction of general high school exam result level using multilayer perceptron neural networks MLPNN and confirms the applicability of the MLPNN in predicting the general high school exam levels. This type of Neural Network can model complex systems and obtain small errors in training and testing prediction was highlighted. The proposed model became a useful tool for the universities; it delivered a percentage of each general high school exam level that allow the universities to plan the acceptance rate and level with high better precision. The proposed model of the neural network obtained excellent results from the testing data. The dataset was composed of 110-time series data with 10 data levels for each year. The most two selected branches of the general high school exam are applied to be predicted using the proposed model which is scientific, and literary. The Accuracy of the proposed model was very high 99.99% with a testing error of ± 0.0005

Agrawal and Mavani (2015) studied student performance prediction using artificial neural networks and found that academic performance is mainly influenced by students' past performances. Their research confirmed the significant impact of past achievements on current performance. Additionally, they discovered that the performance of neural networks improves as the size of the dataset increases.

Urhayati et al. (2018) predicted graduation system using an Artificial Neural Network. In their study, an Educational Data Mining (EDM) system was developed to predict students' graduation at universities based on five parameters: gender, year of entry, GPA semester 1, GPA semester 2, and GPA semester 3. Artificial Neural Network (ANN) was selected as its classifier, where the ANN model used was Multi-Layer Perceptron (MLP). The system was capable of producing good performance with an average accuracy of 94.8%, and the average value of precision and recalls respectively 94.2% and 96.2%. This showed

the ability of ANN (in this case MLP) as a reliable classifier of incomplete data. The non-linear calculation process allows ANN (Artificial Neural Network) to classify complex data.

Figueira (2016) predicted grades by Principal Component Analysis and experimented with a new approach to predict grading and prevent students from failures at early stages. The model was described by implementing within Moodle, but it can be generalized to any system with a fairly reasonable log system. Three types of features were discussed and extracted from the logs to characterize the interaction behavior with the platform, presented an approach to data mining Moodle logs using principal component analysis to detect the relevant features to make predictions.

Prabha & Shanavas (2015) performed classification algorithm on student data and studied how different EDM algorithms for classification works on student records generated from an e-learning domain. The performance of the selected algorithms was analyzed on various criteria. They used students' data from the database of sixth-grade students of a school who worked in Math's tutor. The training set contains 120 records each for a student and 10-fold cross- validations were used for classification. Though Multi-layer Perceptron and J48 has 100% accuracy J48 takes a very minimum time 0.03 seconds compared with 27.19 seconds taken by the other. The multilayer perceptron has 0.0048 MAE and 0.0076 RMSE whereas J48 has 0.

Adewale et al. (2018) researched on predictive modeling and analysis of academic performance of secondary school students using the Artificial Neural Network approach and concluded that, ANN model achieved an accuracy of 90%, which shows the potential efficacy of ANN as a predictive model, a clustering instrument and a selection criterion for candidates seeking admission into a university. Despite the high-level prediction accuracy of ANN in nonlinear phenomena, however, the model does not easily allow the identification of how predictor variables are related to one another in the explanation of the academic outcome.

Aybek & Okur (2018) researched on Predicting Achievement with Artificial Neural Networks: The Case of Anadolu University Open Education System and found that as a result of the analyses, it was found out that networks established through MLPs make more exact predictions. In the prediction of the final exam scores, it was determined that there is a low level of correlation between the actual scores and predicted scores.

Arunachalam and Velmurugan (2018) conducted research on assessing student performance using the Evolutionary Artificial Neural Network Algorithm. Their study found that, based on the simulation results, the optimized Evolutionary Artificial Neural Network provided more precise outcomes compared to other methods that were evaluated for comparison.

2.7.3 Naïve Bayes Approach to Student Performance Prediction

Amra and Maghari (2017) applied KNN and Naïve Bayes algorithms to an educational dataset from secondary schools in the Gaza Strip for the year 2015, collected from the Ministry of Education. The aim of this classification was to assist the ministry in enhancing performance by enabling early prediction of student outcomes. This approach also aids teachers in making appropriate evaluations to improve student learning. The study's experimental results demonstrated that Naïve Bayes outperformed KNN, achieving the highest accuracy rate of 93.6%.

Perez and Perez (2021) utilized the Naïve Bayes classifier to predict students' program completion, using a 70:30 ratio for training and testing data distribution. Correlation analysis was performed to determine the influence of individual attributes on the label attribute, resulting in the selection of four key predictor variables out of ten possible options. These significant attributes impacting program completion were, in order of importance: parents' monthly income, the educational level of both parents, and students' High School GPA. The selected attributes were divided into 70% training data (447 records) and 30% testing data (191 records). Predictions indicated that 84 out of 191 students (44%) would complete the program, while 107 out of 191 students (56%) were

forecasted to not complete the program. The model achieved an accuracy of 84%, with a class precision of 80.46% and a class recall of 83.33% in the testing dataset.

2.7.4 KNN Approach to Student Academic Performance

Jawthari and Stoffová (2021) proposed two new similarity measures to customize the kNN algorithm for mixed data types, focusing particularly on nominal variables. The method enhances the algorithm's handling of outliers by employing different voting rules. Additionally, they developed a distance function that allows classification decisions to be made without converting nominal variables into numeric form. The researchers verified the classifier by experimenting with an educational dataset, comparing the results to those obtained using a standard kNN algorithm with one-hot encoded nominal attributes. Their experiments showed that utilizing Jaccard distance consistently surpassed the standard kNN by 14%, demonstrating the enhanced kNN algorithm's strong accuracy performance.

Maghari (2018) introduced a classification model for predicting student performance based on the marks of two subjects. The study employed modified KNN classifiers, including Cosine KNN, Cubic KNN, and Weighted KNN, to classify students' grades. The dataset was collected from secondary schools in the Gaza Strip and contained grade records for students in the 11th scientific branch. The classification experiment revealed that each algorithm produced different accuracy values and predictions due to the distance metrics used. Changing variables such as distance metrics, distance weight, and the number of neighbors improved the algorithm's efficiency by increasing classification accuracy. Early prediction of student grades allows principals to make decisions to help schools identify students with low academic achievement and provide support for them.

Nugroho et al. (2020) employed the KNN algorithm to classify and monitor teaching and learning activities by processing data on student complaints and evaluating previous learning outcomes. The study used the K-Nearest Neighbor classification algorithm to predict graduation rates based on complaints data, with groups of $k = 1$, $k = 2$, and $k = 3$ using the smallest value possible. Experiments with the KNN method demonstrated

notable accuracy. The research concluded that a reduction in complaints from a student could help minimize the level of student non-graduates at the university, ultimately leading to improved accreditation outcomes.

2.7.5 LR Approach to Student Performance

Urrutia et al. (2016) proposed a logistic regression model to identify the variables that predict academic performance among first-year medical students. The study's findings have practical applications for creating in-person or online preparatory courses tailored for high school students who plan to enter medical school. It also highlights the need for interventions during the first year of medical school. To address academic shortcomings, didactic materials and training courses on teaching strategies should be implemented to help students enhance their self-perception. The model from the study can be applied across all national medical schools to identify students at risk of academic failure and provide effective strategies to ensure they successfully complete their medical training.

Sule & Saporu (2015) used a logistic regression model to examine the factors affecting students' performance in the MTH101 (Element of Calculus) course. The study used data from the grades of students at the 200-400 levels, collected from the department's examination records and a questionnaire distributed to the students. The data analysis revealed that three main factors significantly impact academic performance in the MTH101 course: GPA (students' academic performance), the perceived challenge of the course (students' attitudes toward the course), and the connection of course concepts to real-world experiences (students' motivation). The study recommended focusing intervention strategies on enhancing academic performance, addressing course-related attitudinal issues, and providing adequate motivation to improve the course.

Patrick Soule (2017) used multiple logistic regressions to predict students' performance success. This was accomplished using statistical computing software that employed forward stepwise variable selection methods to identify key variables that accurately forecast student success. Once the logistic model was built with the necessary parameters

and predictors, the inverse logit function provided a probability of student success. In all cases, the logistic prediction models either matched or surpassed the performance of existing prediction methods, while using the same or fewer explanatory variables. The study demonstrated that a statistically sound approach can enhance current prediction methods and revealed the ineffectiveness of certain predictors typically used to estimate student performance.

2.7.6 Decision Tree Approach to Performance Prediction

Mesarić and Šebalj (2016) explored the use of decision trees to predict student success. They developed multiple decision tree models for classifying students' academic performance using various classification algorithms. Their research demonstrated that data mining tools can be effectively utilized by educational institutions to predict student outcomes. The REP Tree decision tree algorithm achieved the highest classification rate, reaching 79%..

In Hamoud's (2016) study, the focus was on selecting the best decision tree algorithm for predicting and classifying students' actions. The study compared the outcomes of using three different decision tree algorithms. Decision tree graphs were influenced by the number of input attributes and the final class attribute. Two main classes were selected for building the tree graph: student success (G3Grade) and the likelihood of pursuing higher education (higher). The results indicated that the J48 decision tree algorithm was the most effective, providing a useful roadmap for predicting and classifying students' actions.

2.7.7 Hybrid Approach to Performance Prediction

Francis & Babu (2019) conducted a study on predicting students' academic performance using a hybrid data mining approach. The results of applying the proposed hybrid algorithm to a student dataset revealed a strong correlation between students' behavior and their academic performance. The hybrid model combined clustering and classification, achieving an accuracy of 0.7547. When the model was applied to the academic,

behavioral, and additional features of the student dataset, it outperformed other existing algorithms.

Al-Shehri et al. (2017) conducted student performance prediction using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms. They applied each algorithm to the dataset to predict students' grades, finding that SVM achieved slightly better results with a correlation coefficient of 0.96, while KNN had a correlation coefficient of 0.95

Opara et al. (2020) developed a hybrid model using k-means and k-representative clustering algorithms to mine students' academic performance for decision-making purposes. The proposed model provided an effective algorithm for clustering students' academic performance and knowledge discovery. The hybrid model improved the K-means clustering algorithm for optimal solutions and efficient clustering of mixed datasets, achieving 99% performance and a clustering error of 0.0025 based on the mix classification matrix table. The study concluded that the proposed hybrid clustering algorithm efficiently mined students' academic performance, aiding constructive decision-making strategies. As a result, the researchers recommended that educational management systems should adopt their model's results for education performance monitoring and assessments. This approach can help universities understand the status, potential challenges, and abilities of students, while also guiding lecturers and academic advisers to help students achieve better performance.

In their study, Limbu & Sah (2019) used a hybrid clustering algorithm to predict students' academic performance. They discovered that while the K-means clustering algorithm, an unsupervised machine learning approach, achieved one of the highest output accuracies, the use of a hybrid model resulted in even higher output accuracy than any individual machine learning algorithm.

Amrieh et al. (2016) employed common ensemble methods such as Bagging, Boosting, and Random Forest (RF) in their research. The results showed a significant correlation between students' behaviors and their academic achievement.

In a study conducted by Ogwoka et al. (2015), a hybrid approach combining K-means clustering and decision trees was utilized to predict students' academic performance. The results showed that using this combination of algorithms enhanced the accuracy of predictions and was straightforward to implement in higher education institutions for assessing students' performance. Additionally, the study uncovered notable academic patterns and characteristics of students.

Hamsa et al. (2016) performed a scholarly performance prediction using Decision tree and Fuzzy Genetic and the result confirmed that prediction of students' tutorial performance in bachelor's and master's degree for each concern was once finished independently the use of decision tree and fuzzy genetic algorithm. The result shape Decision tree algorithm made more students at risk, which makes lectures a selection to take extra care for these college students that help to anticipate a better and almost percentage result from the closing exams. Results from the Fuzzy genetic algorithm offer greater passed students due to the fact of thinking about those who are in between hazard and safe to secure nation that offers students an intellectual satisfaction however the lectures will take attention on them directly.

Sultana et al. (2019) conducted a research on student's performance prediction using deep learning and data mining methods. It was found that classification implemented by MLP Multi-class Classifier, Decision trees, and Random Forest technique was more efficient

compared to other classifiers as seen in the accuracy and precision. Based on the results, the MLP technique was more efficient compared to other techniques in the prediction of students' performance.

Ameen et al. (2019) performed a review on students' academic performance and dropout predictions. It was noted that the essential concerns of the prediction of student academic performance (SAP) and student dropout are the nature of the attributes used in the mining and prediction of student performance of the Data Mining (DM) techniques used. The review showed that serious efforts have not been made in the way of standardizing the DM strategies for SAP and dropout prediction. Another revelation from this assessment was that of the fact that the SAP and dropout datasets are scarcely made public owning the truth that each institution of greater leaning considered SAP facts as too private to be made public.

Hashim et al. (2020) compared the performances of several supervised machine learning algorithms, such as Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Sequential Minimal Optimisation and Neural Network. They trained a model using datasets provided by courses in the bachelor study programmes of the College of Computer Science and Information Technology, University of Basra, for academic years 2017–2018 and 2018–2019 to predict student performance on final examinations. Results indicated that logistic regression classifier was the most accurate in predicting the exact final grades of students (68.7% for passed and 88.8% for failed).

Sawant et al. (2019) conducted research on student placement prediction model using gradient boosted tree algorithm. Their work implemented a gradient boosted tree algorithm that improved the student placement prediction. Through their study, they concluded that the gradient boosted tree algorithm was found to be 100% accurate with the feature importance and with different model evaluation metrics. The educational institution can predict the campus placement of each student and improve the placement of the university

Imran et al. (2019) on student academic performance prediction using supervised learning techniques they presented a student performance prediction model based on supervised learning technique Decision Tree. The performance of student's predictive model was assessed by a set of classifiers namely; J48, NNge, and MLP. In addition, an ensemble method was applied to improve the performance of these classifiers. The result showed that the proposed ensemble model including Decision tree (J48) classifier achieved the high accuracy which was 95.78 %.

Hasheminejad & Sarvmili (2019) conducted research on students' performance prediction based on particle swarm optimization and the research revealed that *S3PSO* performs better in rule detection than the other rule-based classification method like C4.5, ID3, and CART according to the support, confidence, and comprehensibility measurements. Moreover, comparing its results with those obtained by other classification methods reveals that *S3PSO* outperforms other classification methods like SVM and Neural networks. For example, it improved the value of accuracy criterion for the Moodle case study by 9%.

In a study by Fok et al. (2018), researchers explored a predictive model for students' future development using deep learning and the TensorFlow artificial intelligence engine. They found that a deeper model, characterized by a higher number of hidden layers, does not necessarily lead to more accurate results

Govindasamy & Velmurugan (2018) analyzed student academic performance using k-Means, k-Medoids, Fuzzy C Means (FCM) and Expectation Maximization (EM). The main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets with little or none of the background knowledge. The clustering algorithms are evaluated using execution time, purity and NMI. The result shows that FCM and EM algorithm performed well compared with other two clustering algorithms.

Nosseir & Fathy (2020) researched on a mobile application for prompt prediction of learner performance using fuzzy logic and artificial neural networks and concluded that the system tests the learners' basic knowledge in certain area that was related to his specialty or major they select. It develops a mobile app that had a database. It allows the lecturer or tutor to add questions and the system calculates the percentage of correct answers. Based on the real data provided by the university, the system incorporates a neural network that predicts the final GPA of the students. At that point this work is similar to the difference is in the attributes used to predict the performance.

Ajibade et al. (2018) proposed a predictive model evaluated using classifiers such as Naïve Bayesian (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), Discriminant Analysis (Disc), and Pairwise Coupling (PWC). To improve the performance of the classifiers, ensemble methods such as AdaBoost, Bagging, and RUSBoost were employed to enhance the accuracy of the students' performance model. The results demonstrated a strong correlation between students' actions and their academic performance. The proposed model achieved an accuracy of 84.2% when incorporating behavioral features, while without them, it achieved 72.6%. Furthermore, when ensemble methods were applied to the classifiers, the model's accuracy increased to 94.1%, indicating the reliability of the proposed model in predicting academic performance.

The outcome of Adejo & Connolly (2018) showed that the approach of using multiple data sources along with heterogeneous ensemble techniques is very efficient and accurate in prediction of student performance as well as help in proper identification of pupil at risk of abrasion. The research empirically examined and compared the performance accuracy and efficiency of single classifiers and ensemble of classifiers that make use of single and multiple data sources. The study developed a novel hybrid model that can be used for predicting student performance that is high in accuracy and efficient in performance. The research advanced the understanding of the application of ensemble techniques to predicting student performance using learner data.

Ragab et al. (2021) utilized boosting, random forest, bagging, and voting algorithms, which are the normal group of techniques used in studies. By using ensemble methods, good result was achieved that demonstrates the dependability of the proposed model. For better productivity, the various classifiers were gathered and, afterward, added to the ensemble method using the Vote procedure. The implementation results demonstrate that the bagging method accomplished a cleared enhancement with the DT model, where the DT algorithm accuracy with bagging increased from 90.4% to 91.4%. Recall results improved from 0.904 to 0.914. Precision results also increased from 0.905 to 0.915.

Bithari et al. (2020) in this study the predictive model was built on the data of engineering students which includes various attributes. The multi-class classification was used to predict the students into four categories (Excellent, Good, Medium, and Satisfactory). The classification has been done by three individual traditional classifiers first and then the voting was done in the second phase. The result obtained shows significant improvement in the performance when the ensemble method was implemented. It also removes the slight over fitting which was seen in the case of some individual classifier.

In the study by Bithari et al.,(2020) a predictive model was developed using data from engineering students, including various attributes. The model employed multi-class classification to categorize students into four groups: Excellent, Good, Medium, and Satisfactory. Initially, three separate traditional classifiers were used for classification, followed by a voting process in the second phase. The findings indicated a significant improvement in performance when the ensemble method was applied. This approach also helped eliminate slight overfitting observed with some individual classifiers.

A research that was done by Ajibade et al. (2018) revealed that a new performance prediction model for students was proposed which was based on various data mining methods which contains new features known as interactive features. These attributes were associated with the interactivity of learners with the LMS. The predictive model was evaluated based on some classifiers like NB, DT, KNN, DISC and PWC. Furthermore, ensemble techniques were applied to enhance the performance of the classifiers. Bagging,

Ada-Boost and Random under sampling was used. The outcome showed that there exist a significant and strong connection between the behavior of learners and their academic performance. The accuracy of the predictive model using behavioral features was 84.2% and 72.6% without behavioral features. After the ensemble methods was applied it achieved an accuracy of 94.1%.

2.8 Critiques of Existing Literature

The field of classification has seen the use of various classification algorithms with different input datasets. However, scholars have noted key issues in current approaches, especially when using classification for predicting student academic performance. Adewale et al. (2018) explored the potential of artificial neural networks (ANNs) in predicting secondary students' post-UTME grades before their admission into the university system. The model was specifically developed using a feed-forward neural network architecture and based on selected input variables. The ANN model achieved an accuracy of 90%, demonstrating its potential as a predictive tool, a clustering instrument, and a criterion for selecting candidates for university admission. Despite the high prediction accuracy of ANN in nonlinear phenomena, the model did not enable an understanding of how predictor variables relate to each other in explaining academic outcomes. In other words, the ANN model does not provide a clear mathematical model of the relationship between inputs and outputs.

Bithari et al. (2020) predicted the academic performance of engineering students using an ensemble method. They developed a predictive model utilizing traditional classifiers such as Decision Tree, SVM, and Linear Regression, which had shown good results in similar studies. Following that, they implemented an ensemble method known as voting, which is recognized for enhancing the performance of individual classifiers. The voting classifier averages the predictions of base classifiers, combining them to improve accuracy. The results indicated that accuracy, precision, recall, and F1 score were significantly enhanced with the use of ensemble voting compared to using individual classifiers alone. The

research emphasized the importance of ensemble classifiers over single classifiers in predicting academic performance.

Sokkhey & Okazaki (2020) introduced a hybrid approach combining Principal Component Analysis (PCA) with four machine learning (ML) algorithms—Random Forest (RF), C5.0 of Decision Tree (DT), Naïve Bayes (NB) of Bayes network, and Support Vector Machine (SVM)—to enhance classification performance and address the issue of misclassification. However, the model did not fully utilize maximum k-fold cross-validation evaluations to solve prediction and classification challenges.

Abba et al. (2020) developed a hybrid machine learning ensemble technique for modeling dissolved oxygen (DO) concentration in Kinta River, Malaysia. They utilized four different AI-based models, including long short-term memory neural network (LSTM), extreme learning machine (ELM), Hammerstein-Weiner (HW), and general regression neural network (GRNN), to model DO concentration using available water quality parameters. The study proposed exploring other emerging optimization algorithms, deep learning models, and various black box models in conjunction with promising ensemble approaches to further improve prediction accuracy.

Patil et al. (2022) explored the application of various machine learning techniques using ensemble methods such as random forest, gradient boosting, adaptive boosting, and XGBoost for diabetes prediction. The study utilized GridSearch CV for hyper-parameter tuning with both 5-fold and 10-fold cross-validation. The research emphasized the importance of using numerous base learners with low bias and high variance to enhance predictions, as these base learners can correct errors made by previous ones. This approach can lead to more accurate and robust predictions in diabetes prediction tasks.

Verma et al. (2022) conducted a prediction task at the entry-time using four single supervised educational data mining algorithms Decision Tree, Naïve Bayes, k-Nearest Neighbor, and Support Vector Machine along with an ensemble method called "Random Forest." These classifiers were applied to a dataset from an Indian Engineering College,

which included parameters related to students' backgrounds, academics, social factors, and psychological aspects. However, the study had limitations, such as the limited size of the dataset and the fact that it was from just one institution. The researchers suggested that further studies should incorporate larger sample sizes from different regions to explore the effects of various attributes on students' academic performance in greater depth and provide a more general perspective.

Fijani et al. (2021) explored advanced hybrid machine learning algorithms for multistep lake water level forecasting. The hybrid ML techniques yielded more accurate forecasts compared to standalone models, although the performance improvements varied depending on the forecast lead times and whether autoregressive or moving algorithms were employed. This study highlighted the potential of the support vector machine (SVM) algorithm as a novel approach in forecasting problems, demonstrating its ability to produce accurate multistep water level forecasts. The research considered other current work in the hybrid machine learning ensemble field and offered improved solutions to enhance the performance of hybrid ML algorithms. Similarly, Sokkhey and Okazaki (2020) recommended further work on hybrid machine learning for predicting student performance, emphasizing the importance of maximizing k-fold cross-validation evaluations to tackle prediction and classification challenges.

Sokkhey & Okazaki (2020) developed hybrid models incorporating Principal Component Analysis (PCA) and 10-fold cross-validation, yielding highly satisfactory results. They found that combining baseline models with PCA and assessing them using k-fold cross-validation led to the creation of high-performing hybrid models. These models demonstrated potential as effective algorithms for addressing prediction and classification challenges.

Zulfiker et al. (2020) conducted research on predicting student performance at private universities in Bangladesh using machine learning approaches. The study employed seven base classifiers to predict students' final grades and then combined the outputs of the base classifiers using a weighted voting approach. Observations revealed that aggregating the

base classifiers with the weighted voting approach resulted in increased accuracy. The study also noted that, based on the achieved AUC values, the Weighted Voting Classifier was nearly perfect for classifying the collected dataset.

Ragab et al. (2021) developed a predictive model using artificial neural networks, decision trees, and naïve Bayesian methods. The study also investigated the use of bagging and boosting techniques to improve the performance of these classifiers. The findings indicated improvements in the models compared to traditional classifiers. By combining two different classifiers with either bagging or boosting, the researchers achieved better outcomes than previous methods, contributing to advancements in student performance and educational systems. The study recommended applying these models to more datasets and exploring the diverse range of effective classifiers available.

Ajibade et al. (2022) discussed the rapid rise in popularity of e-learning due to the increasing prevalence of the internet. This expansion has resulted in the generation of vast amounts of data, which can be leveraged to enhance e-learning methods through data mining processes. By analyzing data and performing feature selection, the researchers aimed to boost students' academic performance within e-learning systems. They identified behavioral features as crucial because these features reveal how actively learners engage with the e-learning system. Effective feature analysis can lead to the development of a robust prediction model, thereby improving the overall quality and outcomes of e-learning.

Table 2.1: Advantages and Disadvantages of the Traditional Classifiers

Model/Algorithm	Advantages	Disadvantages
K-Nearest Neighbors	According to Liu et al., (2020) KNN is easy to realize without parameter estimation.it is suitable for solving multi-classification problems	Liu et al., (2020) found that KNN is greatly affected by data skew, the computational overhead is very large.
	Ragabet al., (2021) found that KNN works well with non-linearly separable classes and perform well in multimodal classes.	Ragab et al., (2021) stated that KNN uses extra time for determining the nearest neighbor in a large training dataset.
Support Vector Machine	Kadambande et al., (2017) discovered that SVM works really well with clear margin of separation. It is very effective in high dimensional spaces.	Kadambande et al., (2017) confirmed that SVM doesn't perform well, when there is large data set because the required training time is higher. If data set has noisy then system doesn't perform very well. I.e. target classes are overlapping.
Decision Tree	Liu et al., (2020) revealed that Decision Tree is relatively easy to understand and Interpret. they can handle samples with missing values or large scale	According to Liu et al., (2020) Decision Tree is prone to over fitting. It does not support online learning. Unstable to noise. Computationally Expensive on Large Datasets

2.9 Similarities of the Traditional Classifiers

Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN) Multi-Layer Perceptron (MLP) and Naïve Bayesian share the following similarities:

Supervised learning encompasses both Support Vector Machines (SVM) and MLP. This means that before they can make predictions based on data that hasn't been seen, they need to develop a model using labeled training data.

KNN is a supervised learning algorithm. It categorizes instances according to how similar they are to their neighbors using labeled data.

Table 2.2: Advantages and disadvantages of the ensemble's techniques

Ensemble Method	Advantages	Disadvantages
Bagging	According to Amrieh et al., (2016) Bagging reduces variance and improves accuracy, can turn weak learners into strong learners, and works well with high variance models.	Amrieh et al., (2016)found that Bagging Can increase bias, may not work well with low-variance models, and can be computationally expensive for large datasets,
Boosting	Adejo & Connolly (2018) discovered that Boosting improves accuracy and reduces bias, works well with high-bias models and imbalanced data	Adejo & Connolly (2018) revealed that Boosting can over fit with noisy data and outliers, can be computationally intensive
Stacking	According to Adejo & Connolly (2018) Stacking improves prediction accuracy by combining models with different strengths and weaknesses, and can build a more reliable meta-model	Adejo & Connolly (2018) found that Stacking can be complex and time-consuming to implement, especially with large datasets

Table 2.3: Similarities of the Ensemble Techniques

Technique	Bias-Variance Trade-off	Base Models	Model Combination
Boosting	Pandey, M., & Taruna, S. (2014) discovered, that Boosting Focuses on reducing bias by iteratively improving the performance of weak models.	Pandey, M., & Taruna, S. (2014). Found that Boosting makes use of homogeneous models as well, but modifies their weights based on performance.	Pandey, M., & Taruna, S. (2014) revealed that Boosting assigns weights to classifiers based on their performance.
Bagging	Amrieh et al., (2016). Said Bagging Reduces variance by creating additional training data through bootstrapping (sampling with replacement) from the original dataset.	Amrieh et al., (2016) found that Bagging typically uses homogeneous base models (e.g., multiple decision trees)	According to Amrieh et al., (2016) Bagging combines predictions using equal-weight voting, each sample of data is chosen with equal probability.
Stacking		According to Adejo & Connolly (2018) Stacking combines results from heterogeneous base models.	Adejo & Connolly (2018) Introduced a meta-level model to estimate weights for each base model.

2.10 Metrics Measured

The metrics that were measured in this research were tabulated into a table called confusion matrix as shown on figure 2.10. Confusion matrix is a data set that only has two classes, one class as positive and the other class as negative, consisting of four cells, namely True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) as shown in Figure 2.10

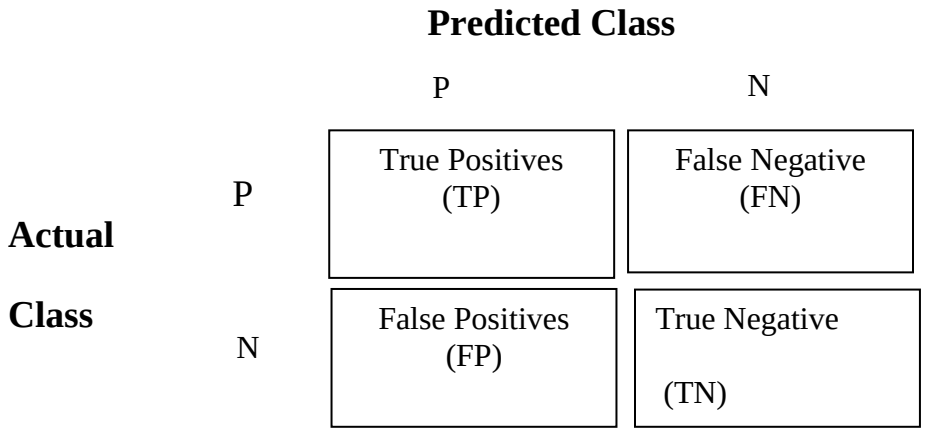


Figure 2.10: The Confusion Matrix

Accuracy this is the percentage of successfully calculated forecasts in the total number of predictions. Precision this is the ratio of correctly classified cases to the total number of misclassified and correctly classified cases. Recall this is the proportion of correctly classified instances to the total number of unclassified and correctly classified cases. Recall and precision was combined using the F-measure, which is regarded a good indicator of their relationship. The following equations were used to calculate accuracy, precision, recall and the F-Measure.

The Area under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the AUC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \text{---------- (2.16)}$$

$$Precision = \frac{TP}{TP+FN} \text{---------- (2.17)}$$

$$Recall = \frac{TP}{TP+FN} \text{---------- (2.18)}$$

$$\mathrm{F-measure} = 2\frac{Precision*Recall}{Precision+Recall}\text{----------} \quad (2.19)$$

2.11 Research Gaps

Al Husaini and Shukor (2022) identified several factors that have a significant impact on students' academic performance. These include low entry grades, family support, accommodation, student gender, prior assessment grades, internal assessment grades, GPA, and engagement in e-learning activities

Qureshi et al. (2023) used structural equation modeling (SEM) to analyze the impact of social factors on students' learning performance. Their findings revealed that social factors such as interaction with peers and teachers, social presence, and social media use positively influence active collaborative learning and student involvement. The research also highlighted the role of double mediation in these relationships. As online learning becomes more prevalent, fostering collaborative learning and engagement through social factors was found to enhance students' learning activities. Therefore, higher educational institutions should integrate these social elements into teaching and learning to positively impact students' academic development.

Batool et al. (2023) conducted an extensive investigation that emphasized students' academic histories and demographic variables as the most reliable indicators of academic performance. The study noted that including irrelevant data points in the dataset not only lowers predictive accuracy but also increases the time needed to process models. Consequently, close to half of the studies used methods to select relevant features before developing prediction models.

Despite the significance of these variables, there has been limited research on student behavior during the learning process and its impact on academic achievement. Educational datasets often include longitudinal information, such as students' academic courses, course histories, and patterns of engagement over time. The research contributed to the literature by using tree-based feature selection methods specifically designed to handle longitudinal

and temporal data. These methods help identify informative features that capture students' evolving learning behaviors and performance over time. By exploring these feature selection methods, the study aimed to improve the predictive accuracy of ensemble models and enable personalized interventions to support students' academic success.

.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Chapter Summary

This chapter describes research design that was adopted during the design and development of the proposed model in student performance prediction, data collection techniques, data analysis processes, research environment and the evaluation of the proposed model. In this research, a student performance model using ensemble approaches was proposed. Ensemble methods are a form of problem-solving approach that employs several models to solve a problem. In contrast to traditional learning approaches, which train data using a single learning model, ensemble methods attempt to train data using a variety of models and then combine them to vote on their outcomes. Ensemble projections are often more accurate than single-model predictions. This method's purpose is to offer an accurate assessment of the features that may influence a student's academic progress. Figure 3.1 gives the flow chat of the steps that were followed to come up with the proposed model.

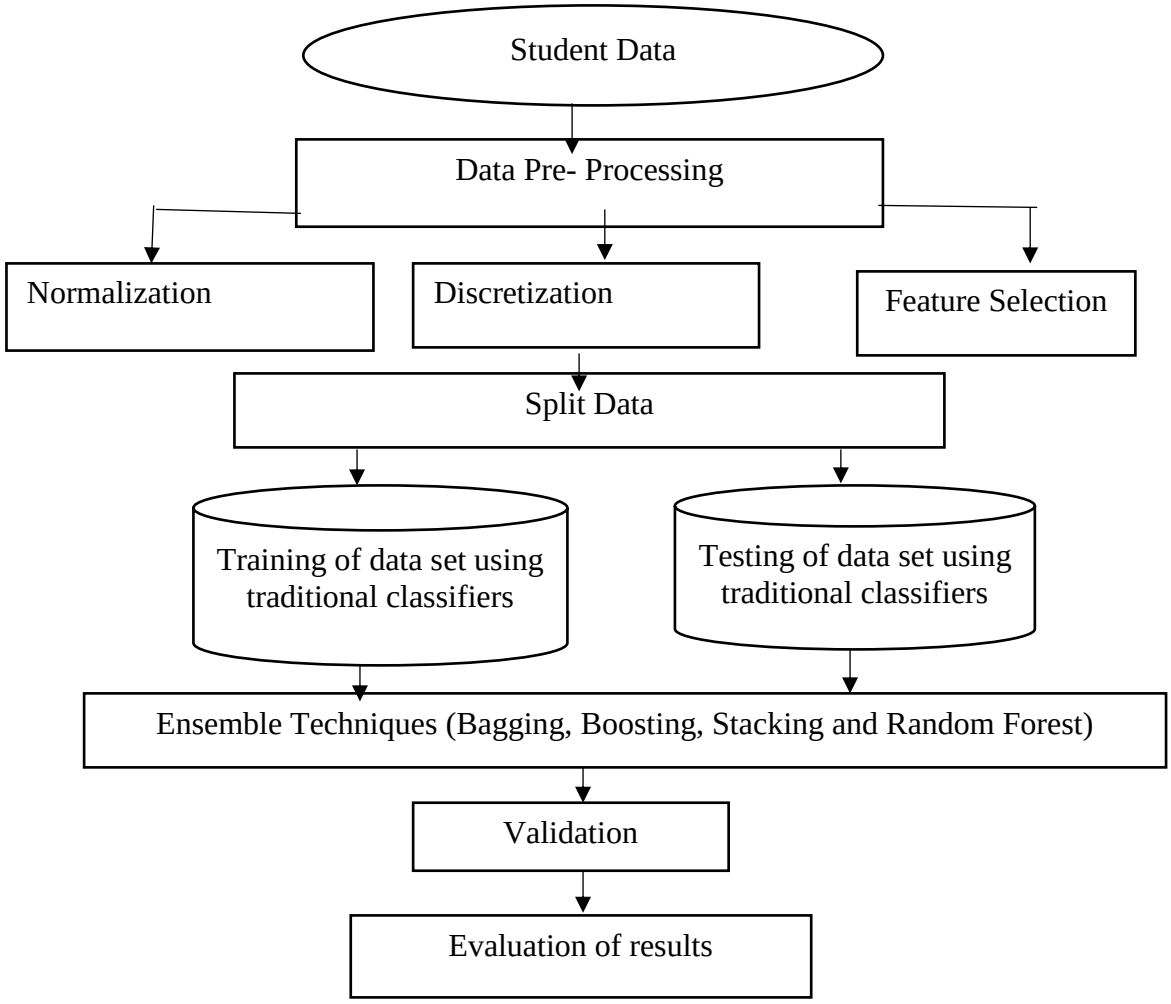


Figure 3.1: Student's Performance Prediction Model Research Steps

3.2 Student Data

Several datasets are available for non-commercial use in order to train neural networks: This research used the existing datasets from Kaggle website because, Kaggle dataset, competition: Challenges are performed on a regular basis, with participants having open access to the dataset (https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics/data); Kaggle bills itself as a "home of data science," and the site's huge collection of datasets definitely lends some validity to that claim.

Table 3.1: Features Used for Student Performance Prediction

No.	Variables	Description	Type
1	School	Student's School	Binary (Msongari or Valley Road)
2	Sex	Student's Sex	Binary ("F" Female or "M" Male)
3	Age	Student's Age	Numeric (From 15 to 22)
4	Address	Students' Home Address	Binary ("U" Urban or "R" Rural)
5	Family Size	Family Size	Binary (: "LE3" Less or equal to 3 or "GT3"Greater than 3)
6	Parent Status	Parent's Cohabitation Status	Binary (: "T" Living together or "A" - Apart)
7	Mother's Education	Mother's Education	Numeric:(0 None, 1 Primary Education (4^{th} Grade), 2-5^{th} to 9^{th} grade, 3-Secondary Education or 4-Higher Education
8	Father's Education	Father's Education	Numeric (: 0 None, 1 Primary Education (4th grade), 2 – 5th to 9th grade, 3 – Secondary Education or 4 – Higher Education)
9	Mother's Job	Mother's Job	Nominal ("Teacher", "Health" Care Related, Civil "Services" (e.g. Administrative or Police), "At home" or "Other").
10	Father's Job	Father's Job	Nominal ("Teacher", "Health" Care Related, civil "Services" (e.g. Administrative or police), "at Home" or "Other")
11	Reason	Reason for Choosing the School	Nominal: Close to "Home", School "Reputation", "Course" Preference or "Other")
12	Guardian	Student Guardian	Nominal ("Mother", "Father" or "Other")
13	Travel Time	Home to School Travel Time	Numeric1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	Study Time	Weekly Study Time	Numeric1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	Failures	Number of Past Class Failures	Numeric n if 1<=n<3, else 4)
16	School Support	Extra Educational Support	Binary (Yes or No)
17	Family Support	Family Educational Support	Binary (Yes or No)
18	Paid	Extra Paid Classes within the	Binary: (Yes or No)
19	Activities	Extra-Curricular Activities	Binary: (Yes or No)
20	Nursery	Attended Nursery School	Binary: (Yes or No)
21	Higher	Wants to Take Higher Education	Binary: (Yes or No)
22	Internet	Internet Access at Home	Binary: (Yes or No)
23	Romantic	With a Romantic Relationship	Binary: (Yes or No)

No.	Variables	Description	Type
24	Family Relationship	Quality of Family Relationships	Numeric: 1 Very bad,5 Excellent
25	Free Time	Free Time After School	Numeric:1 Very low, 5 very high
26	Going out with friends	Going Out with Friends	Numeric:1 Very low, 5 very high
27	Work Day Alcohol	Workday Alcohol Consumption	Numeric: 1 Very low,5Very high
28	Weekend Alcohol	Weekend Alcohol Consumption	Numeric:1 Very bad 5Very good
29	Health	Current Health Status	Numeric: 1Very bad 5Very good
30	Absences	Number of School Absences	Numeric: 0 to 93
31	First Grade Period	First Period Grade	Numeric: 0 to 20
32	Second Grade Period	Second Period Grade	Numeric: 0 to 20
33	Final Grade	Final Grade (Out Put Target)	Numeric: 0 to 20

3.2.1 Data Visualization

Data visualization is an important aspect of the preprocessing process, which use graphs to simplify complex data. Python libraries such as Seaborn and Matplotlib were used to visualize the data set. Graphical representations can assist instructors in better understanding their students and monitoring what is going on in student's classes.

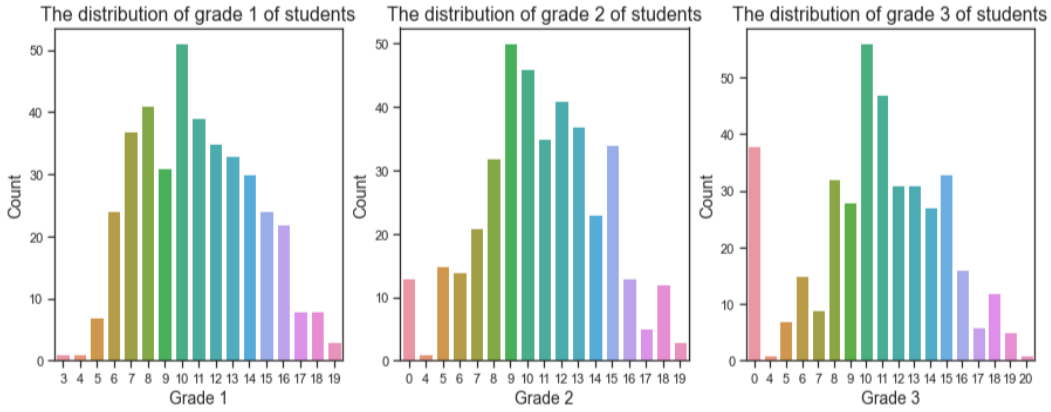


Figure 3.2: Histogram Showing the Grade Distribution of the Students

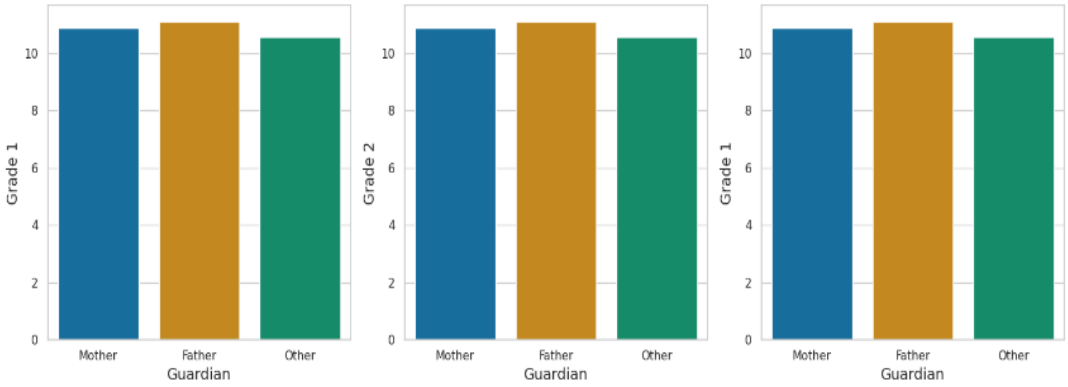


Figure 3.3: Histogram Showing How Both Parents Impact Student Grades

The students with their fathers or mother as guardians got higher grades. Fathers can serve as role models for their children, demonstrating the value of education through their own actions and attitudes towards learning. Mothers often play a central role in a child's early cognitive and social development. Engaging in educational activities and providing a nurturing environment can stimulate a child's curiosity and readiness for school.

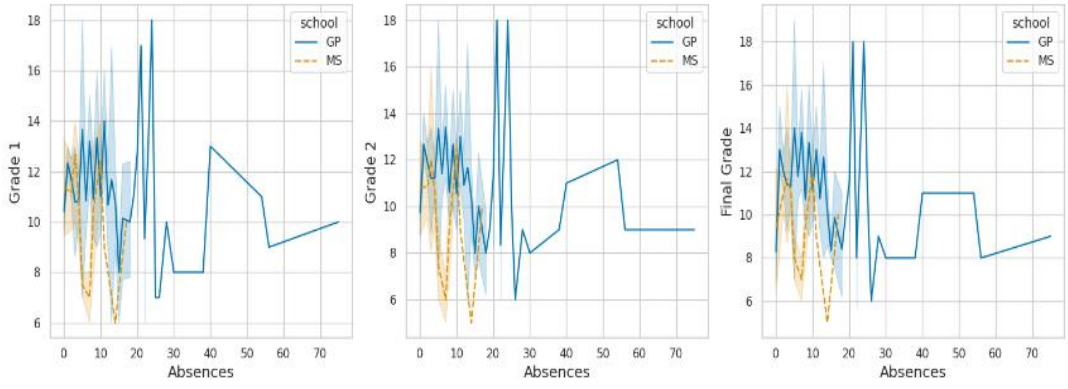


Figure 3.4: Line Graph Showing How Absences Affect Student Grades

From the above diagram it shows the graph of three grades that is Grade1, Grade2, and Grade 3 respectively. The Students of GP who were absent for around 20-24 days, tends

to do better on exam. Students of MS who were absent for around 3 days or 10 days, tends to do better on exam.

Heatmap

This research used Heatmap, Heatmap is a graphical representation of data using colors to visualize the value of the matrix. In this, to represent more common values or higher activities brighter colors basically reddish colors are used and to represent less common or activity values, darker colors are preferred. Heatmap is also defined by the name of the shading matrix.

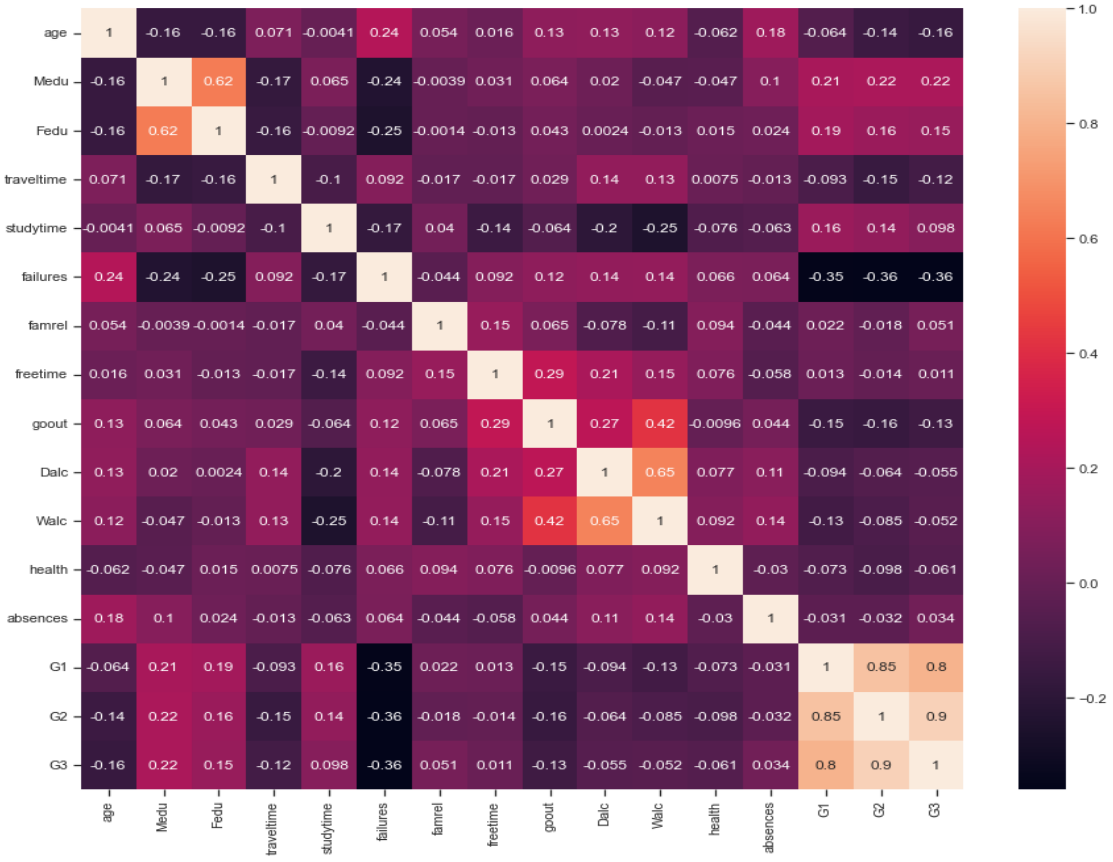


Figure 3.5: Heat Map Showing How the Parameters Contribute to Performance Prediction

From the Heat map above, the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

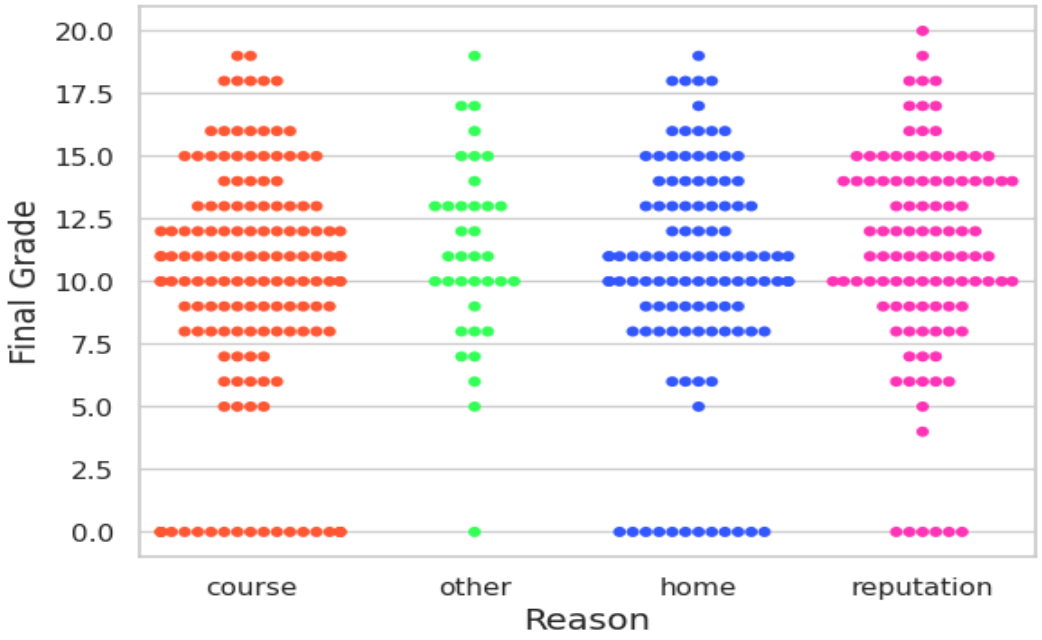


Figure 3.6: A Graph Showing Relationship between Grades and Reason for Choosing the School

We can see that the students who chose to join this school because of specific course, close to their home, or because of school reputation got final grades slightly higher than others.

3.3 Data Pre-Processing

Data preprocessing followed where the original data was transformed to a suitable shape to be used by a particular mining algorithm. The dataset contained 650 records and 33 features categorized into numerical (17 features) and categorical (16 features). Normalization was done to change the categorical data to numerical, so that all the values

were brought within the range [0.0-1.0]. This process was used to speed up the learning process by preventing attributes with large ranges from outweighing attributes with smaller ranges. This increased the number of features to 59. Discretization was used to convert the performance of student from numerical values to nominal values, and this signifies the class labels of the classification problem. To carry out Discretization, the data set was split into three nominal intervals (Excellent, Good, Fair) which is based on the final grade of students like fair interval consist of score ranging from 0 to 9, while good interval consists of scores that range from 9 to 14 and Excellent consist values from 14–19. The dataset after discretization comprises of 301 students with Fair, 154 students with Good and 194 students with Excellent.

3.4 Feature Importance Analysis

Feature importance is a step in building a machine learning model that involves calculating the score for all input features in a model to establish the importance of each feature in the decision-making process. The higher the score for a feature, the larger effect it has on the model to predict a certain variable. In this research, a feature importance analysis to improve the model's interpretability was performed. Table 3.2 shows the contribution of each variable to prediction performance. On the basis of an actual data collection, a feature significance analysis was used to discover relevant features. Base-learners create feature significance scores in a variety of methods. The Decision Tree feature importance is the weight vector, which contains the coefficients, and these coefficients define an orthogonal vector to the hyperplane. The RF model calculates the associated out-of-bag (OOB) error to determine the relevance of the feature. The XGBoost computes feature scores by determining the normalized total decrease in mean squared error caused by that feature, with the sum of all feature significance levels equaling one. The results of these techniques are normalized, and the ultimate level of feature relevance is determined by the average of the three

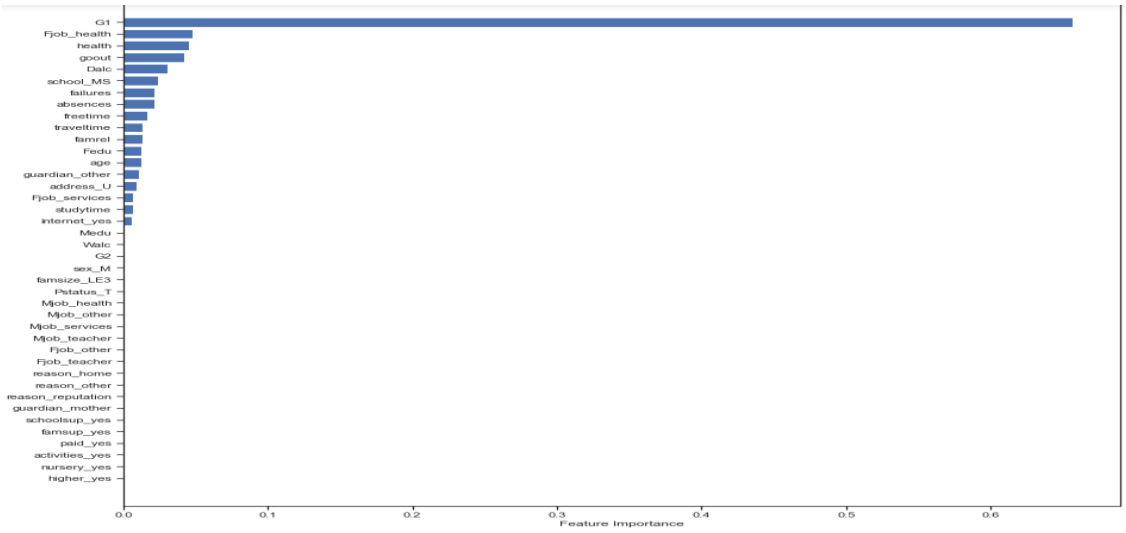


Figure 3.7: Feature Importance from Decision Tree Classifier

Based on the figure 3.7, the most important features, in descending order of importance, are as follows: G1, Fjob-health

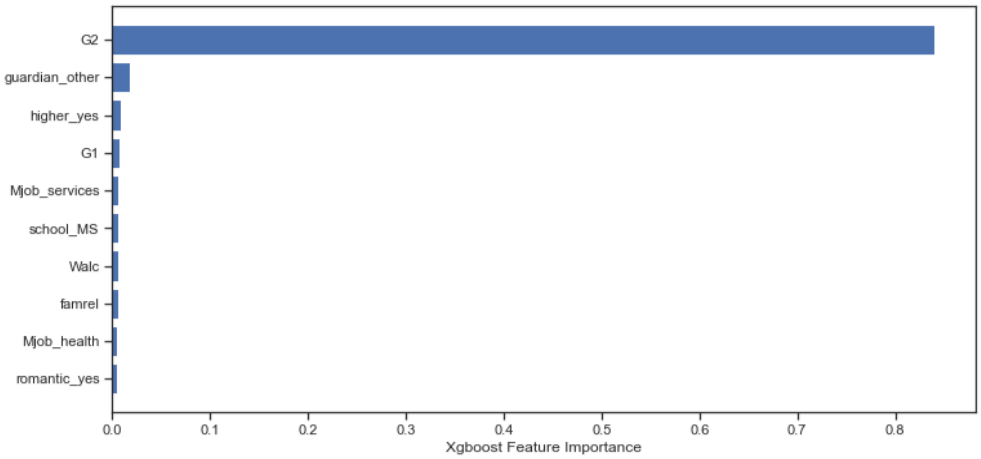


Figure 3.8: Feature Importance from XGboost Classifier

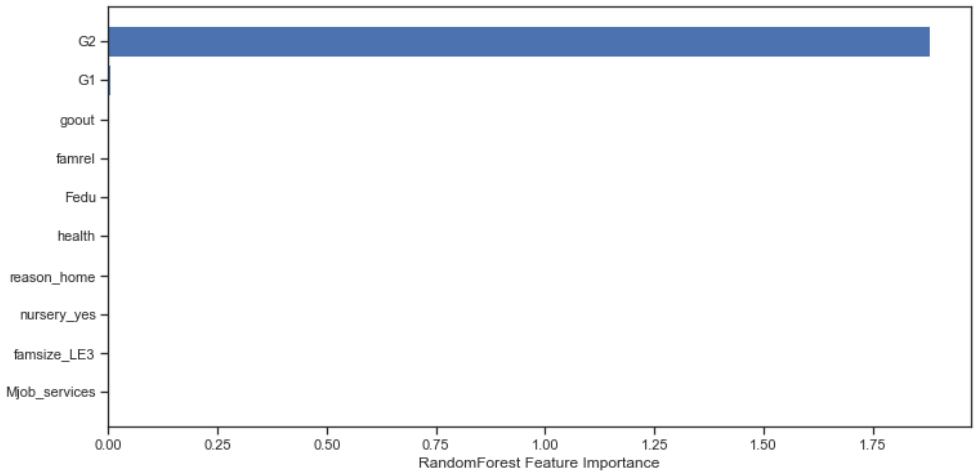


Figure 3.9: Feature Importance from Random Forest Classifier

3.5 Data Splitting

After selection was done and the features were obtained, the dataset was then split to train with a size of 0.7 and test with a test size of 0.3.

3.6 Setting up of the Environment Required for the Experiment

The required software tools were gathered differently. The required hardware computing environment required specifications were a minimum of 4GB RAM, a webcam camera, 500 gigabyte internal hard disk space and a CPU of 3.2 gigahertz speed. The following programming language were installed on the hardware computer. Python3 which is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python3 build the entire structure code developed for predicting students' academic performance. Python3 made it to achieve high-level interaction nature of scientific libraries for a good platform of development.

The Jupyter Notebook was also used which is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience. Data science can be learned using, machine learning and python projects without necessary install of other packages, because many packages like numpy, pandas,

scikit learn come with anaconda. Anaconda is an open-source distribution of the Python and R programming languages for data science that aims to simplify package management and deployment. Package versions in Anaconda are managed by the package management system, conda, which analyzes the current environment before executing an installation to avoid disrupting other frameworks and packages.

The Anaconda distribution comes with over 250 packages automatically installed. Over 7500 additional open-source packages can be installed from PyPI as well as the conda package and virtual environment manager. It also includes a GUI (graphical user interface), Anaconda Navigator, as a graphical alternative to the command line interface. Anaconda Navigator is included in the Anaconda distribution, and allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages, install them in an environment, run the packages and update them.

Scikit-learn Library Scikit-learn is another actively used machine learning library for Python. It includes easy integration with different Machine Learning (ML) programming libraries like NumPy and Pandas. Scikit-learn comes with the support of various algorithms such as: Classification, Regression, Clustering Dimensionality Reduction, Model Selection, and Preprocessing. Built around the idea of being easy to use but still be flexible, Scikit-learn is focused on data modelling and not on other tasks such as loading, handling, manipulation and visualization of data. It is considered sufficient enough to be used as an end-to-end ML, from the research phase to the deployment.

3.7 Training of the Model

Each of the Machine Learning (ML) algorithms mentioned was trained using Python 3 and Scikit-learn library to obtain the accuracy corresponding to the dataset. Then 10-fold cross validation (CV) was performed to the above mentioned Machine Learning (ML) algorithms by Shuffling the entire samples randomly, then the sample was split into k sub folds, In the split into k sub folds, 1 fold was held as test set, then the remaining k -1 folds

as the training set, the assessment score was kept and discarded the model as the training set, and the process was repeated until every single fold was considered as a testing set. Finally, computing the average score of all the scores.

3.8 Testing of the Traditional Classifiers

After training the traditional classifiers, testing of the traditional classifiers followed. The testing of the classifiers was done until the desired output was achieved.

3.9 Applying of the Ensemble Techniques

After training and testing the traditional classifiers. And ensuring the base models are diverse and are able to capture different aspects of data. The appropriate hyper parameters are chosen for the base model and avoid over fitting by using regularization techniques and cross-validation. Monitor the performance of the ensemble on both the training and validation sets.

Experiment with different ensemble methods and combinations of base models to find the best-performing ensemble for your problem

3.10 Validation of the Results

The validation process begun after the classification model had been trained. The validation process is the final step in building a predictive model; it is used to assess the model's performance by comparing it to real-world data.

3.11 Evaluation of Results

The Model was evaluated on Accuracy, Precision, Recall, F1-score and Area under the Curve (AUC) by first recording the results of the classifiers when they were trained and tested for the first time. Followed by the results of the base classifiers on 10- fold Cross Validation (CV). Then after applying the ensemble techniques.

CHAPTER FOUR

RESULTS ANALYSIS AND DISCUSSION

4.1 Introduction

This chapter gives a presentation of the findings obtained from student performance prediction that were carried out.

Table 4.1: The Results of the Traditional Classifiers with and without Behavior Feature

Evaluation Measure	DT		KNN		SVM	
Behavioral Feature Extraction	BF	WBF	BF	WBF	BF	WBF
Accuracy	0.87	0.84	0.82	0.81	0.86	0.83
Precision	0.85	0.77	0.80	0.67	0.81	0.70
Recall	0.86	0.78	0.79	0.56	0.81	0.73
F-Measure	0.86	0.77	0.69	0.59	0.75	0.71

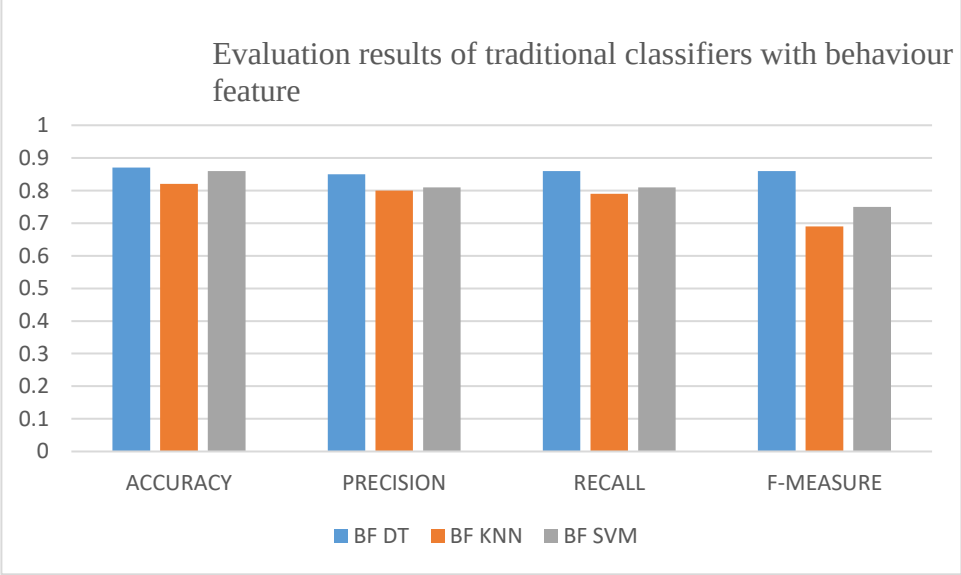


Figure 4.1: Results of the Traditional Classifiers with Behavior Feature

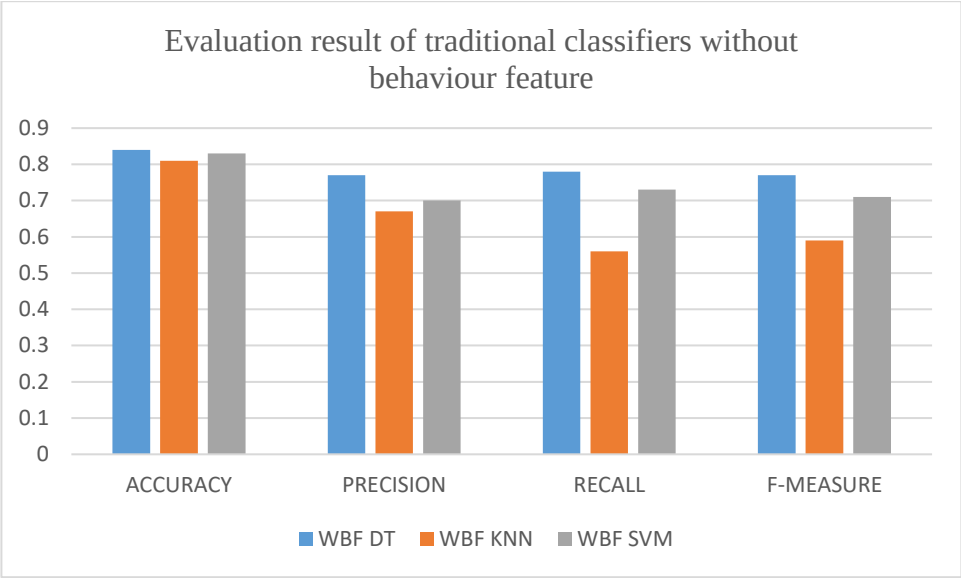


Figure 4.2: Results of the Traditional Classifiers without Behavior Feature

4.2 Evaluation Results Using Traditional Classifiers

Numerous factors influence the model when forecasting student achievement. In this research, behavioral characteristics have been identified as important factors that can influence student success. Table 4.1, Figure 4.1 and figure 4.2 displays the results of classification methods (DT, KNN, and SVM) to show the impact of behavioral features. The classification findings are divided into two categories. Classification results with student behavioral features (BF) and Classification results without student behavioral features (WBF). According to Table 4.1, Figure 4.1 and figure 4.2, the DT model outperforms other data mining techniques. The DT model obtained 87.1% BF accuracy and 84.4% WBF accuracy. In terms of precision, the model achieved 85.2% with BF and 77.7% with WBF. The recall measure yields 86.3% with BF and 78.1% with WBF. The values for F-Measure are 86.0% with BF and 77.1% with WBF. As a result of the preceding analysis, the results show that learner behavior has a significant impact on students' academic success.

Table 4.2: Results of the Ensemble Classifiers

Evaluation Measure	Classification Methods			Bagging			Boosting			Random Forest
Classifier	DT	KNN	SVM	DT	KNN	SVM	DT	KNN	SVM	DT
Accuracy	0.87	0.82	0.86	0.86	0.85	0.86	0.89	0.86	0.88	0.89
Precision	0.85	0.80	0.81	0.86	0.79	0.80	0.87	0.83	0.82	0.84
Recall	0.86	0.79	0.81	0.87	0.80	0.82	0.87	0.83	0.84	0.84
F-Measure	0.86	0.69	0.75	0.88	0.69	0.76	0.89	0.74	0.80	0.73

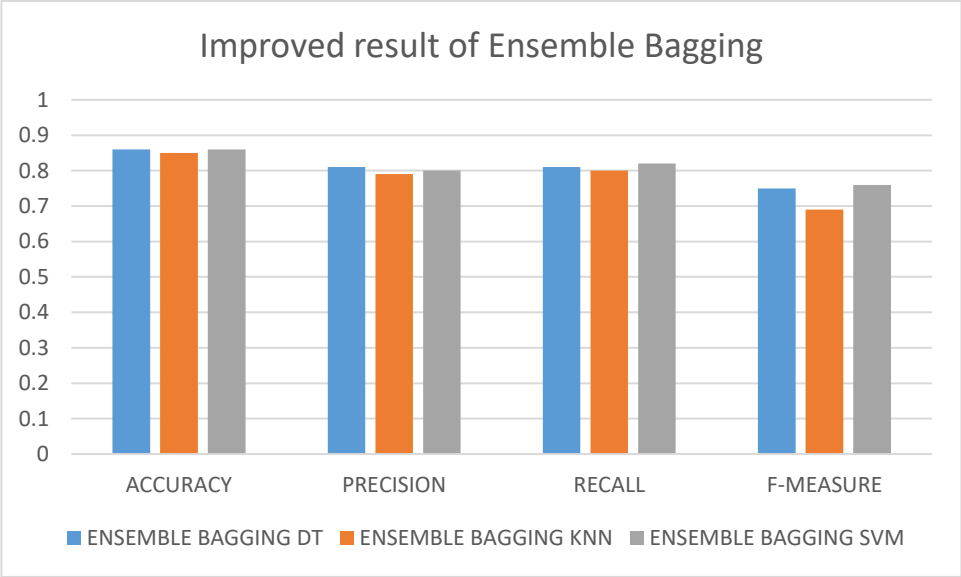


Figure 4.3: Results of the Ensemble Classifier Bagging

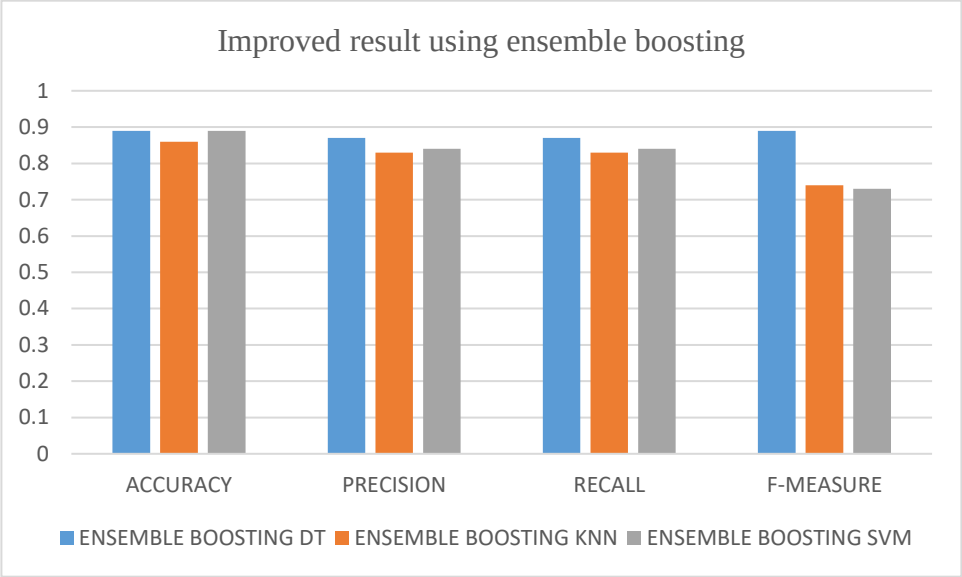


Figure 4.4: Results of the Ensemble Classifier Boosting

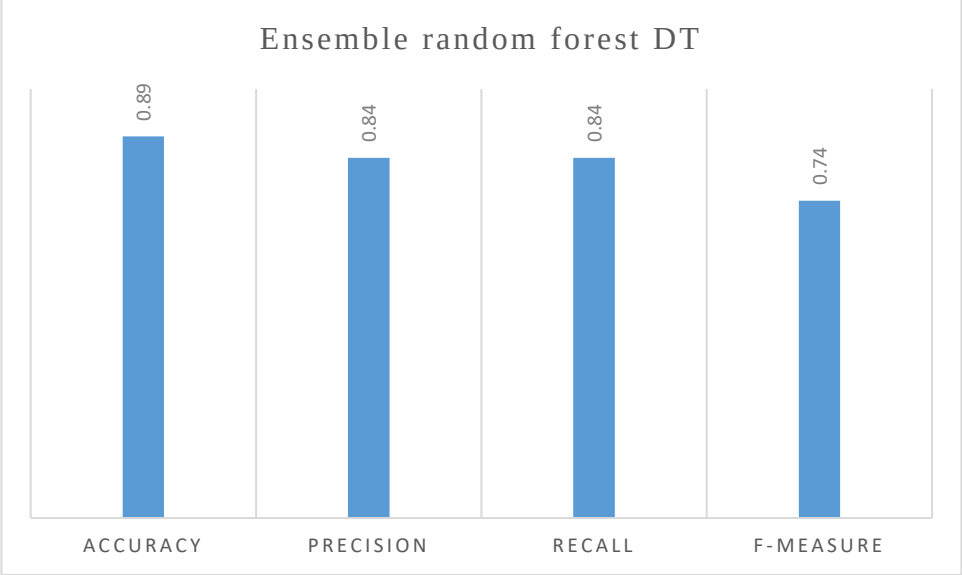


Figure 4.5: Results of the Ensemble Classifier Random Forest

4.3 Evaluation Results Using Ensemble Methods

In this section, ensemble approaches were used to improve the accuracy of Traditional classifiers Data Mining techniques' evaluation outcomes. Table 4.3 displays the improved

results obtained by combining ensemble methods with three traditional classifiers (DT, KNN, and SVM). Each ensemble trains three classifiers and then uses a majority voting technique to combine the findings in order to attain the best prediction performance of students. Boosting techniques outperform other ensemble methods in the cases of DT, KNN, and SVM, but DT provided the best results, with accuracy increasing from 0.87 to 0.89, precision increasing from 0.85 to 0.87, and recall increasing from 0.86 to 0.87, and F-measure results improved from 0.86 to 0.89.

Table 4.3: Results After Testing and Validation of the Model

Evaluation Measure	Testing Results			Validation Results		
Classifier	DT	KNN	SVM	DT	KNN	SVM
Accuracy	0.87	0.82	0.86	0.91	0. 85	0.88
Precision	0.85	0.80	0.81	0.87	0.83	0.84
Recall	0.86	0.79	0.81	0.89	0.82	0.84
F-Measure	0.86	0.69	0.75	0.89	0.75	0.82

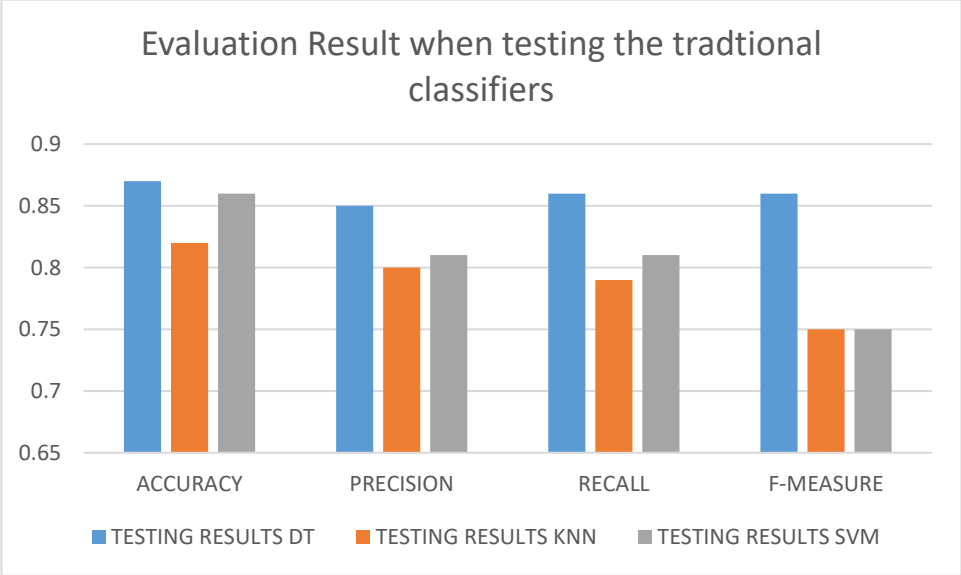


Figure 4.6: Results of the Traditional Classifiers after Testing

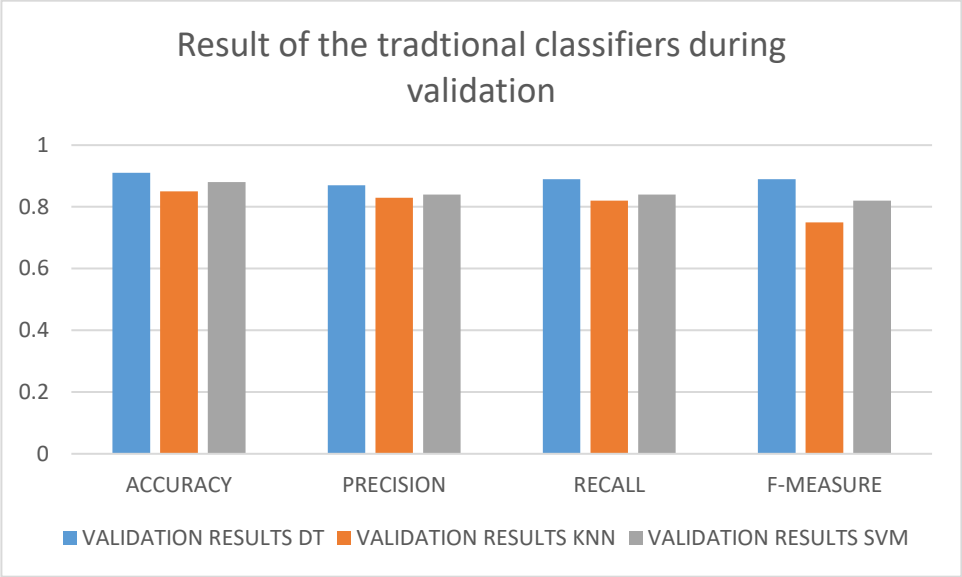


Figure 4.7: Results of the Traditional Classifiers after Validation

The validation process begun once the classification model was trained using 10-fold cross validation. The validation procedure is a critical step in the development of predictive models since it determines the predictive models' correctness. Table 4.4, Figure 4.5 and 4.6 shows the outcomes of evaluation using classification approaches (DT, KNN, and SVM) during the testing and validation phases.

As shown in Table 4.4, the proposed model achieves 91.5% accuracy during the validation phase. The proposed model outperformed Adejo, O. W., & Connolly, T. (2018) which had an accuracy of 82.2%. As a result, the validation phase result validates the proposed model's reliability.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS AND FUTURE RESEARCH

5.1 Conclusion

The primary goal of this research was to develop a student performance prediction model that leverages on longitudinal and temporal features using an ensemble model of machine learning algorithms. To improve the performance of single machine learning classifiers, ensemble approaches were applied. Three machine learning classifiers were utilized as traditional learning algorithms: Decision Tree, Support Vector Machine, and K-nearest Neighbor followed by three ensemble approaches. Bagging, Boosting and Random Forest to improve the performance of single-traditional classifiers. The accuracy of these various machine learning classifiers was 86.83% for SVM, DT and 91.5% for boosting ensemble approach.

Furthermore, DT, Random Forest, and XGBoost were used to identify the relevant factors required for prediction. It was shown that there was a correlation between the identified variables and that performance prediction model improvements can be achieved using ensemble boosting technique, with the resultant effect of increased accuracy, reduced error rate, and increased predictive efficiency. A summary of the findings showed that the ensemble classifier outperforms the traditional classifiers in terms of accuracy, precision, recall, and f-measure. The heterogeneous Stacking Ensemble Model (K-NN, SVM, DT) improves the homogeneous model by delivering 91% accuracy.

The findings of this research can be utilized to identify underperforming students and focus more attention on them in order to enhance their performance. This has the potential to increase the quality of higher education while also benefiting higher education institutions.

5.2 Recommendation

Based on the findings of this research, this new method of predicting student academic achievement should be pursued. The suggested ensemble model could be utilized as a tool for reliable and error-free prediction of student performance, as well as early detection of students in danger of attrition. However, the research's findings highlight the need for more detailed and generalized research in this area. This should include the addition of extra variables as well as the integration of variables from other sources before implementing ensemble methods.

5.3 Limitations

There are certain limitations to this research that should be mentioned. The research relies on publicly available datasets rather than a student dataset. Furthermore, the dataset was limited, with only a few hundred records. More data-driven research may yield more conclusive results. The majority of EDM researchers are currently reticent to release their study dataset for two reasons: The first is concerned with privacy, integrity, and legality; the second is with dataset collecting, which is a laborious, time-consuming, and costly operation. Based on a combination of privacy protection, economic impact, and scholarly ramifications, we urge that machine learning researchers disclose more educational datasets. This research employed offline data, but an increasing amount of online data remains untapped, allowing us to train the model to predict online student performance in real-time. A distinct educational dataset that our models can examine. Right away, if we are given a significant dataset, we may use the most recent big data technologies to construct a new model and validate the outputs. Furthermore, we can collect more data and apply deep learning approaches to improve model performance by incorporating new variables, such as assessing how students' use of social media affects their performance. Furthermore, additional experiments could be carried out by employing other machine learning techniques, such as clustering. This research made use of classification, DT, KNN, and ensemble approaches such as Boosting. Other methods, like clustering and deep neural networks, can be employed to increase perception of the significance of method

selection in classification or regression problems. Another area that may be addressed is the feature engineering process. The amount of feature engineering that can be done with limited data is also constrained.

5.4 Future Research

Predicting students' academic success has long been a source of concern for higher education organizations worldwide. The data acquired contains some secret knowledge that is being used to improve students' academic performance. A novel performance prediction model for students was proposed in this research, which was based on multiple data mining approaches and includes additional features known as behavioral features. The prediction model is tested using classifiers such as DT, KNN, and SVM. In addition, ensemble approaches were used to improve the performance of the classifiers. Out of Bagging, Boosting, and Random Forest based on the forward sequential selection was built, which was used to choose the most relevant features, the prediction model's accuracy was 91.5%, which was higher. In future efforts, analyze student data to find more factors that will identify students with lower success and performance will be done. Optimization techniques such as Differential Evolution, Genetic Algorithms, and others could also be used to improve student performance models in educational data mining.

REFERENCES

Adam, N. L., Rosli, N. H., & Soh, S. C. (2021, September). Sentiment analysis on movie review using Naïve Bayes. In *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 1-6). IEEE.

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61–75.

Adewale, A. M., Bamidele, A. O., & Lateef, U. O. (2018). Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach. *International Journal of Science and Technology Education Research*, *9*(1), 1-8.

Adil, S. H., Raza, K., & Hashmani, M. A. (2016). A hybrid cuckoo algorithm for lot scheduling problem using extended basic period and power of two policy. *Mehran University Research Journal of Engineering & Technology*, *35*(2), 229-238.

Adilah, M. T., Supendar, H., Ningsih, R., Muryani, S., & Solecha, K. (2020, November). Sentiment Analysis of Online Transportation Service using the Naïve Bayes Methods. In *Journal of Physics: Conference Series* (Vol. 1641, No. 1, p. 012093). IOP Publishing.

Agrawal, H., & Mavani, H. (2015). Student performance prediction using machine learning. *International Journal of Engineering Research and Technology*, *4*(03), 111-113.

Ajibade, S. S. M., Ahmad, N. B., & Shamsuddin, S. M. (2018, December). A data mining approach to predict academic performance of students using ensemble techniques. In *International Conference on Intelligent Systems Design and Applications* (pp. 749-760). Springer, Cham.

Ajibade, S. S. M., Dayupay, J., Ngo-Hoang, D. L., Oyebode, O. J., & Sasan, J. M. (2022). Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students. *Journal of Optoelectronics Laser*, *41*(6), 48-54.

Yahaya, C. A. C., Yaakub, C. Y., Abidin, A. F. Z., Ab Razak, M. F., Hasbullah, N. F., & Zolkipli, M. F. (2020, February). The prediction of undergraduate student performance in chemistry course using multilayer perceptron. In *IOP conference series: Materials science and engineering* (Vol. 769, No. 1, p. 012027). IOP Publishing. https://doi.org/10.1088/1757-899X/769/1/012027

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Alhiyafi, J., & Olatunji, S. O. (2017). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1–4. https://doi.org/10.1109/CCECE.2017.7946847

Altaf, S., Soomro, W., & Rawi, M. I. M. (2019). Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining. *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, 59–64.

Ambusaidi, M. A., He, X., Nanda, P., & Tan, Z. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE transactions on computers*, *65*(10), 2986-2998.

Ameen, A. O., Alarape, M. A., & Adewole, K. S. (2019). Students'academic performance and dropout prediction. *Malaysian Journal of Computing*, *4*(2), 278–303.

Amin, M. Z., & Ali, A. (2017, February). Application of multilayer perceptron (mlp) for data mining in healthcare operations. In *3rd International Conference on Biotechnology* (p. 9).

Amra, I. A. A., & Maghari, A. Y. (2017, May). Student's performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 18.

Anuradha, C., & Velmurugan, T. (2016, January). Feature selection techniques to analyse student acadamic performance using Naïve Bayes classifier. In *The 3rd international conference on small & medium business* (pp. 345-350).

Arif, M., Jahan, A., Mau, M. I., & Tummarzia, R. (2021). An Improved Prediction System of Students' Performance Using Classification model and Feature Selection Algorithm. *International Journal of Advances in Soft Computing & Its Applications*, *13*(1).

Arunachalam, A. S., & Velmurugan, T. (2018). Analyzing student performance using evolutionary artificial neural network algorithm. *International Journal of Engineering & Technology, 7*(2.26), 67-73.

Augusstine, N. M. J., & Samy, S. R. N. (2018). Smart healthcare monitoring system using support vector machine. *Australian Journal of Science and Technology*, *2*(3), 1-8.

Awad, M., & Ewais, A. (2018). *Prediction of General High School Exam Result Level Using Multilayer Perceptron Neural Networks*. *13*(10), 10.

Aybek, H. S. Y., & Okur, M. R. (2018). *Predicting Achievement with Artificial Neural Networks: The Case of Anadolu University Open Education System*. Same as 80

Babu, R. L., & Vijayan, S. (2016). Wrapper based feature selection in semantic medical information retrieval. *Journal of Medical Imaging and Health Informatics*, *6*(3), 802-805.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905-971.

Battineni, G., Chintalapudi, N., & Amenta, F. (2019). Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked*, *16*, 100200.

Bhogan, S., Sawant, K., Naik, P., Shaikh, R., Diukar, O., & Dessai, S. (2017). Predicting Student Performance Based On Clustering and Classification. *IOSR Journal of Computer Engineering (IOSR-JCE)*.

Bibireddy, K. (2017). *Design and Implementation of Real-time Student Performance Evaluation and Feedback System* (Doctoral dissertation).

Bikku, T. (2020). Multi-layered deep learning perceptron approach for health risk prediction. *Journal of Big Data*, *7*(1), 1-14.

Bithari, T. B., Thapa, S., & Hari, K. C. (2020). Predicting academic performance of engineering students using ensemble method. *Technical Journal*, *2*(1), 89-98.

Bohra, H., Arora, A., Gaikwad, P., Bhand, R., & Patil, M. R. (2017). Health prediction and medical diagnosis using Naive Bayes. *International Journal of Advanced Research in Computer and Communication Engineering*, *6*(4), 32-35.

Borges, V. R. P., Esteves, S., de Nardi Araújo, P., de Oliveira, L. C., & Holanda, M. (2018). Using Principal Component Analysis to support students' performance prediction and data analysis. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática Na Educação-SBIE)*, *29*(1), 1383.

Borucka, A. (2020). Logistic regression in modeling and assessment of transport services. *Open Engineering*, *10*(1), 26-34.

Burman, I., & Som, S. (2019). Predicting Students Academic Performance Using Support Vector Machine. *2019 Amity International Conference on Artificial Intelligence (AICAI)*, 756–759.

Chen, C. W., Tsai, Y. H., Chang, F. R., & Lin, W. C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, *37*(5), e12553.

Chong, P., Elovici, Y., & Binder, A. (2019). User authentication based on mouse dynamics using deep neural networks: A comprehensive study. *IEEE Transactions on Information Forensics and Security*, *15*, 1086-1101.

Cortez, P., & Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance.*

Domladovac, M. Comparison of Neural Network with Gradient Boosted Trees, Random Forest, Logistic Regression and SVM in predicting student achievement. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 211-216). IEEE.

El Bilali, H. (2019). The multi-level perspective in research on sustainability transitions in agriculture and food systems: A systematic review. *Agriculture*, *9*(4), 74.

Farissi, A., & Dahlan, H. M. (2019, September). Genetic algorithm based feature selection for predicting student's academic performance. In *International Conference of Reliable Information and Communication Technology* (pp. 110-117). Springer, Cham.

Figueira, A. (2016). Predicting grades by principal component analysis: A data mining approach to learning analyics. *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, 465–467.

Fithri, D., & Latifah, N. (2018, November). Prediction of Determination of Rice Farming Production using the Naïve Bayes Method. In *The 1st International Conference on Computer Science and Engineering Technology Universitas Muria Kudus*.

Fok, W. W., He, Y. S., Yeung, H. A., Law, K. Y., Cheung, K. H., Ai, Y. Y., & Ho, P. (2018). Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. *2018 4th International Conference on Information Management (ICIM)*, 103–106.

Francis, B. K., & Babu, S. S. (2019). Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *Journal of Medical Systems*, *43*(6), 162. https://doi.org/10.1007/s10916-019-1295-4

Gerritsen, L., & Conijn, R. (2017). Predicting student performance with Neural Networks. *Tilburg University, Netherlands*.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predictingstudents' performance using machine learning techniques. *IEEE Access*, *8*, 67899-67911.

Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2020). A wrapper-filter feature selection technique based on ant colony optimization. *Neural Computing and Applications*, *32*(12), 7839-7857.

Govindasamy, K., & Velmurugan, T. (2018). Analysis of student academic performance using clustering techniques. *International Journal of Pure and Applied Mathematics*, *119*(15), 309-323.

Gupta, G. K. (2014). *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd.

Haiyang, L., Wang, Z., Benachour, P., & Tubman, P. (2018, July). A time series classification method for behaviour-based dropout prediction. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)* (pp. 191-195). IEEE.

Hamoud, A. (2016). Selection of best decision tree algorithm for prediction and classification of students' action. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, *16*(1), 26-32.

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.

Hasheminejad, S. M., & Sarvmili, M. (2019). S3PSO: Students' performance prediction based on particle swarm optimization. *Journal of AI and Data Mining*, *7*(1), 77–96.

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing.

Hasudungan, R. A. (2020). Using MDA to Improve Naïve Bayes Classification for Students Performance Prediction. *JSE Journal of Science and Engineering*, *1*(2), 65-70.

Heuer, H., & Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. *DeLFI 2018-Die 16. E-Learning Fachtagung Informatik*

Hui, K. H., Ooi, C. S., Lim, M. H., Leong, M. S., & Al-Obaidi, S. M. (2017). An improvedwrapper-based feature selection method for machinery fault diagnosis. *PloS one*, *12*(12), e0189143.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience*, *2018*.

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning (IJET)*, *14*(14), 92–104.

Ising, C., Venegas, C., Zhang, S., Scheiblich, H., Schmidt, S. V., Vieira-Saecker, A., & Heneka, M. T. (2019). NLRP3 inflammasome activation drives tau pathology. *Nature*, *575*(7784), 669-673.

Iyanda, A. R., Ninan, O. D., Ajayi, A. O., & Anyabolu, O. G. (2018). Predicting student academic performance in computer science courses: A comparison of neural network models. *International Journal of Modern Education and Computer Science (IJMECS)*, *10*(6), 1–9.

Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, *77*(5), 5198-5219.

Jawthari, M., & Stoffová, V. (2021). Predicting students' academic performance using a modified kNN algorithm. *Pollack Periodica*, *16*(3), 20-26.

Jithender, B., Upendar, K., & Nickhil, C. (2019). Chapter-8 Post-Harvest Quality of Fresh Produce. *Chief Editor Dr. RK Naresh*, *19*, 129.

Junyou, Z., Fanyu, W., & Shufeng, W. (2018). Application of support vector machine in bus travel time prediction. *Int. J. Syst. Eng*, *2*, 21-25.

Kadambande, A., Thakur, S., Mohol, A., & Ingole, A. M. (2017). Predicting students' performance system. *International Research Journal of Engineering and Technology*, *4*(5), 2814-2816.

Kanchana, J. S., & Sujatha, S. (2016). Integrating heterogeneous agriculture information using naive Bayes and FCA. *Advances in Natural and Applied Sciences*, *10*(10 SE), 308-312.

Kapur, D. R. (2018). Factors influencing performance and job satisfaction of teachers in secondary schools in India. *Research Gate*, 1-25.

Karasu, S., Altan, A., Bekiros, S., & Ahmad, W. (2020). A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. *Energy*, *212*, 118750.

Karthikeya, H. K., Sudarshan, K., & Shetty, D. S. (2020). Prediction of Agricultural Crops using KNN Algorithm. *International Journal of Innovative Science and Research Technology*, *5*(5), 1422-1424.

Kok, Z. H., Shariff, A. R. M., Alfatni, M. S. M., & Khairunniza-Bejo, S. (2021). Support Vector Machine in Precision Agriculture: A review. *Computers and Electronics in Agriculture*, *191*, 106546.

Kumar, S., Jain, A., & Mahalakshmi, P. (2018). Enhancement of Healthcare Using Naïve Bayes Algorithm and Intelligent Datamining of Social Media. *International Journal of Applied Engineering Research*, *13*(6), 4109-4112.

Kumar, S., Kumar, V., & Sharma, R. K. (2019). Rice yield forecasting using support vector machine. *International Journal of Recent Technology and Engineering*, *8*(4), 2588-2593.

Kumari, P., Jain, P. K., & Pamula, R. (2018, March). An efficient use of ensemble methods to predict students' academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Limbu, J., & Sah, S. (2019). Prediction on Student Academic Performance Using Hybrid Clustering Algorithm. *LBEF Research Journal of Science, Technology and Management 1*(1), 1-22

Liu, C., & Chao, Z. (2021, October). Supervised learning and unsupervised learning on music data with different genres. In *2021 IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom)* (pp. 7-12). IEEE.

Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D., & Liu, H. (2020). A review of android malware detection approaches based on machine learning. *IEEE access*, *8*, 124579-124607.

Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, *6*(3), 703-715.

Liu, J., Lin, Y., Lin, M., Wu, S., & Zhang, J. (2017). Feature selection based on quality of information. *Neurocomputing*, *225*, 11-22.

Ma, L., Li, M., Gao, Y., Chen, T., Ma, X., & Qu, L. (2017). A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation. *IEEE Geoscience and Remote Sensing Letters*, *14*(3), 409-413.

Maghari, A. (2018). Prediction of student's performance using modified KNN classifiers. In *Alfere, SS, & Maghari, AY (2018). Prediction of Student's Performance Using Modified KNN Classifiers. In The First International Conference on Engineering and Future Technology (ICEFT 2018)* (pp. 143-150).

Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014). Evaluating performance and dropouts of undergraduates using educational data mining. In *Proceedings of the Twenty-Ninth Symposium on Applied Computing*.

Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review*, (33), 235-266.

Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. *Croatian Operational Research Review*, *7*(2), 367–388.

Mohanapriya, K., & Balasubramani, M. (2019, October). Recognition of Unhealthy Plant Leaves Using Naive Bayes Classifier. In *IOP Conference Series: Materials Science and Engineering* (Vol. 561, No. 1, p. 012094). IOP Publishing.

Murty, M. N., & Devi, V. S. (2011). *Pattern recognition: An algorithmic approach*. Springer Science & Business Media.

Nagahisarchoghaei, M., Dodd, J., Nagahi, M., Ghanbari, G., & Poudyal, S. (2020, October). Analysis of a Warranty-Based Quality Management System in the Construction Industry. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)* (pp. 1-5). IEEE.

Naraei, P., Abhari, A., & Sadeghian, A. (2016, December). Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. In *2016 Future Technologies Conference (FTC)* (pp. 848-852). IEEE.

Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R., & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, *8*(2), 221–228.

Noercholis, A., & Zainuddin, M. (2020). Comparative Analysis of 5 Algorithm Based Particle Swarm Optimization (PSO) for Prediction of Graduate Time Graduation. *MATICS: Jurnal Ilmu Komputer dan Teknologi Informasi (Journal of Computer Science and Information Technology)*, *12*(1), 1-9.

Nosseir, A., & Fathy, Y. M. (2020). A Mobile Application for Early Prediction of Student Performance Using Fuzzy Logic and Artificial Neural Networks. *iJIM*, *14*(2), 4-18.

Nugroho, A., Riady, O. R., Calvin, A., & Suhartono, D. (2020). Identification of student academic performance using the KNN algorithm. *Engineering, MAthematics and Computer Science Journal (EMACS)*, *2*(3), 115-122.

Nurhayati, O. D., Bachri, O. S., Supriyanto, A., & Hasbullah, M. (2018). Graduation prediction system using artificial neural network. *International Journal of Mechanical Engineering and Technology*, *9*(7), 1051–1057.

Ogwoka, T. M., Cheruiyot, W., & Okeyo, G. (2015). A model for predicting students' academic performance using a hybrid of K-means and decision tree algorithms. *International Journal of Computer Applications Technology and Research*, *4*(9), 693-697.

Oloruntoba, S. A., & Akinode, J. L. (2017). Student academic performance prediction using support vector machine. *International Journal of Engineering Sciences and Research Technology*, *6*(12), 588-597.

Opara, C. C., Eze, U. F., & Oleji, C. P. (2020). Hybrid Data Mining Model for Knowledge Discovery on Students Academic Performance. *Hybrid Data Mining Model for Knowledge Discovery on Students Academic Performance*, *47*(1), 9–9.

Osmanbegović, E., Suljić, M., & Agić, H. (2014). Determining dominant factor for student's performance prediction by using data mining classification algorithms. *Tranzicija*, *16*(34), 147-158.

Paas, W., & Groot, J. C. (2017). Creating adaptive farm typologies using Naive Bayesian classification. *Information Processing in Agriculture*, *4*(3), 220-227.

Pandey, M., & Taruna, S. (2014). A comparative study of ensemble methods for students' performance modeling. *International Journal of Computer Applications*, *103*(8).

Pedraza, D., & Beruvides, M. (2016). The Relationship between Course Assignments and Academic Performance: An Analysis of Predictive Characteristics of Student Performance. *2016 ASEE Annual Conference & Exposition Proceedings*, 26217. https://doi.org/10.18260/p.26217

Perez, J. G., & Perez, E. S. (2021). Predicting Student Program Completion Using Naïve Bayes Classification Algorithm. *International Journal of Modern Education & Computer Science*, *13*(3).

Phiophuead, T., & Kunsuwan, N. (2019). Logistic regression analysis of factors affecting travel mode choice for disaster evacuation. *Engineering Journal*, *23*(6), 399-417.

Prabha, S. L., & Shanavas, A. M. (2015). Performance of Classification Algorithms on Students' Data–A Comparative Study. *International Journal of Computer Science and Mobile Applications*, *3*(9), 1-8.

Rachburee, N., & Punlumjeak, W. (2015, October). A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. In *2015 7th international conference on information technology and electrical engineering (ICITEE)* (pp. 420-424). IEEE.

Ragab, M., Abdel Aal, A. M., Jifri, A. O., & Omran, N. F. (2021). Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques. *Wireless Communications and Mobile Computing*, *2021*, 1-9.

Ramaswami, G. S., Susnjak, T., Mathrani, A., & Umer, R. (2020, December). Predicting Students Final Academic Performance using Feature Selection Approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-5). IEEE.

Ren, K., Fang, W., Qu, J., Zhang, X., & Shi, X. (2020). Comparison of eight filter-based feature selection methods for monthly streamflow forecasting–three case studies on CAMELS data sets. *Journal of Hydrology*, *586*, 124897.

Rezaei, A., Yahya, S. I., Noori, L., & Jamaluddin, M. H. (2022). Designing high-performance microstrip quad-band bandpass filters (for multi-service communication systems): a novel method based on artificial neural networks. *Neural Computing and Applications*, 1-15.

Saifudin, A., & Desyani, T. (2020, March). Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes. In *Journal of Physics: Conference Series* (Vol. 1477, No. 3, p. 032007). IOP Publishing.

Salah Hashim, A., Akeel Awadh, W., & Khalaf Hamoud, A. (2020). Student Performance Prediction Model based on Supervised Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*, *928*, 032019. https://doi.org/10.1088/1757-899X/928/3/032019

Santoso, L. W. (2020). *Predicting student performance in higher education using multi-regression models* [PhD Thesis]. Petra Christian University.

Satyanarayana, A., & Nuckowski, M. (2016). Data mining using ensemble classifiers for improved prediction of student academic performance.

Sawant, T. U., Pol, U. R., & Patankar, P. S. (2019). Student Placement Prediction Model Using Gradient Boosted Tree Algorithm. *Journal of Emerging Technologies and Innovative Research 6*(5)-499

Sharma, R., & Ranjan, P. (2021, December). A Review: Machine Learning Based Hardware Trojan Detection. In *2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)* (pp. 1-4). IEEE.

Siddiqi, M. A., & Pak, W. (2020). Optimizing filter-based feature selection method flow for intrusion detection system. *Electronics*, *9*(12), 2114.

Sikder, M. F., Uddin, M. J., & Halder, S. (2016). Predicting student's yearly performance using neural network: A case study of BSMRSTU. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 524–529.

Smirani, L. K., Yamani, H. A., Menzli, L. J., & Boulahia, J. A. (2022). Using ensemble learning algorithms to predict Student failure and enabling customized educational paths. *Scientific Programming*, *2022*.

Sokkhey, P., & Okazaki, T. (n.d.). *Hybrid Machine Learning Algorithms for Predicting Academic Performance*.

Solomon, D. (n.d.). *Predicting Performance and Potential Difficulties of University Student using Classification: Survey Paper*. 6.

Stewart, R., & Ermon, S. (2017, February). Label-free supervision of neural networks with physics and domain knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 2576-2582).

Strecht, P., Cruz, L., Soares, C., & Mendes-Moreira, J. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.

Suh, M. H., & Jeong, M. (2022). Development of Bus Routes Reorganization Support Software Using the Naïve Bayes Classification Method. *Sustainability*, *14*(8), 4400.

Sule, B. O., & Saporu, F. W. O. (2015). A logistic regression model of students' academic performance in university of Maiduguri, Maiduguri, Nigeria. *Mathematical Theory and Modeling*, *5*(10), 124-136.

Sultana, J., Rani, M. U., & Farquad, M. A. H. (2019). *Student's Performance Prediction using Deep Learning and Data Mining Methods*. *8*(1), 4.

Tillmanns, S., & Krafft, M. (2021). Logistic regression and discriminant analysis. In *Handbook of market research* (pp. 329-367). Cham: Springer International Publishing.

Urrutia-Aguilar, M. E., Fuentes-García, R., Martínez, V. D. M., Beck, E., León, S. O., & Guevara-Guzmán, R. (2016). Logistic regression model for the academic

performance of first-year medical students in the biomedical area. *Creative Education*, *7*(15), 2202.

Utku, A., & Kaya, S. K. (2022). Multi-layer perceptron based transfer passenger flow prediction in Istanbul transportation system. *Decision Making: Applications in Management and Engineering*.

Verma, S. K., Thakur, R. S., & Jaloree, S. (2017). Fuzzy association rule mining-based model to predict students' performance. *International Journal of Electrical and Computer Engineering (IJECE)*, *7*(4), 2223–2231.

Vora, D. R., & Rajamani, K. (2019). A hybrid classification model for prediction of academic performance of students: A big data application. *Evolutionary Intelligence*. https://doi.org/10.1007/s12065-019-00303-9

Wang, G., & Kim, J. (2016, November). The prediction of traffic congestion and incident on urban road networks using naive bayes classifier. In *Australasian Transport Research Forum (ATRF), 38th*.

Wang, P., Hafshejani, B. A., & Wang, D. (2021). An improved multilayer perceptron approach for detecting sugarcane yield production in IoT based smart agriculture. *Microprocessors and Microsystems*, *82*, 103822.

Wanjau, S. K. (2016). Data mining model for predicting student enrolment in stem courses in higher education institutions.

Wasif, M., Waheed, H., Aljohani, N. R., & Hassan, S. U. (2019). Understanding student learning behavior and predicting their performance. In *Cognitive Computing in Technology-Enhanced Learning* (pp. 1-28). IGI Global.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). Delving deeper into MOOC student dropout prediction. *arXiv preprint arXiv:1702.06404*.

Xiao, W., Ji, P., & Hu, J. (2021). RnkHEU: A Hybrid Feature Selection Method for Predicting Students' Performance. *Scientific Programming*, *2021*.

Xin, J., & Chen, S. (2016). Bus dwell time prediction based on KNN. *Procedia engineering*, *137*, 283-288.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indones. J. Electr. Eng. Comput. Sci*, *16*(3), 1584-1592.

Yahaya, C. A. C., Yaakub, C. Y., Abidin, A. F. Z., Ab Razak, M. F., Hasbullah, N. F., & Zolkipli, M. F. (2020, February). The prediction of undergraduate student performance in chemistry course using multilayer perceptron. In *IOP Conference Series: Materials Science and Engineering* (Vol. 769, No. 1, p. 012027). IOP Publishing

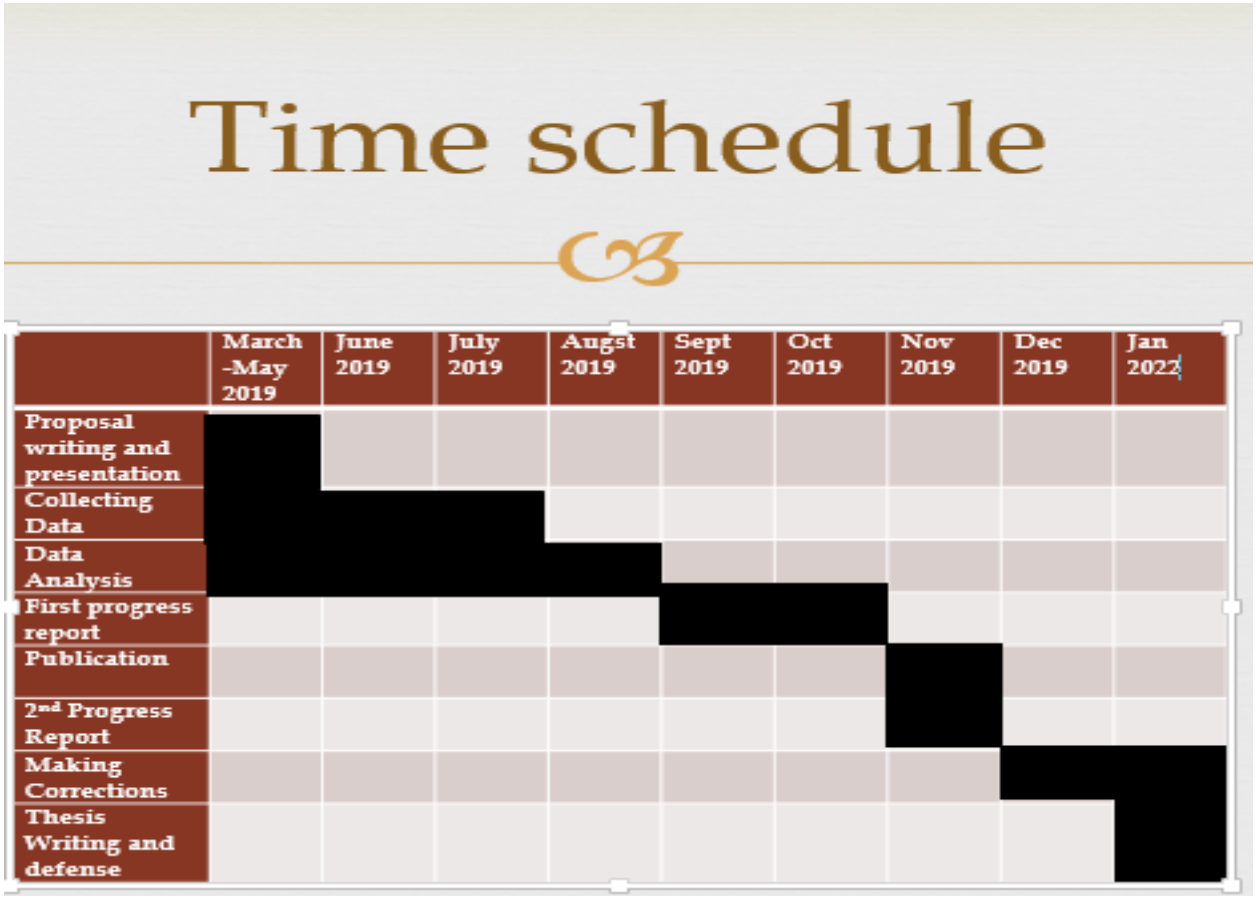
Zaffar, M., Hashmani, M. A., & Savita, K. S. (2017, November). Performance analysis of feature selection algorithm for educational data mining. In *2017 IEEE Conference on Big Data and Analytics (ICBDA)* (pp. 7-12). IEEE.

Zaffar, M., Savita, K. S., Hashmani, M. A., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students' academic performance. *Int. J. Adv. Comput. Sci. Appl*, *9*(5), 541-549.

Zulfiker, M. S., Kabir, N., Al Amin Biswas, P. C., & Rahman, M. M. (n.d.). *Predicting Students' Performance of the Private Universities of Bangladesh using Machine Learning Approaches*

APPENDICES

Appendix I: Activity Schedule (Gantt chart)



Appendix II: Cost and Materials

Budget

Item /Activity	Description	Total Cost
Transport	From Home To Juja then School	10,000
Publication	Two publications@15,000	30,000
Internet access		50,000
Data analysis		40,000
Printing (proposal, publications and thesis)	@10,000	30,000
Field assistance	@20,000	60,000
Miscellaneous	10% of the total cost	22,000
Total		242,000

Appendix III: Feature Selection Code

```
data.rename(columns ={'G3':'Class'}, inplace=True)
```

```
data
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	Class
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
644	MS	F	19	R	GT3	T	2	3	services	other	...	5	4	2	1	2	5	4	10	11	10
645	MS	F	18	U	LE3	T	3	1	teacher	services	...	4	3	4	1	1	1	4	15	15	16
646	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1	1	5	6	11	12	9
647	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	6	10	10	10
648	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	4	10	11	11

649 rows × 33 columns

```
print(data.columns)
```

```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_excel('Data.xlsx')

data
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
644	MS	F	19	R	GT3	T	2	3	services	other	...	5	4	2	1	2	5	4	10	11	10
645	MS	F	18	U	LE3	T	3	1	teacher	services	...	4	3	4	1	1	1	4	15	15	16
646	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1	1	5	6	11	12	9
647	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	6	10	10	10
648	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	4	10	11	11

```
onehot,values = create_one_hot(data)
onehot
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	...	activities-yes	nursery-yes	nursery-no	higher-yes	higher-no	internet-no	internet-yes	ror
0	18.0	4.0	4.0	2.0	2.0	0.0	4.0	3.0	4.0	1.0	...	0.0	1.0	0.0	1.0	0.0	1.0	0.0	
1	17.0	1.0	1.0	1.0	2.0	0.0	5.0	3.0	3.0	1.0	...	0.0	0.0	1.0	1.0	0.0	0.0	1.0	
2	15.0	1.0	1.0	1.0	2.0	0.0	4.0	3.0	2.0	2.0	...	0.0	1.0	0.0	1.0	0.0	0.0	1.0	
3	15.0	4.0	2.0	1.0	3.0	0.0	3.0	2.0	2.0	1.0	...	1.0	1.0	0.0	1.0	0.0	0.0	1.0	
4	16.0	3.0	3.0	1.0	2.0	0.0	4.0	3.0	2.0	1.0	...	0.0	1.0	0.0	1.0	0.0	1.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
644	19.0	2.0	3.0	1.0	3.0	1.0	5.0	4.0	2.0	1.0	...	1.0	0.0	1.0	1.0	0.0	0.0	1.0	
645	18.0	3.0	1.0	1.0	2.0	0.0	4.0	3.0	4.0	1.0	...	0.0	1.0	0.0	1.0	0.0	0.0	1.0	
646	18.0	1.0	1.0	2.0	2.0	0.0	1.0	1.0	1.0	1.0	...	1.0	1.0	0.0	1.0	0.0	1.0	0.0	
647	17.0	3.0	1.0	2.0	1.0	0.0	2.0	4.0	5.0	3.0	...	0.0	0.0	1.0	1.0	0.0	0.0	1.0	
648	18.0	3.0	2.0	3.0	1.0	0.0	4.0	4.0	1.0	3.0	...	0.0	0.0	1.0	1.0	0.0	0.0	1.0	

649 rows × 59 columns

```
#for colum in onehot.columns:
    #print(onehot[colum].dtype)
```

```
#for colum in onehot.columns:
    #print(onehot[colum].dtype)
```

```
def create_normalization(data,normalizationtype="minmax"):
    df = data.copy() # create a copy of the data
    df = df.drop(columns=["CLASS"])
    df.select_dtypes(include=['float','int'])
    if(normalizationtype == "minmax"):
        dicti = {}
        for x in df.columns:
            mini = df[x].min()
            maxi = df[x].max()
            df[x] = [(y - mini)/(maxi-mini) for y in df[x]]
            dicti[x] =("minmax",mini,maxi)
        df = df.assign(CLASS=data["CLASS"].values)
        return df,dicti
    elif(normalizationtype =="zscore"): #The function is written for the zscore normalization though not used.
        dict2 = {}
        for x in df.columns:
            meane = df[x].mean()
            stdi = df[x].std()
            df[x] = df[x].apply(lambda x:(x-meane)/stdi)
            dict2[x] =("zscore",meane,stdi)
        df = df.assign(CLASS=data["CLASS"].values)
        return df,dict2
```

```
normalized, normalizer = create_normalization(onehot, normalizationtype="zscore")
```

```
X = normalized.iloc[:,:-1]
y = normalized.iloc[:,-1]


from sklearn.svm import SVC

model = SVC(kernel='linear')


model.fit(X,y)

pd.Series(abs(model.coef_[0]), index=X.columns).nlargest(10).plot(kind='barh')

# feat_importances = pd.Series(model.coef_, index=X.columns)
# feat_importances.nlargest(58).plot(kind='bar')
plt.show()
```