BIRZEIT UNIVERSITY

Faculty of Graduate Studies and Research كلية الدراسات العليا والأبحاث

Predicting Students' Performance Using Machine Learning Techniques: Faculty of Engineering at Birzeit University as a Case Study

Prepared by:

Ahmad Ibrahem Bearat

1195488

Committee:

Name	Signature
Dr. Hassan Abu Hassan	
Dr. Tareq Sadeq	
Dr. Majdi Mafarja	

Submitted in partial fulfillment of the requirements for the "Master's Degree in Applied Statistics and Data Science" from the faculty of Graduate Studies at Birzeit University-Palestine

Birzeit University

Dec 2022

جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Graduate Studies and Research كلية الدراسات العليا و الأبحاث

Predicting Students' Performance Using Machine Learning Techniques: Faculty of Engineering at Birzeit University as a Case Study

Prepared by:

Ahmad Ibrahem Bearat

1195488

Committee:

Name	Signature
Dr. Hassan Abu Hassan	
Dr. Tareq Sadeq	
Dr. Majdi Mafarja	

Submitted in partial fulfillment of the requirements for the "Master's Degree in Applied Statistics and Data Science" from the faculty of Graduate Studies at Birzeit University-Palestine

Birzeit University

Dec 2022

Acknowledgment

To God and my Family.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

DT	Decision Tree
RF	Random Forest
K-NN	K-Nearest Neighbour
ANN	Artificial Neural Network
SVM	Support Vector Machine
LR	Logistic Regression
NB	Naïve Base
DM	Data Mining
Twjeehi	High School Diploma
GPA	Cross Point Average
CGPA	Cumulative Cross Point Average
BZU	Birzeit University
MICE	Multivariate Imputation by Chained Equations
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
ML	Machine learning
DM	Data Mining
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
SMO	Sequential Minimal Optimization
MLP	Multi-Layer -Perceptron

Abstract

It is known that students' academic performance is the core of the educational process in universities. The educational institutions as well as students themselves constantly seek to raise their academic performance. The results of students' academic performance depend on several factors and variables, including personality, demographic, social, economic, educational, core curriculum, academic staff, and many other variables.

The following research looks into these factors and variables and focuses on three main objectives. The first one identifies the most significant variables that affect students' academic performance in The Faculty of Engineering at Birzeit University. Secondly, it identifies the most significant variables that affect students' retention. The last one is building several models using machine learning techniques to predict students' academic performance and make comparisons between them. The machine learning algorithms that were utilized in this study were: (DT, RF, SVM, K-NN, and ANN).

Furthermore, the study data was collected via a questionnaire. It was distributed as a stratified sample that included all disciplines of the Engineering Faculty at Birzeit University during the second semester of 2022. The sample size was 397 students. The data was analyzed using SPSS and R. The research found that the RF algorithm was the finest algorithm for predicting students' academic performance among the used algorithms. In addition, the research found that the most significant factors that affected students' academic performance were: Physics 1 grade, Calculus 1 grade, Number of not passed courses, Calculus 2 grade, A 12th-grade student's average in school, Student's grade in Twjeehi, Physics 2 grade, The 10th-grade student's average in school, Student's year at university, Student's absence in class. Also, the factors that affected students' retention were: Student's GPA, A 10th-grade student's average in school, Physics 1 grade, Student's grade in Twjeehi, Calculus 1 grade, The 12th-grade student's average in school, Calculus 2 grade, Student's computer experience, Physics 2 grade, Student's satisfaction with academic staff.

ملخص

يعتبر الأداء الأكاديمي للطلاب في الجامعات جوهر العملية التعليمية. إذ تسعى المؤسسات التعليمية وكذلك الطلاب أنفسهم بشكل دائم إلى رفع الأداء الأكاديمي ويعتمد الأداء الأكاديمي للطلاب على عدة عوامل ومتغيرات، منها: الشخصية، والديموغرافية، والاجتماعية، والاقتصادية، وبيئة التعليم والمتغيرات المرتبطة بها من مواد تدريسية وهيئة أكاديمية وغيرها الكثير.

هناك ثلاثة أهداف رئيسية لهذا البحث، هي: دراسة المتغيرات التي تؤثر في أداء الطلاب الأكاديمي في كلية الهندسة في جامعة بيرزيت وتحديد أهمها، وكذلك دراسة المتغيرات التي تؤثر في بقاء الطلاب في التخصصات وتحديدها. وبناء عدة نماذج باستخدام تقنيات تعلم الآلة للتنبؤ بأداء الطلاب الأكاديمي وعمل مقارنات فيما بينها، حيث استخدمت عدة خوارزميات في هذه الدراسة وهي (DT, RF, SVM, K-NN, and ANN).

جمعت بيانات الدراسة عن طريق استبانة، وقد وزعت على عينة طبقية تشمل كل تخصصات كلية الهندسة في جامعة بيرزيت خلال الفصل الدراسي الثاني 2022، وكان حجم العينة 397 طالب. وقد استخدم برنامج R وSPSS لتحليل البيانات. حيث حددت خوارزمية RF كأفضل خوارزمية للتنبؤ بأداء الطلاب الأكاديمي من بين الخوارزميات المستخدمة. وكانت العوامل الأكثر ارتباطاً بأداء الطلاب الأكاديمي هي: (الفيزياء 1، والتفاضل والتكامل 1، وعدد المقررات التي لم يتم النجاح فيها، والتفاضل والتكامل 2، ومعدل الطالب في الصف الثاني عشر في المدرسة، وعلامة الطالب في التوجيهي، والفيزياء 2، ومعدل الطالب في الصف العاشر في المدرسة، والسنة الدراسية في الجامعة، ومقدار غياب الطالب عن المحاضرات). والعوامل التي تؤثر في بقاء الطلاب في التخصصات هي: (المعدل التراكمي للطالب، ومعدل الطالب في الصف العاشر في المدرسة، والفيزياء 1، ومعدل الطالب في التوجيهي، والتفاضل والتكامل 1، ومعدل الطالب في الصف الثاني عشر في المدرسة، والتفاضل والتكامل 2، مهارة الطالب في استخدام الحاسوب، والفيزياء 2، رضى الطالب عن الطاقم الأكاديمي).

1 Chapter One

1.1 Introduction

The use of technology and electronic devices has recently become very common. Technology is indispensable in all fields. Yet it can be a burden. This burden manifests in the difficulty of dealing with the massive amount of data (BD) from devices by humans. However, there is an opportunity that lies in the ability of computers to analyze this big data. Furthermore, computers can detect patterns and draw useful information from this data since it is the language that computers can handle (Pojon, 2017).

Shingari et al. (2017) defined data mining as data archaeology. He pointed out that it is a technique for extracting hidden patterns and relationships from big data. Also, Han et al. (2011) discussed the massive amount of data collected daily by technology revelation. Thus, in recent years, this field has gained importance and it has been classified as one of the most crucial modern data sciences. This is because big data constitutes a complex problem to solve without using learning analysis and it is done through computerized programs. Furthermore, these programs are used by computers that gain a high ability to analyze such data quickly and accurately which provides researchers with valuable results. Therefore, modern science is concerned with the methods of collecting and analyzing data. Also, in addition, a massive amount of information can be extracted from analyzing this data.

The researcher can use this data in what is known as machine learning to develop several models and use these models to predict and classify some of the needed variables. However, several ways are used in dealing with big data. The most common are prediction, clustering, classification, and relationship (Madnaik, 2020).

Moreover, data mining technology is applied in many fields, including economic, medical, and educational ones. The database for these and other areas increases over time. The importance of analyzing this data using data mining and machine learning techniques lies here.

Data mining and machine learning methods are used in analyzing big education data and thus contribute significantly to extracting functional patterns about students' academic performance and are used for prediction. (Shingari et al., 2017).

Ünal (2020) said it’s possible to predict students' academic performance by many variables, including personal, social, economic, environmental, demographic, emotional, and psychological, in addition to other variables like the educational environment, materials, educational tools, and many other variables. Consequently, identifying these variables and studying their impact on students' performance can help manage and develop the educational process. Furthermore, the prediction of students' performance using these variables is essential in increasing their success and raising their performance by making a greater effort to provide appropriate support for low achievers. This is an important goal for academic institutions.

One of the essential methods used to predict students' academic performance in higher education is machine learning techniques. This prediction is made based on the several mentioned variables as the inputs for these models. Moreover, the importance of these variables lies in using them to build prediction models by learning from the big data available in the databases of educational institutions (Altabrawee et al., 2019).

In this research, the researcher seeks to build several models to predict academic performance by determining the best variables and factors that affect the students’ performance. This goal will be achieved by using machine-learning techniques. The research subsumed independent variables (listed in the methodology) and one dependent variable (Academic Performance). Additionally, after building the machine learning models, and based on the findings, the researcher will make a comparison between the results of these models. Recommendations will also be made at the end some of them may contribute to the development of students’ academic performance at the university and may be generalized to other universities.

1.2 Research Problem

Due to the lack of local studies that deal with predicting students' academic performance in higher education using machine learning techniques in Palestine, this research aims at identifying the most significant variables that affect students' academic performance. The researcher will use these variables to build several prediction models using machine learning algorithms and compare the accuracy of these models to determine the one with the best predictive ability.

1.3 Research Objectives

The main objective of this study is to build several models that predict students' academic performance using machine learning algorithms based on several variables and attributes of the students.

Additionally, this study seeks to identify the most significant variables that affect students' academic performance. The researcher can summarize the objectives of the study as follows:

1. Identifying the best variables and factors which can affect a students' performance.
2. Offering several prediction models for students' performance using machine learning techniques.
3. Comparing these different machine learning models in terms of accuracy and obtaining the best model among them.
4. Formulating recommendations for improving students' academic performance.
5. Identifying the variables that predict the decision of some engineering students to change their majors.

1.4 Research Questions

1. What are the most significant variables and factors that affect students' academic performance at the Faculty of Engineering at BZU?
2. What is the best predictive model among ML algorithms?
3. What are the most significant variables and factors that force engineering students at BZU to change their major?

1.5 Research Significance

To improve students' academic performance, the researcher should determine the variables and factors that affect this performance (Yassein et al., 2017).

Similarly, a fundamental problem for many students at universities is changing their specialization. Therefore, student retention is an essential issue in higher education. It is necessary to know students' academic performance in advance to solve this problem. This is possible by focusing on low achievers and providing them with more support. This requires identifying the variables and factors that significantly affect students' academic performance.

Additionally, many variables may significantly affect students' academic performance. This research seeks to identify these variables and study their impact on students' academic performance and provide recommendations for university administration and decision-makers in the engineering faculty to set policies and procedures that may contribute to developing academic performance based on these factors.

Moreover, predictive models can give advanced predictions of students' academic performance after they complete their first year of study at the university. This may contribute to students' retention. The prediction models can also be used by academic advisors and department heads to discover the level of their current students and take proactive steps in teaching strategies. Moreover, this research may be useful in applying appropriate academic strategies within the academic

programs of the departments in particular and the faculty of engineering in general.

Finally, it provides those who are interested in higher education research or planners in this field with recommendations that are based on scientific results that can raise the quality of the educational process. (Alturki et al., 2020).

2 Chapter Two

Background and Literature Review

2.1 Background

2.1.1 Machine Learning Algorithms

After the modern scientific revolution, machines cannot be dispensed with all aspects of life. However, the main difference between a machine and a human is intelligence. A human can think and accordingly make the appropriate decision (Bonaccorso, 2017). As for the machine, it is not intelligent. It is managed by humans to analyze the data, and then humans make the decision. With rapid technological development, the issue of artificial intelligence began to emerge, and scientists were wondering whether it was possible to make machines that can think. Besides, this led to rapid development in computer science in this field, such as image processing, pattern recognition, clustering, variable interpretation, and prediction.

Murphy (2012) defined machine learning as a set of techniques and methods that can detect patterns in the data. After that, he used those patterns to predict future data, which enables us to make the appropriate decision. Livingston (2005) mentioned that machine learning can be defined as machines that can learn without direct programming. Moreover, the philosophy of machine learning is to sample data to represent the entire population. Also, this data is usually divided into two sets. The first set of data for machine learner development is called training data, and the second set of data for evaluation is called testing data. Additionally, there are two main ways of machine learning: supervised, and unsupervised. Maldonado (2019) defined Supervised Machine Learning as building a model in which the training or input data for the model contains the correct or desired output. He also defined Unsupervised Machine learning as building a model in which the training or input data for the model does not contain the correct or desired output, just segmentation or grouping based on its similarities. Murphy (2012) pointed out the aim of supervised learning is mapping an input x to output y, given a labelled data set called

training data. While if y is nominal or categorical then the researcher has a classification issue, and if the researcher has real or ordinal value, then the researcher has a regression issue. Also, the second type of machine learning is unsupervised learning, where the researcher just has the input data, and the need is to find the common pattern in these data.

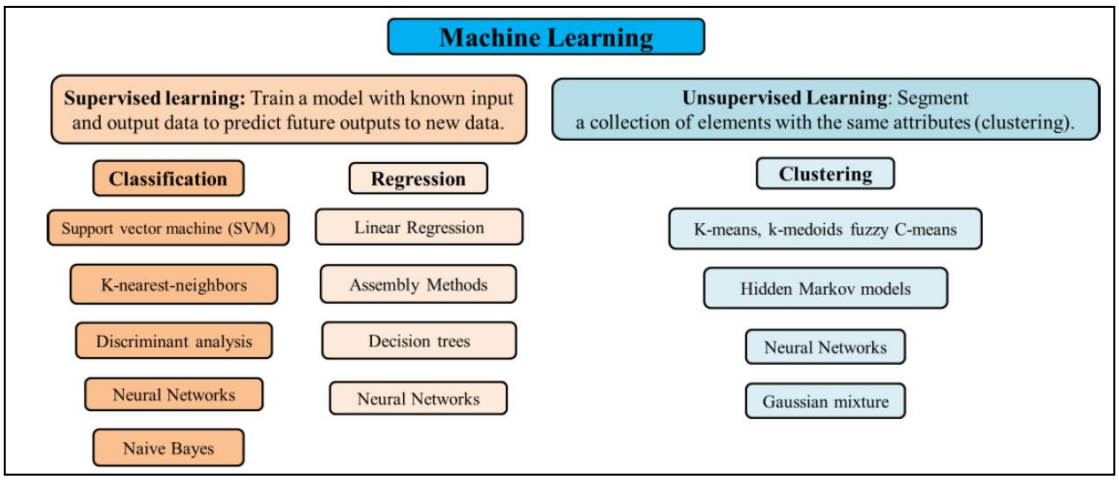


Figure 1: Supervised and Unsupervised learning, By Maldonado (2019)

Figure 1 describes some of the machine learning definitions and types.

2.1.2 Decision Tree (DT)

A decision tree, like its name, is a tree diagram used to define a course of action. Each branch in the tree represents a possible decision. According to Tarik et al. (2021) decision tree is classified as supervised learning. Furthermore, it handles regression and classification models and aims to stratify a population into homogeneous groups based on different distinct characteristics.

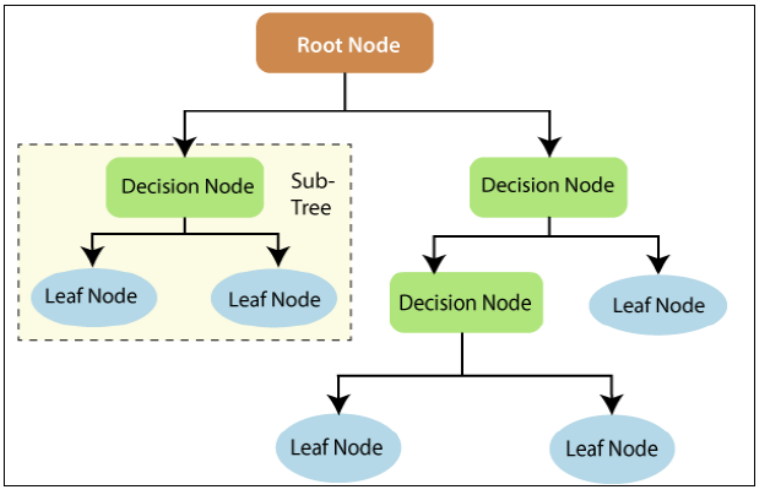


Figure 2: Decision Tree By Charbuty and Abdulazeez (2021)

Charbuty and Abdulazeez (2021) pointed out that a DT consists of three main types of nodes, as shown in Figure 2, the root node, the internal node or decision node, and the leaf node. The root node represents the community without any incoming edges. Decision nodes represent variables with incoming and outgoing edges. Finally, the outer nodes (leaves) represent the final decision. Also, there will be a class for the classification, and a number for regression. According to Charbuty and Abdulazeez (2021), DT has been widely used in image processing and pattern identification. Additionally, the researcher mentioned that there are many types of decision tree algorithms, such as CART, ID3, C4.5, CHAID, and QUEST. These different algorithms are varied in dependent variable type, pruning, and splitting technique.

There are two significant definitions in the decision tree which are entropy and information gain. Entropy is a measure of randomness or homogeneity, and information gain is the measure of the decrease in entropy while the splitting is done for the data. Also, he pointed out that decision tree algorithms like ID3 use information gain to select the candidate variable at each split while building the tree (Mitchell, 1997).

There are some advantages to using a decision tree. First, it is easy to follow and understand since it is self-exploratory. Also, it can handle both numerical and nominal data. Additionally, it can be used with data having errors and missing values. Furthermore, it is also considered a nonparametric method, so there is no need to test any

assumption before analyzing the data. On the other hand, there are two main disadvantages to the decision tree. These are overfitting problems and sensitivity to training data (Resende & Drummond, 2018).

2.1.3 Random Forest (RF)

According to Yeşilkanat (2020), a random forest machine learner is a supervised learning algorithm used for both regression and classification. The researcher explained RF consists of several decision trees which are used to classify input data. Furthermore, the final result for RF is the average of all trees for regression, and it is the majority output for classification. Additionally, RF is one of the most common and best-accurate machine learning algorithms. Resende and Drummond (2018) defined RF as a machine learning algorithm that predicts using a multi-decision tree. Furthermore, the researchers pointed out that for creating each decision tree resampling bootstrap approach is used, and the variable at each node is selected randomly for splitting. Also, as mentioned before, the majority for classification and the average for regression give the final result. Additionally, Yeşilkanat (2020) mentioned that the data is divided into two main parts, training data, and it is called in-bag data used for learning. The second is testing data, used for validation and it is called the out-of-bag data. Resende and Drummond (2018) said there are two RF tuning, one is for the number of decision trees and the other one is for the number of randomly selected variables or features in each split.

The most significant advantage of the RF is the correction of the overfitting problem in the decision trees, and the second advantage is that it has very good accuracy compared to other machine learning algorithms (Yeşilkanat, 2020). However, Resende and Drummond (2018) added that RF graphical representation is not available as in the case of DT. Also, RF is a little bit slow since it builds several decision trees and this is one of the biggest disadvantages. Also, since RF is non-parametric there is no need for formal distributional assumptions.

2.1.4 K- Nearest Neighbour (K-NN)

k-Nearest Neighbour is a supervised machine learning algorithm. K-NN is a memory-based algorithm that classifies objects based on the closest features (Hastie et al., 2009). However, it classifies that the new data points out based on measuring the distances between the data and its closest points. Usually, the number of points taken is K. Anuradha and Velmurugan (2015) said the K-NN algorithm determines which points in the data are the same when making a prediction. Also, it chooses the data points closest to the new observation and takes the most common ones among those. That is why it is called the k- nearest neighbour algorithm. Moreover, Anuradha and Velmurugan (2015) summarized that an odd number K should be chosen, then the data closest to the point classified can be determined, but its amount must be equal to the value K. Finally, for classification, the majority of the nearest data is chosen as a final class, and the average is taken in the regression. Cunningham and Delany (2020) pointed out that K-NN is also called a lazy learning technique since it takes time to calculate the distances between the points.

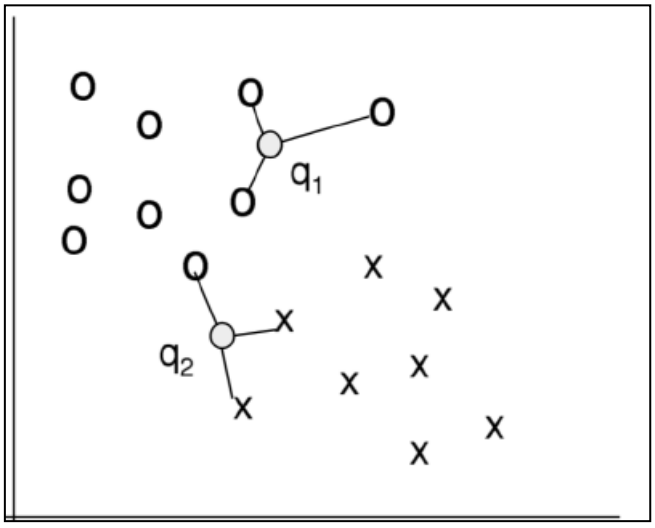


Figure 3: KNN, by Cunningham and Delany (2020)

Figure 3 shows a two-dimension space with two variables and three nearest neighbors. It points out that all of the nearest for q1 is O

so it is classified as O. for q2 since the nearest neighbors are two X and one O, the researcher can consider the majority so it is classified as X.

The value K should be wisely selected since the chosen K will change the output result. There are different ways to select K, one of them is by the $\sqrt{n}$ where n is the total number of points.

K-NN is easy to implant and debug. Additionally, K-NN is effective with noise reduction technology and gives higher accuracy with noisy data. On the other hand, K-NN has poor time performance for large data, and K-NN is very sensitive to redundant variables (Cunningham & Delany, 2020). Hastie et al. (2009) pointed out that K-NN gives a good prediction for several classification problems like handwriting digits, and satellite image processing. Other wide uses for K-NN are anomaly detection, text mining, and recommendation systems (Amazon, Netflix).

2.1.5 Artificial Neural Network (ANN)

A neural network is a supervised machine learning algorithm that works like the way of human neurons. However, it processes information instead of signals (Nasser & Abu-Naser, 2019). ANN can be used not only for classification but also for regression.

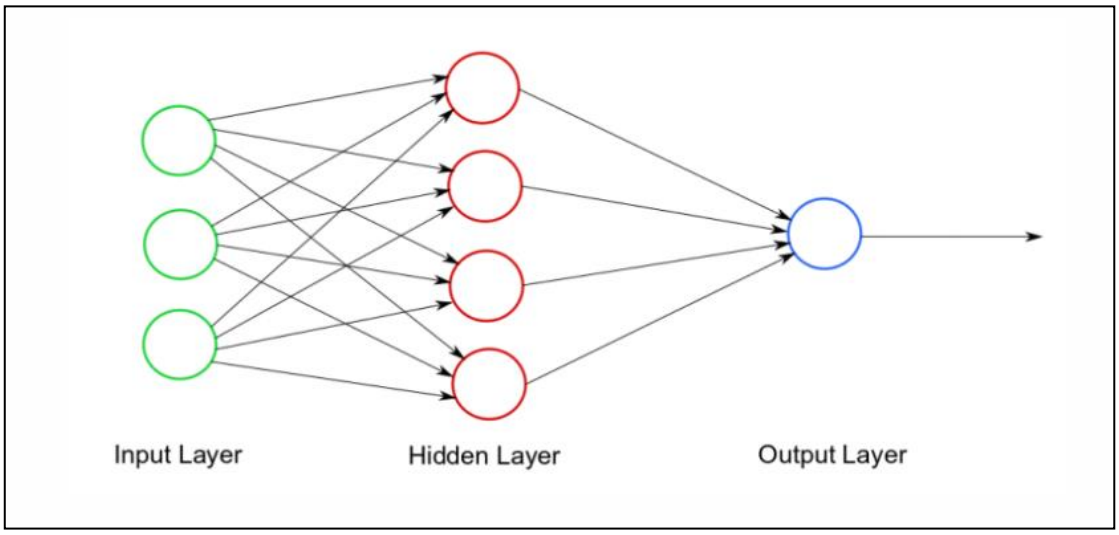


Figure 4: ANN, By Shah (2021)

Nasser and Abu-Naser (2019) indicated that there are three main layers of ANN, as shown in Figure 4. They explained that the input layer represents the incoming information that income to the neuron. Usually, their number is equal to the number of variables. Also, its job is to duplicate the received value to the all-hidden layers. Furthermore, maybe one or more hidden layers are a connection between the input and output layers. In addition, it is especially important to determine the weight of each node in the neural structure. So, it is called a weighted connection since it multiplies each value in the hidden node by its weight then all the nodes add to gather to produce a single number. Finally, the output layer depends on two main things the weight between the hidden and output and the activating function. The result for the output layer is the prediction for the dependent variable, it combines all values of the hidden layers and returns one output value.

Additionally, there are many advantages to ANN. The most important are: it can handle linear and nonlinear data, classified as non-parametric so there is no need to test any assumption; is very effective in higher-dimension space; ANN reduces the over and underfitting problem since it has a powerful tuning option (Otchere et al., 2021). Another strength of ANN compared to other machine learning algorithms is that it limits the effect of outliers. On the other hand, there are some disadvantages for ANN. One of them is the poor model performance with noisy and overlapped data. Mitchell (1997) mentioned that ANN is widely used in handwritten character recognition, spoken word recognition, and face recognition. Also, it is very suitable for complex data, such as cameras and microphone data. Additionally, it is very suitable for data with missing values and errors.

2.1.6 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that can handle both regression and classification analysis. Huang et al. (2018) defined SVM as a machine learning algorithm that can maximize separating margins, it creates a boundary between different classes which is called a hyperplane. As shown in figure5.

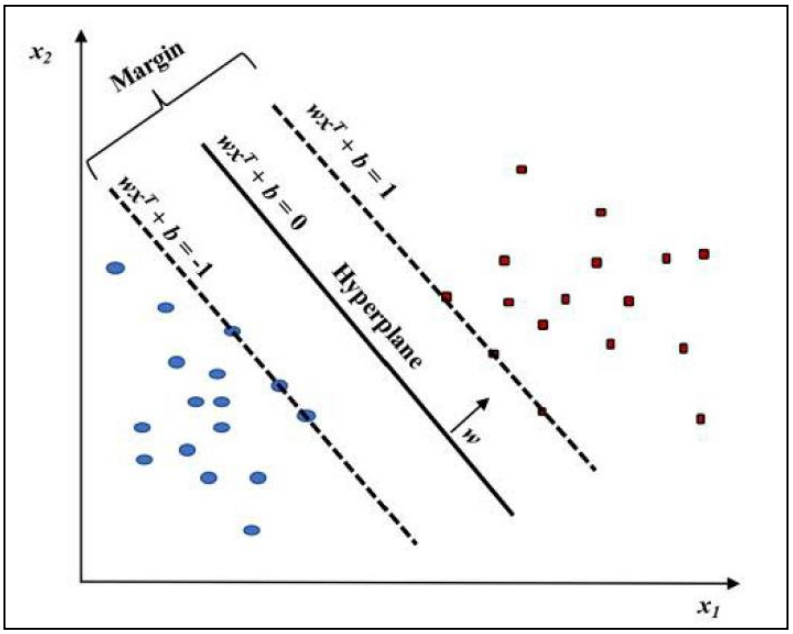


Figure 5: Linear SVM model, Two classes, By Huang et al. (2018)

This hyperplane tries to be as far as possible from the support vector where the support vector is the extreme point in each class. As shown in Figure 5.

SVM aims to separate the class, so different classes are as far as possible from each other. Also, the distance between the support vector and the hyperplane should be as far as possible. Additionally, the margin is the distance between the two-support vector, then by finding the largest distance margin, the researcher can find the optimal hyperplane.

The kernel function is the process of transformation of the data from one dimension to another, to achieve the best separation between the different classifications (Huang et al., 2018). The kernel function works to give the best separation between the different classes. Consequently, if the researcher cannot separate two classes in a good way, the kernel function will move the data to a higher dimension space, and it will find a support vector that can classify the data. Also, there are too many kernel functions, and choosing the best one is difficult and affects the final result.

According to Otchere et al. (2021) SVM is a very powerful algorithm for detecting patterns in big data. It is effective in high-

dimensional spaces and it is a very stable algorithm. Furthermore, SVM can efficiently handle non-linear data using the kernel trick. Also, using SVM prevents the over-fitting problem. Finally, the disadvantages of SVM are as follows: it does not perform very well with noisy and overlapped data sets, it has extensive memory requirements, it requires a long training time, and choosing the kernel function is very difficult as well. Lee (2007) pointed out SVM algorithm is widely used in several financial applications, such as credit cards, time-series data prediction, insurance claims, and fraud detection.

Finally, this research will predict the dependent variable students' performance using regression models based on the following machine learning algorithms (DT, K- NN, RF, SVM, and ANN). The output for these regression models will be the student's GPA.

This research will use different machine learning algorithms to build predictive models for students' performance. Then a comparison will be made between the results of these models using some evaluated criteria, like RMSE and accuracy.

2.2 Literature Review

Artificial intelligence is a science that depends on mathematical equations and models that lets the machine take the decision instead of the people (Tarik et al., 2021). Additionally, using this method improves the experience and rapid scientific development in several areas. One of these areas is predicting academic performance.

One of the most critical challenges facing the educational process is to reduce students' attrition and raise their academic performance at the same time (Adejo & Connolly, 2018). Therefore, prior knowledge of students' academic performance will save time and money for students who may find themselves at some point at a dead end where

they are unable to continue with their studies or forced to change their major.

Educational institutions strive to achieve the best quality education for their students. To accomplish this, it is necessary to identify the variables affecting academic performance (Yassein et al., 2017). Furthermore, according to Altabrawee et al. (2019), the goal of any educational institution is to increase the efficiency of the educational process and increase students' knowledge. Therefore, the researcher used machine learning techniques to determine low achievers in addition to several classification models that were built to predict students' performance. Among the essential results added by the decision tree model the following are the best variables used for classification: "Computer Grades-Course1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency". Finally, the researcher found that the best machine-learning algorithm to predict students' performance is Artificial Neural Network and Decision Tree.

Ünal (2020) said it is essential to predict students' academic performance because it will raise their achievement, develop the efficiency of the educational process, and contribute to strategic planning. Moreover, the researcher used several personal, economic, social, demographic, educational, and environmental variables. Also, the researcher used the Wrapper method to select the best variable for machine learning algorithms. Furthermore, the researcher used decision tree (DT), random forest (RF), and naïve base (NB) machine learning algorithms to predict students' performance in five-level and two-level (Binary: P/F) grading. Finally, he found that the best accuracy is by using the RF algorithm.

Educational data mining is vital in extracting a pattern from old data and using it for predicting future data. This could help in reducing failure and give spatial attention as required for students (Madhumitha S, 2018). Furthermore, Madhumitha S (2018) used several personal, economic, demographics, social and educational variables to predict students' performance. The final results found that the most correlated variables with students' performance were "student's 10th, 12th, degree

marks in each semester, assignment, study hours, parent's education, income".

Predicting students' performance is essential for developing educational services and vital for low achievers to know their abilities (Tran et al., 2017). Additionally, he proposed different regression machine learning models for predicting students' performance, and these models were built using machine learning algorithms LR, ANN, DT, and SVM. Finally, he found that the most significant variable was grade point average (GPA) for the previous semester and the most predictive model was using the SVM algorithm.

Madnaik (2020) predicted students' academic performance using several purely academic variables, and other non-academic variables. Moreover, he used these variables and the following machine learning algorithms to build his prediction models (RF, DT, K-NN, and NB). Furthermore, he found that Random Forest gives the best prediction. Moreover, the most significant variables that affected students' performance were the variables that were related to participation in the class. On the other hand, social variables like parents' education and jobs had a minor but important effect on students' performance.

Shahiri and Husain (2015) pointed out the importance of systems to analyze and monitor students' academic performance, analyze the rapid increase in big educational data, and use this data to predict the future performance of new students. Furthermore, he predicted students' academic performance by using classification data mining techniques and focused on the variables that affect students' performance based on previous studies. Additionally, the researcher found that the most significant variable was cumulative grade point average (CGPA). Finally, the prediction model built was based on the following machine learning algorithms (ANN, DT, SVM, K-NN, and NB) and was sorted according to the best prediction results.

Amrieh et al. (2016) discussed the importance of extracting hidden patterns from big educational data and its result in improving the educational process and methods. Also, he predicted students' academic performance by building prediction models using machine learning algorithms (ANN, NB, and DT). Furthermore, he found that

the best prediction method is using the ANN algorithm. Also, the researcher applied ensemble methods (Bagging, Boosting, and RF) which found that this method improves the accuracy, and using the ANN algorithm still gives the best accuracy. Additionally, the researcher used several variables: demographical, academic, parents' participation in the learning process, and student behavior to build the prediction models. Lastly, he found that the most significant variable was students' behavior in class.

Tarik et al. (2021) mentioned that after high school, students are led to choose their major of study at the university based on their abilities. Therefore, students are usually confused and afraid of making a wrong decision that will affect their personal and professional life in the future. Thus, the role of academic guidance and the development of systems to predict students' academic performance is vital which enables them to choose their course of study and future profession. Additionally, the researcher used three regression machine learning algorithms (LR, DT, and RF) to predict students' academic performance. He found that the best machine learning model is the RF algorithm since it gave much more accuracy than other algorithms.

Tran et al. (2017) proposed different models using machine learning algorithms (LR, ANN, DT, SVM) for predicting students' performance. The researcher found the best prediction model using a support vector machine algorithm (SVM).

Education data mining is essential in identifying low-level students to determine steps to address and raise their performance (Acharya & Sinha, 2014). So, Acharya and Sinha (2014) proposed prediction models, using the following machine learning algorithms: Decision Tree (C4.5), Sequential Minimal Optimization (SMO), Naïve Base classifier (NB), 1-Nearest Neighbourhood (1-NN), and Multi-Layer -Perceptron (MLP). Furthermore, to predict CGPA the researcher used 14 independent variables. Finally, he found that DT (C4.5) was the best predictive model since it gave the best accuracy.

3 Chapter Three

3.1 Research Methodology

3.1.1 Introduction

This study worked to determine the best variables that affect students' academic performance at the Engineering Faculty of Birzeit University and the best variables that affect students' retention. The researcher predicts students' performance based on several machine-learning algorithms. Finally, a comparison of Regression models was conducted to identify the best predictive one and identify the best variables that affect students' performance. Furthermore, classification models are used to identify the best variables that affect students' retention. As shown in Figure 6.

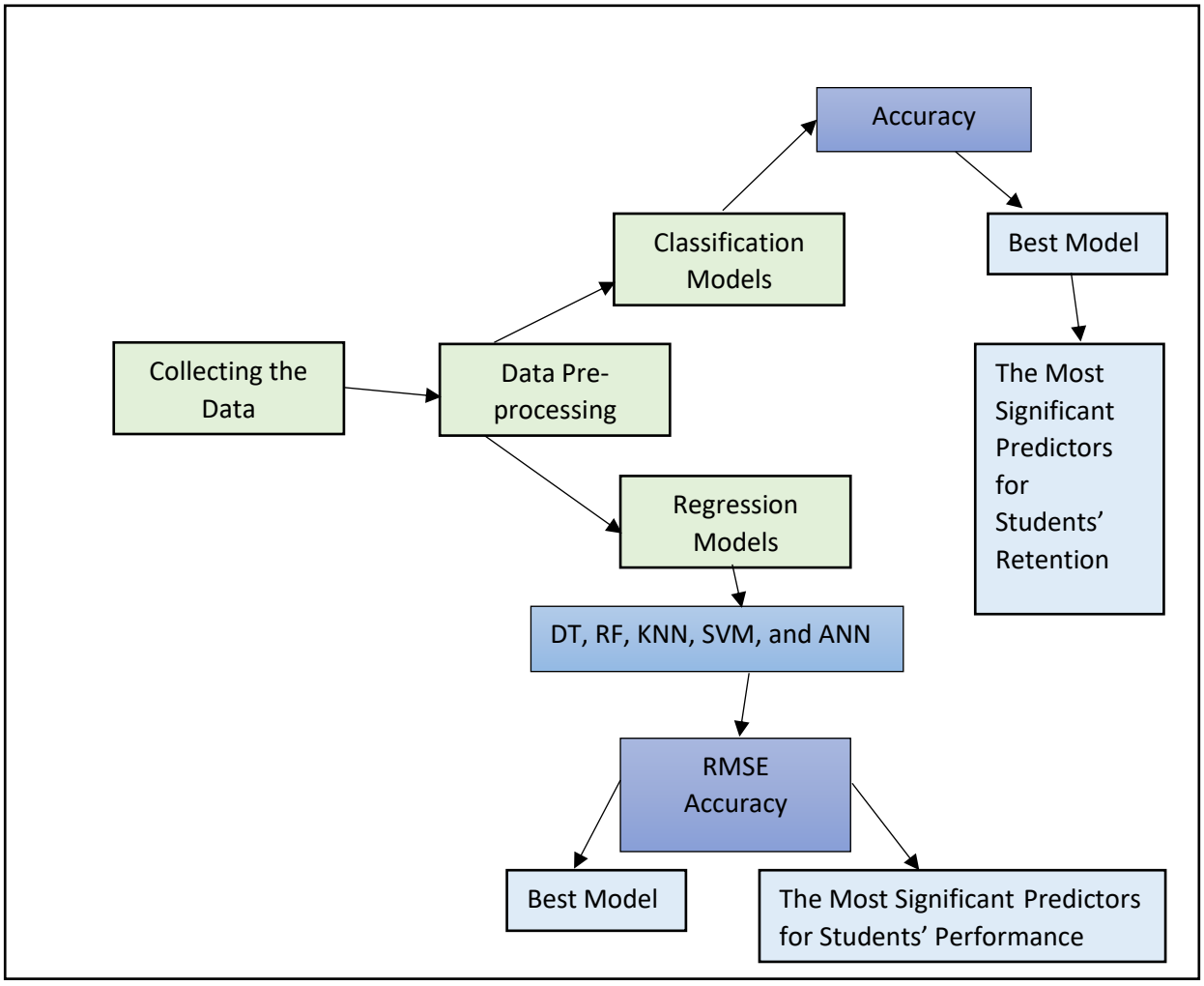


Figure 6: Proposed Approach

3.1.2 Pilot study

Before the main survey, a pilot study of 49 students was made, to calculate the response rate, and the questionnaires were examined. The following are the findings:

1. Response rate equal to 85.7%. as shown in equation 3:

$$Response\ Rate = \frac{I}{I + P + R} x\ 100\% = \frac{42}{42 + 5 + 2} x\ 100\% = 85.7\% \quad \text{Equation 1}$$

I: Completed interviews

$$P: Partial\ completed$$

$$R: Refused\ to\ participate$$

$$Nonresponse\ Rate = 1 - Response\ Rate = 14\%$$

2. Based on the exanimation, some variables were removed from the study (students' social status, first language, high school branch, students' disability, and the availability of the internet at home) since there is no variability in these variables.

3.1.3 Population

The population of the study consists of the students of the faculty of engineering at Birzeit University. These students are studying for a bachelor's degree in the academic year 2021/2022 and have completed their first academic year. The population size of this study is (2617) students.

3.1.4 Sample

The Engineering Faculty at Birzeit University has 2617 students studying in the 2021/2022 academic year and have completed their first academic year. Since the researcher cannot take every student's point of view, given that the number is large, a representative random sample of students was drawn for this purpose.

The researcher used the (Yamane, 1967) equation to calculate the sample e as shown in equation 4:

$$n = \frac{N}{1 + Ne^2}$$ Equation 2

Where n is the sample size, N is the population size and e are the margins of error. Let e=0.05 and N=2617

Then our required sample size is 347 as shown in equation 5:

$$n = \frac{N}{1 + Ne^2} = \frac{2617}{1 + 2617 * 0.05^2} = 347$$ Equation 3

Based on the pilot study conducted before the main survey, it was found that the non-response rate was 14% (4% refusal rate, 10% partially completed). Hence the adjusted sample size will be 403 as shown in equation 6:

$$adjusted\ n = \frac{347}{(1 - 0.14)} = 403$$ Equation 4

Furthermore, a stratified random sample used to collect the sample

for different majors as shown in table 1:

Table 1: Sample of students by major in the Engineering Faculty

	Engineering Faculty Major	**Number of students**	**Percentage**	**Roundup Sample Size**
1	Computer Science	609	0.232709	94
2	Computer systems engineering	692	0.264425	106
3	Mechatronics Engineering	131	0.050057	21
4	Mechanical Engineering	168	0.064196	26
5	Electrical Engineering	220	0.084066	34
6	Civil Engineering	334	0.127627	52
7	Architectural Engineering	355	0.135652	55
8	Urban planning and design	108	0.041268	17
	Total	2617	1	405

In parallel to collecting primary data from the students themselves by paper questionnaire, the researcher tried to explore the possibility of collecting the data on students from an online questionnaire. The response rate was very small. All social media platforms, including university platforms, were used without any progress.

3.1.5 Dataset

The data consists of one dependent variable, the students' performance. The definition of students' performance is the measurement of students' achievement across the years of study at the university (Grade point average – GPA), Furthermore, there are two types of achievements for students: student's GPA (an interval scale variable) and student rating (an ordinal scale variable). In this research, a student's GPA is used to predict students' performance.

This research will study 50 independent variables including personal, social, economic, environmental, demographic, emotional, and psychological studied variables, in addition to variables related to the educational environment, materials, educational tools, and many other variables based on previous studies and as listed in table 2.

Table 2: Study Variables Description

ID	Variable Name	Description	Domain
1.	Sex	Student's sex (Binary)	Female:0, male:1
2.	Address	Student's current address (Binary)	City:1, Village:2, Refuge Camp:3
3.	Sacom	Student's accommodation (Binary)	Dorms:0, with family:1
4.	Fsize	Family size (Numeric)	
5.	Smag	Student's major (Nominal)	Computer Science:1, Computer Engineering:2, Mechatronics Engineering:3, Mechanical Engineering:4,

			Electrical Engineering:5, Civil Engineering:6, Architectural Engineering:7, Urban Planning, Design:8, Environmental Engineering:9
6.	Syer	Student's year at university (Numeric)	
7.	Ssch	Type of school attended at the higher secondary level (Binary)	Private school:0, public school:1
8.	STLang	The language of study at school (Nominal)	Arabic:0, English:1, Others:3
9.	S10tha	The 10^{th}-grade student's average in school (Numeric)	
10.	S12tha	A 12^{th}-grade student's average in school (Numeric)	
11.	PhysG1	Physics 1 grade (Numeric)	
12.	PhysG2	Physics 2 grade (Numeric)	
13.	CalcG1	Calculus 1 grade (Numeric)	
14.	CalcG2	Calculus 2 grade (Numeric)	
15.	STgrad	Student's grade in Twjeehi (Numeric)	
16.	Sgpa	**Student's GPA is** the dependent variable (Numeric)	
17.	Sarbl	Level of student's Arabic entrance exam	ARAB 135: ARAB 136:2
18.	Sengl	Level of student's English entrance exam	A1:1, A2:2, B1:3, B2:4, C:5
19.	Snpass	Number of not passed courses(Numeric)	
20.	Schois	Reasons for choosing the major (Numeric)	Self-interest:1, family choice:2, reputation in the labor market:3, easy to get high grades:4, Tawjihi GPA:5, Other:6
21.	Sspt	Do you play any sports regularly? (Binary)	No:0, yes:1
22.	Sjob	Does the student have a job (Binary)	No:0, yes:1
23.	Mjob	Does your mother work? (Binary)	No:0, yes:1
24.	Fjob	Does your father work? (Binary)	No:0, yes:1
25.	Fsupport	Family education support (Binary)	No:0, yes:1

26.	Eclass	Extra paid classes (Binary)	No:0, yes:1
27.	Hedu	Wants to complete higher education (Binary)	No:0, yes:1
28.	GPAonl	Did online courses significantly increase your GPA (Binary)	No:0, yes:1
29.	Ssponser	Does the student have any sponsorship?	No:0, yes:1
30.	Smchn	Did you change your study major? (Binary)	No:0, yes:1
31.	Simchn	Do you intend to change your major of study? (Binary)	No:0, yes:1
32.	Hstat	Does the student have health issues (Binary)	No:0, yes:1
33.	Ftime	Free time after university (Binary)	No:0, yes:1
34.	Medu	Mother's education (Numeric)	Master or Higher:3, Bachelor: 2, College:1, High School or less:0
35.	Fedu	Father's education (Numeric)	Master or Higher:3, Bachelor: 2, College:1, High School or less:0
36.	Ttime	How long does it take you to reach the university (Numeric)	<15 min:1, 15 to 30min:2, 30min to 1 h:3, >1h:4
37.	Stime	daily study time (Numeric)	Just for exams:1, 1h:2, 2-3h:3, ≥4h:4
38.	Scex	Student's computer experience (Numeric)	Very Good:5, Good:4, Average:3, Poor:2, Very Poor:1
39.	Sfs	Student's financial status (Numeric)	Very Good:5, Good:4, Average:3, Poor:2, Very Poor:1
40.	SMsat	Student's satisfaction with the major of study (Numeric)	Very satisfied:5, Satisfied:4, Neither:3, Dissatisfied:2, Very dissatisfied:1
41.	SLsat	Students' satisfaction with the available logistics. (Numeric)	Very satisfied:5, Satisfied:4, Neither:3, Dissatisfied:2, Very dissatisfied:1
42.	SASsat	Student's satisfaction with academic staff (Numeric)	Very satisfied:5, Satisfied:4, Neither:3, Dissatisfied:2, Very dissatisfied:1

43.	SCRsat	Student satisfaction with the curriculum and resources (Numeric)	Very satisfied:5, Satisfied:4, Neither:3, Dissatisfied:2, Very dissatisfied:1
44.	SONLsat	Student's satisfaction with online teaching (Numeric)	Very satisfied:5, Satisfied:4, Neither:3, Dissatisfied:2, Very dissatisfied:1
45.	Timgout	Going out with friends (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5
46.	Timsm	Time on social media (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5
47.	Timtv	Time on TV (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5
48.	Timss	Time spent with social service (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5
49.	Timfr	Time is given for free reading (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5
50.	Timps	Student's political involvement (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5
51.	Sabsn	Student's absence in class (Numeric)	Never:1, Rarely:2, Sometimes:3, Frequently:4, Always:5

3.1.6 Data pre-processing

The researcher used several data mining techniques, such as data cleaning, and outlier detection. These prepared student data to be used by machine learning algorithms for predicting students' academic performance.

Several data entry errors were modified from questionnaires when reviewed. After correcting the wrong data entry, the second step is to check univariate and multivariate outliers. Therefore, to identify univariate outliers a Zscore was calculated for all variables, and to check multivariate outlier Mahalanobis distance was calculated using SPSS.

By finding the Zscore for all variables it turns out that there are outliers in the following variables (Fsize, Snpass, STLang, and Simchn) since their Zscore value is larger than 3. But The researcher decided to keep this data and not delete it because it seems logical.

By finding the Mahalanobis distance and comparing it with the multiplication of the number of predictors and the threshold. it turns out that there are no multivariate outliers since all Mahalanobis distances are less than 125 (the threshold used is 2.5 and the number of predictors is 50).

Furthermore, since the researcher had data more than the required sample size (347) so, the missing value has been deleted and the data size becomes 397 after this deletion. Also, the missing value could be imputed using Multivariate Imputation by Chained Equations in R (MICE) for future work, several methods could be used in this package for imputation like the PMM method for a continuous variable, and the LDA method for a categorical variable.

The R package (mice) imputes multivariate missing data by a set of conditional models according to the different variable types (Van Buuren & Groothuis-Oudshoorn, 2011).

3.1.7 Data Analysis

Splitting the data: The data set is randomly divided into two main data sets: training data and testing data.

Training data is a subset of 70% of the main data set for training the models and parameter estimation.

Testing data is a subset of 30% of the main data set for models' evaluation.

3.1.8 Models' Evaluation

For prediction models, and according to Hodson (2022) Root Mean Square Errors (RMSE) are widely used for model evaluation, so this research applied it as one of the criteria for model evaluation and the second criteria is the mean absolute percentage error (MAPE). Khair et al. (2017) pointed out that MAPE calculates how many errors are in predicting compared with the actual value. On the other hand, for classification models, Sensitivity, Specificity, and classification Accuracy were used for model evaluation.

RMSE and MAPE used for model evaluation are calculated by the following equations:

$$RMSE = \sqrt{\frac{\sum_1^n (Actual - Predicted\ )^2}{n}} \qquad \text{Equation 5}$$

$$MAPE = \frac{100\%}{n} \sum_1^n \left| \frac{(Actual - Predicted\ )}{Actual} \right| \qquad \text{Equation 6}$$

$$Accuracy = 1 - MAPE \qquad \text{Equation 7}$$

Where

n: Sample Size

Actual: is the real data

Predicted: is the predicted data

Classification model diagnostic tests (sensitivity and specificity) depend on positive versus negative test results (Genders et al., 2012). The sensitivity and specificity used for model evaluation are as shown in equation 10, and equation 11:

$$Sensitivity = \frac{TP}{TP + FN} \qquad \text{Equation 8}$$

And

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad \text{Equation 9}$$

Murphy (2012) defined True Positive as when both actual data and predicted data were positive. Also, False Positive when the actual data was negative but the predicted data was positive. Additionally, True Negative is when both the actual and predicted values were negative. Finally, False Negative when the actual value was positive but the predicted value was negative.

Finally, the accuracy of the classification model can be calculated using the equation.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Equation 10}$$

4 Chapter Four

Results and Discussions

4.1 The most related attributes that affect students' performance

In this research, twenty-five variables were found to have bivariate significant relationships with students' performance. The researcher believes that this large number of significant relationships was achieved after a closer look into previous studies and choosing the variables accordingly. Additionally, this large number of significant variables will have a positive effect on the models' accuracy, which predicts academic performance. A bivariate Person correlation between variables was calculated using R and the results are shown in Table 3 below:

Table 3: Correlation Table

ID	Variable	Correlation Coefficient	P-value
1.	PhysG1	0.61	0
2.	CalcG1	0.605	0
3.	CalcG2	0.515	0
4.	Snpass	-0.511	0
5.	STgrad	0.479	0
6.	S12thg	0.446	0
7.	PhysG2	0.407	0
8.	S10thg	0.266	0
9.	Sengl	0.246	0
10.	Sarbl	0.191	0
11.	Hedu	0.188	0
12.	Syer	-0.178	0
13.	SASsat	0.158	0.002
14.	Medu	0.149	0.003
15.	Stime	0.136	0.007
16.	Fsupport	0.129	0.01
17.	Smag	-0.125	0.013
18.	SMsat	0.125	0.012
19.	Sjob	-0.123	0.014
20.	Sabsn	-0.122	0.015
21.	SCRsat	0.121	0.016
22.	Smchn	-0.114	0.023
23.	Sex	-0.113	0.024
24.	Scex	0.103	0.04

To test multicollinearity, a regression model was performed using an Enter method on SPSS, the overall regression was statistically significant (R-square = .828, F(49, 347) = 15.437, p < .000). It was found that there is no multicollinearity problem since all values for VIF in the coefficients table are less than five as shown in table 4, table 5, and table 6.

Table 4: Model Summary

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.828[a]	.686	.641	3.608406411624314

a. Predictors: (Constant), Sabsn, Smag, Medu, Timsm, Ssponser, Eclass, Sacom, Sarbl, Timss, SCRsat, Schois, Fsupport, GPAonl, Smchn, Fsize, Hstat, CalcG2, Scex, Sspt, Ftime, Ttime, Fjob, Simchn, Hedu, Address, Sjob, STLang, S10thg, Timtv, Stime, Sfs, Timfr, PhysG1, Syer, Ssch, Timgout, Sex, SMsat, Timps, Fedu, Mjob, SLsat, Sengl, Snpass, STgrad, PhysG2, SASsat, CalcG1, S12thg

Table 5: ANOVA

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9848.805	49	200.996	15.437	.000[b]
	Residual	4518.147	347	13.021		
	Total	14366.952	396			

a. Dependent Variable: Sgpa

b. Predictors: (Constant), Sabsn, Smag, Medu, Timsm, Ssponser, Eclass, Sacom, Sarbl, Timss, SCRsat, Schois, Fsupport, GPAonl, Smchn, Fsize, Hstat, CalcG2, Scex, Sspt, Ftime, Ttime, Fjob, Simchn, Hedu, Address, Sjob, STLang, S10thg, Timtv, Stime, Sfs, Timfr, PhysG1, Syer, Ssch, Timgout, Sex, SMsat, Timps, Fedu, Mjob, SLsat, Sengl, Snpass, STgrad, PhysG2, SASsat, CalcG1, S12thg

Table 6: Coefficients

Coefficients[a]			
		Collinearity Statistics	
Model		Tolerance	VIF
1	Sex	.637	1.569
	Address	.742	1.347
	Sacom	.842	1.188
	Fsize	.767	1.304
	Smag	.729	1.372
	Syer	.656	1.525
	Ssch	.659	1.519
	STLang	.737	1.356
	S10thg	.633	1.580
	S12thg	.369	2.708
	PhysG1	.451	2.216
	PhysG2	.545	1.834
	CalcG1	.454	2.203
	CalcG2	.524	1.909
	STgrad	.370	2.705
	Sarbl	.739	1.353
	Sengl	.620	1.612
	Snpass	.620	1.613
	Schois	.806	1.241
	Sspt	.820	1.219
	Sjob	.777	1.287
	Mjob	.636	1.572
	Fjob	.837	1.195
	Fsupport	.787	1.271
	Eclass	.857	1.167
	Hedu	.786	1.272
	GPAonl	.846	1.182
	Ssponser	.783	1.276
	Smchn	.808	1.238
	Simchn	.801	1.249
	Hstat	.819	1.221
	Ftime	.792	1.262
	Medu	.470	2.127

	Fedu	.639	1.564
	Ttime	.833	1.201
	Stime	.661	1.512
	Scex	.693	1.444
	Sfs	.712	1.404
	SMsat	.587	1.704
	SLsat	.576	1.735
	SASsat	.511	1.958
	SCRsat	.579	1.728
	Timgout	.685	1.460
	Timsm	.698	1.432
	Timtv	.726	1.378
	Timss	.715	1.398
	Timfr	.689	1.452
	Timps	.639	1.566
	Sabsn	.715	1.399
a. Dependent Variable: Sgpa			

4.2 Machine Learning Models

There are several outputs for each model, including graphs or other comparison criteria used to differentiate between models. So, after building our prediction model the researcher calculate RMSE and accuracy for each model to do the comparison. Below are the results of each machine-learning model.

4.2.1 Decision Tree (rpart)

After building the Decision tree model using the (rpart) package in R, the researcher got an accuracy equal to 96.11, and RMSE equal to 2.086.

The most ten related attributes that affect students' performance by using DT Algorithm are sorted by importance in descending order, as shown in table 7:

Table 7: Variables Importance by DT

ID	Variable	Description
1.	PhysG1	Physics 1 grade
2.	CalcG2	Calculus 2 grade
3.	STgrad	Student's grade in Twjeehi
4.	CalcG1	Calculus 1 grade
5.	S12thg	A 12^{th}-grade student's average in school
6.	Snpass	Number of not passed courses
7.	PhysG2	Physics 2 grade
8.	S10thg	A 10^{th}-grade student's average in school
9.	Sengl	Level of student's English entrance exam
10.	SMsat	Student satisfaction with the major of study

The decision tree output is shown in Figure 6

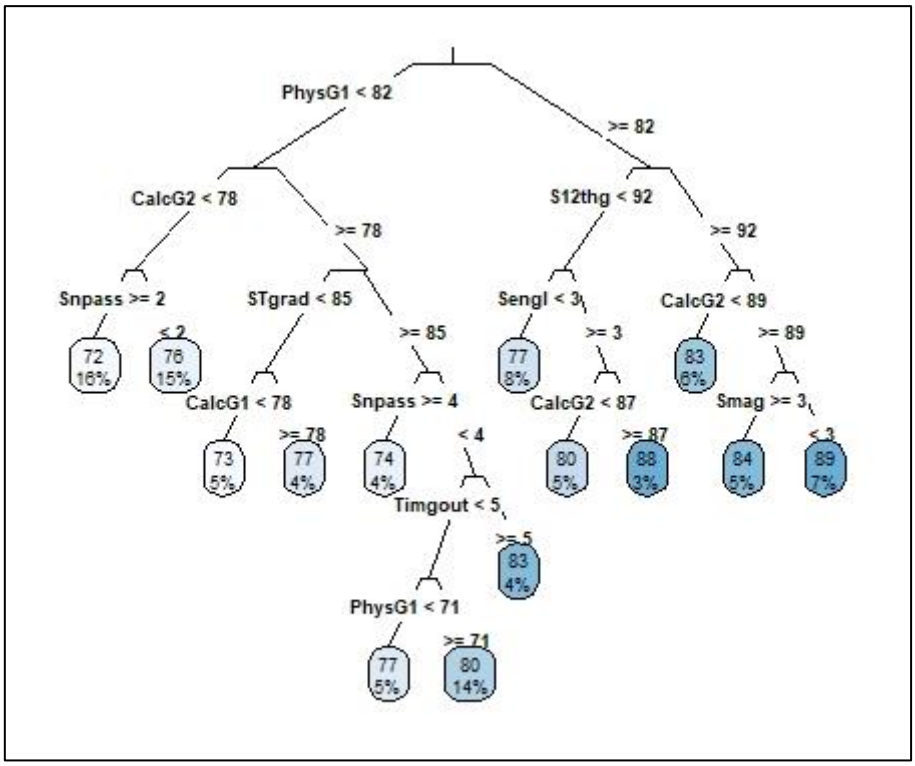


Figure 7: Decision Tree

4.2.2 K-Nearest Neighbour (KNN)

The second used machine learning algorithm was K-NN. The researcher used the KNN method by cart package in R, and the researcher got an accuracy equal to 96.25, and RMSE equal to 1.91.

The most ten related attributes that affect students' performance by using K-NN Algorithm are sorted by importance in descending order, as shown in Figure 7, and table 8:

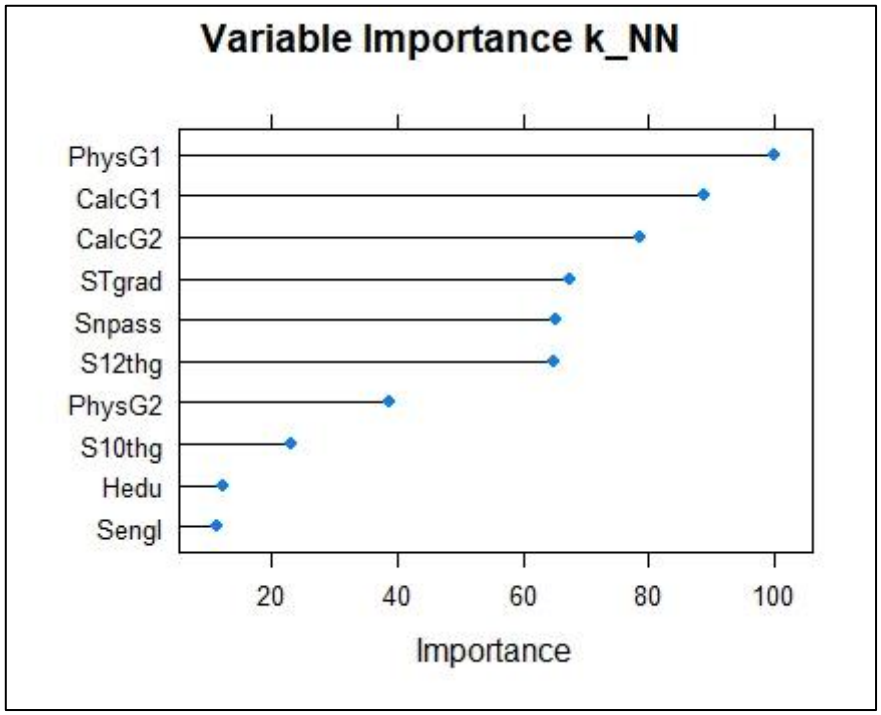


Figure 8: Variable Importance of K-NN

Table 8: Variables Importance by K-NN

ID	Variable	Description
1.	PhysG1	Physics 1 grade
2.	CalcG1	Calculus 1 grade
3.	CalcG2	Calculus 2 grade
4.	STgrad	Student's grade in Twjeehi
5.	Snpass	Number of not passed courses
6.	S12tha	A 12^{th}-grade student's average in school
7.	PhysG2	Physics 2 grade
8.	S10thg	The 10^{th}-grade student's average in school
9.	Hedu	Student wants to complete higher education
10.	Sengl	Level of student's English entrance exam

4.2.3 Support Vector Machine (SVM)

The third machine learning algorithm was SVM, using the (svmRadial) method by caret package in R. This algorithm got an accuracy equal to 96.597, and RMSE equal to 1.77.

The most related attributes that affect students' performance by using the SVM Algorithm are sorted by importance in descending order as shown in Figure 8, and table 9:

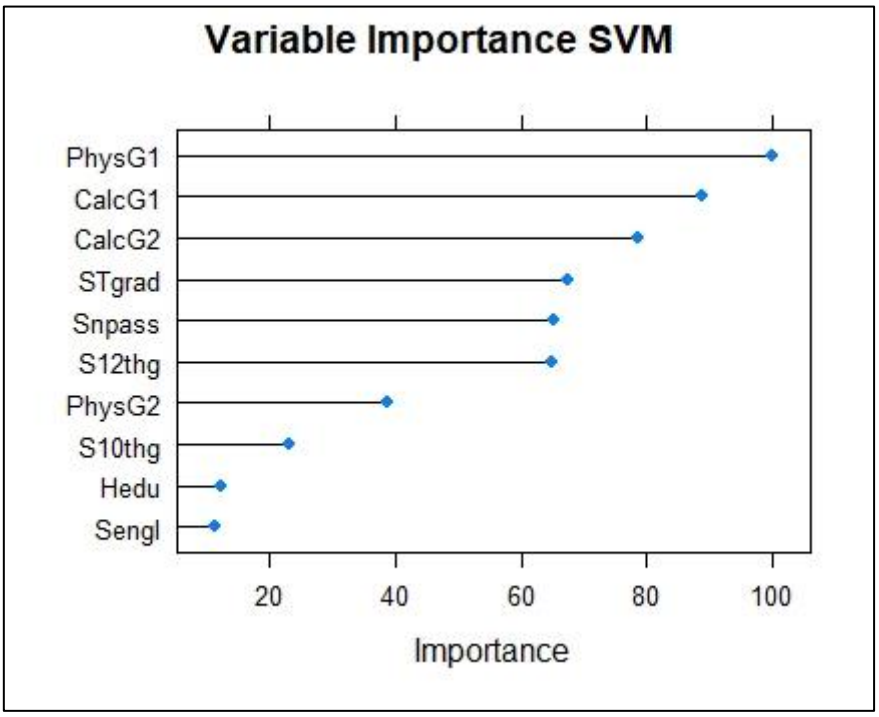


Figure 9: Variable Importance of SVM

Table 9: Variables Importance by SVM

ID	Variable	Description
1.	PhysG1	Physics 1 grade
2.	CalcG1	Calculus 1 grade
3.	CalcG2	Calculus 2 grade
4.	STgrad	Student's grade in Twjeehi
5.	Snpass	Number of not passed courses
6.	S12tha	A 12th-grade student's average in school
7.	PhysG2	Physics 2 grade

8.	S10thg	The 10th-grade student's average in school
9.	Hedu	Student wants to complete higher education
10.	Sengl	Level of student's English entrance exam

4.2.4 Random Forest (RF)

The fourth machine learning algorithm was RF, the researcher used the (rf) method in R, and found that by using this algorithm the researcher got an accuracy equal to 96.60, and RMSE equal to 1.72.

The most related ten attributes that affect students' performance by using RF Algorithm are sorted by importance in descending order as shown in Figure 9, and table 10:

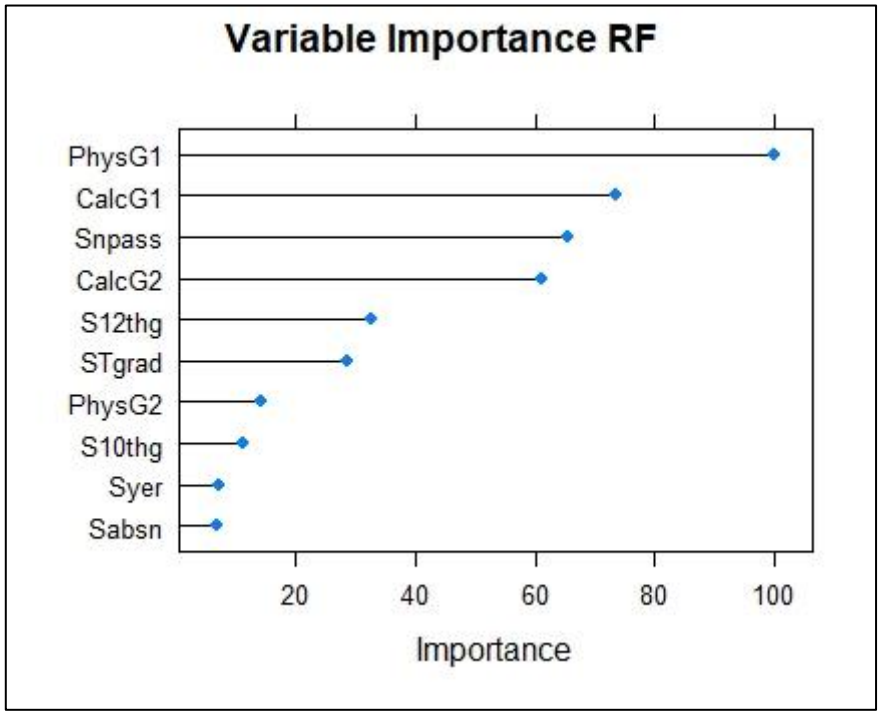


Figure 10: Variable Importance of RF

Table 10: Variables Importance by RF

ID	Variable	Description
1.	PhysG1	Physics 1 grade
2.	CalcG1	Calculus 1 grade
3.	Snpass	Number of not passed courses
4.	CalcG2	Calculus 2 grade
5.	S12tha	A 12^{th}-grade student's average in school
6.	STgrad	Student's grade in Twjeehi
7.	PhysG2	Physics 2 grade
8.	S10thg	The 10^{th}-grade student's average in school
9.	Syer	Student's year at university
10.	Sabsn	Student's absence in class

4.2.5 Artificial Neural Network (ANN)

The last machine learning algorithm was ANN, the researcher used a (neuralnet) package in R. Furthermore, the ANN model needed some pre-processing for the data. Therefore, a function on R was built to do normalization for the data. Based on this algorithm, the researcher got an accuracy equal to 84.23, and RMSE equal to 5.89. The most related attributes that affect students' performance by using ANN Algorithm are as follows :

Table 11: Variables Importance by ANN

ID	Variable	Description
1.	Snpass	Number of not passed courses
2.	S12thg	A 12^{th}-grade student's average in school
3.	CalcG2	Calculus 2 grade
4.	CalcG1	Calculus 1 grade
5.	STgrad	Student's grade in Twjeehi
6.	PhysG1	Physics 1 grade
7.	Sfs	Student's financial status
8.	Medu	Mother's education
9.	Fsize	Family size
10.	Timss	Time spent with social service

The artificial neural network output is shown in Figure 11.

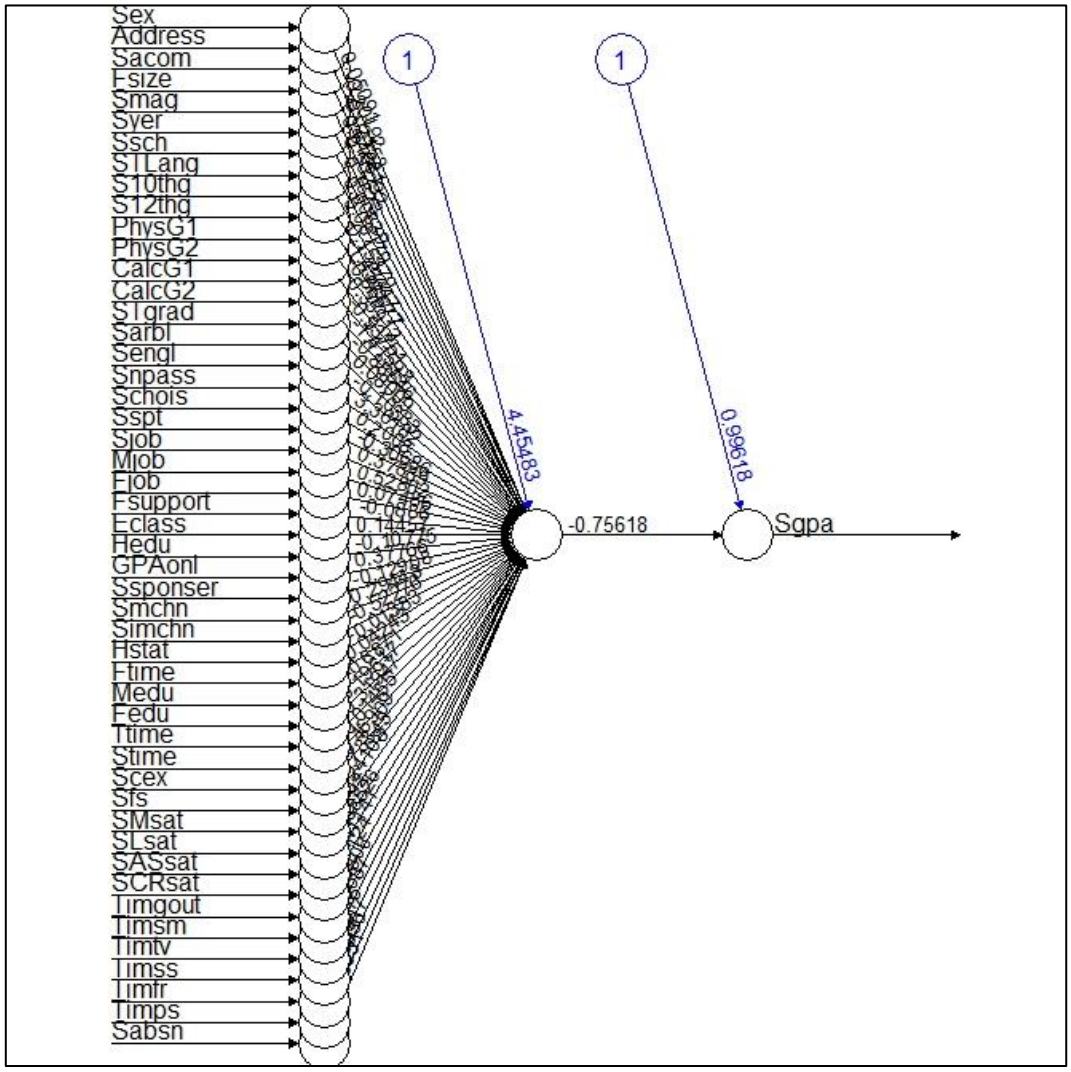


Figure 11: ANN Model

4.3 Models Comparison (DT, K-NN, SVM, RF, and ANN)

Table 12, compares RMSE and Accuracy for (DT, K-NN, SVM, RF, and ANN)

Table 12: Models Comparison

	DT	**K_NN**	**SVM**	**RF**	**ANN**
RMSE	2.0859	1.9086	1.7676	1.7250	5.8866
Accuracy	96.1100	96.2483	96.5971	96.6025	84.2296

From the above table, the researcher found that the RF algorithm has the lowest RMSE and the highest accuracy. So, it is the best predictive model among the used machine learning algorithms in this study. Also, the second one is the support vector machine.

Comparing the most ten related attributes that affect students' performance for different models, as shown in table 13:

Table 13: Common Variables Between Models

ID	DT	K_NN	SVM	RF	ANN
1.	PhysG1	PhysG1	PhysG1	PhysG1	Snpass
2.	CalcG2	CalcG1	CalcG1	CalcG1	S12thg
3.	STgrad	CalcG2	CalcG2	Snpass	CalcG2
4.	CalcG1	STgrad	STgrad	CalcG2	CalcG1
5.	S12thg	Snpass	Snpass	S12tha	STgrad
6.	Snpass	S12tha	S12tha	STgrad	PhysG1
7.	PhysG2	PhysG2	PhysG2	PhysG2	Sfs
8.	S10thg	S10thg	S10thg	S10thg	Medu
9.	Sengl	Hedu	Hedu	Syer	Fsize
10.	SMsat	Sengl	Sengl	Sabsn	Timss

From the above table, the researcher found that the most common five variables among models are (PhysG1, CalcG1, CalcG2, STgrad, and S12tha).

4.4 The most related attributes that affect students' retention

To answer the research question (What are the factors that force engineering students at BZU University to change their major of study?), the researcher performed two RF classifier models one for the variable Smchn as a dependent variable (Did the student change major of study) and the second for the variable Simchn as a dependent variable (Did the student intend to change major of study).

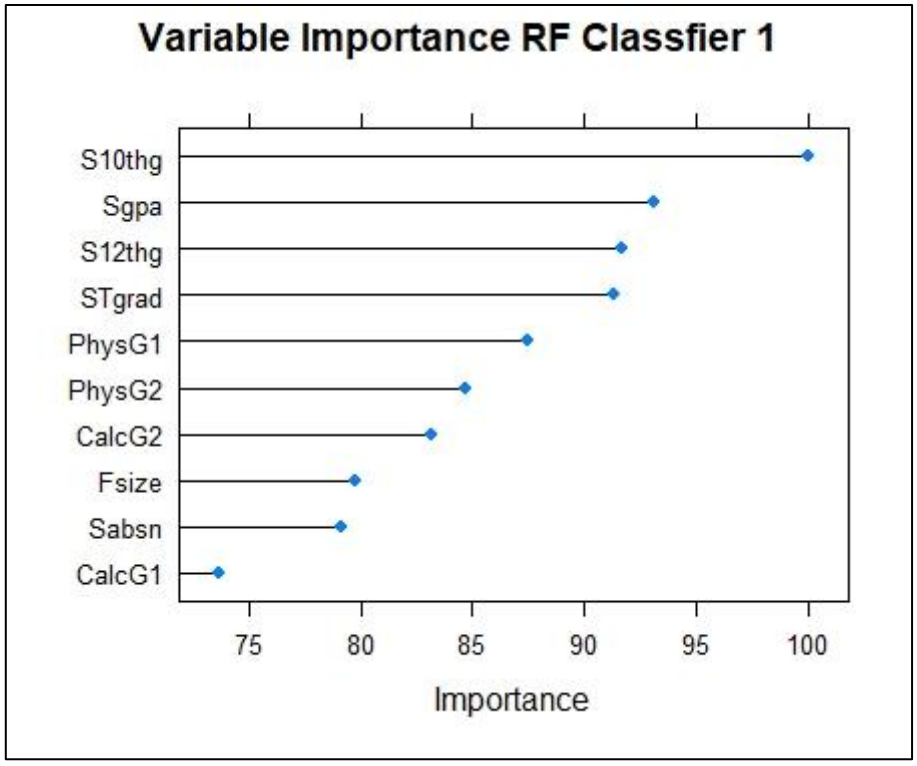


Figure 12: RF Classifier for Smchn

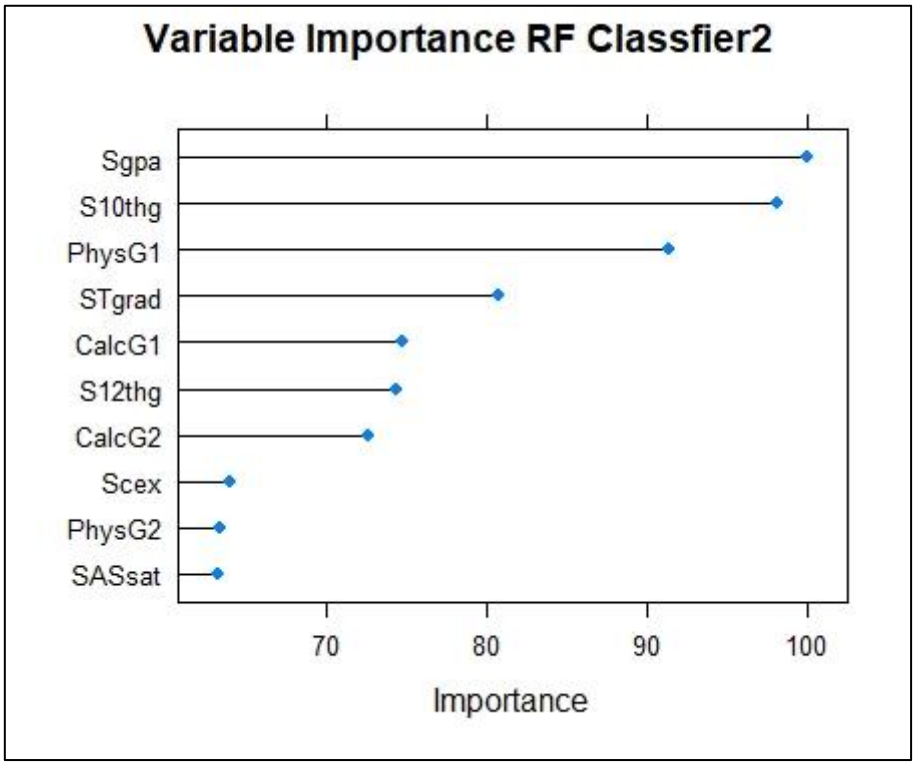


Figure 13: RF Classifier for Simchn

As shown in Figure 12 and Figure 13, the output of the two RF models are identical concerning the variables found among the best 10 in importance but the order of the variables was different in the two models, these variables are listed in table 14, and table 15:

Table 14: RF classifier for Smchn

ID	Variable	Description
1.	S10thg	The 10th-grade student's average in school
2.	Sgpa	Student's GPA
3.	S12tha	A 12th-grade student's average in school
4.	STgrad	Student's grade in Twjeehi (High school diploma)
5.	PhysG1	Physics 1 grade
6.	PhysG2	Physics 2 grade
7.	CalcG2	Calculus 2 grade
8.	Fsize	Family size
9.	Sabsn	Student's absence in class
10.	CalcG1	Calculus 1 grade

Table 15: RF classifier for Simchn

ID	Variable	Description
1.	Sgpa	Student's GPA
2.	S10thg	The 10th-grade student's average in school
3.	PhysG1	Physics 1 grade
4.	STgrad	Student's grade in Twjeehi
5.	CalcG1	Calculus 1 grade
6.	S12tha	A 12th-grade student's average in school
7.	CalcG2	Calculus 2 grade
8.	Scex	Student's computer experience
9.	PhysG2	Physics 2 grade
10.	SASsat	Student's satisfaction with academic staff

Table 16 below shows a comparison between the two RF classifiers based on Accuracy. The researcher notes that model 2 has better accuracy.

Table 16: RF classifier models Comparison

	Accuracy
Model1_Smchn	0.8318584
Model2_Simchn	0.9646018

5 Chapter Five

Discussion, Conclusions, Recommendations, and Future Work

5.1 Discussion and Conclusions

In this research, there were 50 attributes to build prediction models. The results of these models were evaluated to select the best predictive one. Following that, the best variables and factors were identified for creating the models. Also, the factors that influenced students to change their majors in the study were identified. The purpose of this chapter is to provide a discussion and conclusions based on the models' evaluation and their results.

In this study, RF was found to be the best model for predicting students' performance. The results of the study intersect with the studies of Ünal (2020), Madnaik (2020), and Tarik et al. (2021). Additionally, as mentioned (Resende & Drummond, 2018) RF model is a little bit slow, and this research found that the time required to extract results using RF was about five minutes. On the other hand, other models did not take that much time. It was also mentioned that ANN has poor performance for noisy and overlapped data, and in this research, it was noticed that ANN had the lowest model performance. Furthermore, some variables were important for building the models in this study, and in previous studies which include: students' grades in 10th, and 12th by Madhumitha S (2018), Students' GPA as mentioned by Tran et al. (2017), participation and not skipping class by Madnaik (2020), and CGPA is mentioned by Shahiri and Husain (2015). Moreover, other variables had significant importance in previous studies but did not appear to have significant importance in this study including parents' education and income by Madhumitha S (2018), parent's jobs by Madnaik (2020), and student behavior in class by Amrieh et al. (2016).

The following are the main conclusions of the study:

1. One of the results of the predictive models is that the entrance exam variable for the Arabic language level is inessential to predict the students' academic performance. However, regarding the results of the English language admission test, the importance of this test appeared in three models: (DT, SVM, and KNN) found that this variable was among the top ten important variables to predict students' performance.
2. All the used machine learning models (DT, K-NN, SVM, RF, and ANN) in this study, gave a high predictive accuracy. So, it can be used effectively to predict academic performance.
3. Machine learning models can be used effectively to identify the factors that cause students to change their majors.
4. The most significant predictors of students' performance found in this study using the RF model were: (Physics 1 grade, Calculus 1 grade, Number of not passed courses, Calculus 2 grade, A 12th-grade student's average in school, Student's grade in Twjeehi, Physics 2 grade, The 10th-grade student's average in school, Student's year at university, Student's absence in class).
5. The most significant factors that predict the decision of some engineering students to change their majors are (Student's GPA, A 10th-grade student's average in school, Physics 1 grade, Student's grade in Twjeehi, Calculus 1 grade, The 12th-grade student's average in school, Calculus 2 grade, Student's computer experience, Physics 2 grade, Student's satisfaction with academic staff).

5.2 Recommendations

Below are some recommendations for future studies and these are related to the development of academic performance at the Engineering Faculty at Birzeit University.

1. The Admissions and Registration Department can apply machine learning models to study and identify variables that affect the academic performance of students and reconstruct the Engineering Faculty-students' admission criteria and student retention. This study showed that the entrance exam variable for Arabic does not have essential importance regarding the students' academic performance. Therefore, it is recommended to re-study the feasibility of Engineering Faculty students taking this exam and the consequent courses that they must study. Regarding the English language level test, its importance appeared in three predictive models, so, it is recommended to replace the general English courses with more specialized subjects like scientific writing or research. Furthermore, this study also clarified the impact of the achievement variable of students in the tenth and twelfth grades at school on their performance at the university. Therefore, the researcher recommends adding these variables to the admission criteria for the Engineering Faculty.

2. The researcher recommends focusing on the important variables that were found in the study. This will develop the students' academic performance at the Engineering Faculty. Additionally, they will help focus on variables that influence students' decisions in changing their majors for student retention.

3. It is recommended to create a database that contains all student variables that can assist in building machine learning models. This can be achieved through cooperation between the various university departments, including the Department of Admissions and Registration and the Department of Student Affairs.

4. The predictive machine learning models showed that the variable of skipping classes has a significant effect on the students' academic performance, so one of the important recommendations for students is to avoid skipping lectures as much as possible, this issue must also be followed up administratively.

5.3 Limitations

When calculating the required sample size, it was planned to collect data from 403 students with an anticipated non-response rate of 14% to reach a completed data set of around 347 cases. We were able to get 397 completed cases. This number is more than the planned sample size. However, one could have collected data on more students or even from other colleges, but due to time constraints, no more data was collected. As indicated in the recommendations, a database must be created that contains variables that can be used to build prediction models, as it is known that the accuracy of the models depends on the size of the data that the model is trained on. The most important limitations of the study are our inability to collect huge data to train the models.

5.4 Future work

Based on the good accuracy of the results of the prediction models, there are future tasks that can be made in the same field:

1. It is possible to make predictive machine learning models for faculties other than the Engineering faculty and several universities other than Birzeit University.
2. It is recommended to develop an application that can be used by students, academic advisors, and department heads to predict students' performance based on students' information.

6 APPENDIXES

APPENDIX (A): Students' Questionnaire

تعتبر هذه الاستبانة والتي هي جزء من رسالة الماجستير **التنبؤ بأداء الطلاب الأكاديمي باستخدام تقنيات تعلم الالة، كلية الهندسة في جامعة بيرزيت كحالة دراسية،** واحدة من الدراسات التي ستناقش المتغيرات التي تؤثر على أداء الطلاب الأكاديمي واستخدامها لبناء نماذج تنبئ.

لذا نرجو منكم القيام بتعبئة هذه الاستبيان بكل صدق وموضوعية. مع العلم بأن هذا الاستبيان يهدف لجمع المعلومات لغرض البحث العلمي فقط، وسيتم التعامل مع البيانات بسرية تامة، ونشكر لكم تعاونكم.

معلومات عن المبحوث:			
V01	الجنس؟	☐	1. ذكر 2. أنثى
V02	مكان الإقامة؟	☐	1. حضر 2. ريف 3. مخيم لاجئين
V03	نوع السكن؟	☐	1. سكنات الطلبة 2. مع العائلة
V04	عدد أفراد الاسرة؟	☐☐	
V05	ما هو تخصص الدراسة؟	☐☐	1. علم الحاسوب 2. هندسة أنظمة الحاسوب 3. هندسة الميكاترونكس 4. الهندسة الميكانيكية 5. الهندسة الكهربائية 6. الهندسة المدنية 7. الهندسة المعمارية 8. هندسة التخطيط والتصميم الحضري
V06	سنة الدراسة؟	☐	1. أولى 2. ثانية 3. ثالثة 4. رابعة 5. خامسة فأكثر
V07	نوع المدرسة في المرحلة الثانوية؟	☐	1. مدرسة خاصة 2. مدرسة حكومية 3. وكالة UNRWA
V08	لغة الدراسة في المدرسة؟	☐	1. عربي 2. انجليزي 3. غير ذلك
V09	ما هو المعدل بعد انهاء الصف العاشر؟	☐☐	
V10	ما هو المعدل بعد انهاء الصف الثاني عشر؟	☐☐	
V11	ما هو معدل مادة الفيزياء1 في الجامعة؟	☐☐	
V12	ما هو معدل مادة الفيزياء 2 في الجامعة؟	☐☐	

V13	ما هو معدل مادة الرياضيات 1 في الجامعة؟	☐☐	
V14	ما هو معدل مادة الرياضيات 2 في الجامعة؟	☐☐	
V15	ما هو معدل الثانوية العامة التوجيهي؟	☐☐	
V16	ما هو المعدل التراكمي بالجامعة؟	☐☐	
V17	ما هو نتيجة امتحان مستوى اللغة العربية في الجامعة	☐☐	1. ARAB 135 2. ARAB 136
V18	ما هو نتيجة امتحان مستوى اللغة الانجليزية في الجامعة؟	☐☐	1. A1 2. A2 3. B1 4. B2 5. C
V19	ما هو عدد المساقات التي لم تتمكن من النجاح بها في الجامعة؟	☐☐	
V20	ما هي أسباب اختيار التخصص؟		1. أسباب شخصية. 2. اختيار عائلي. 3. سمعة البرنامج في السوق المحلية. 4. سهولة الحصول على العلامة. 5. معدل التوجيهي. 6. أخرى (حدد.....).
V21	هل تمارس التمارين الرياضية؟	☐	1. نعم 2. لا
V22	هل لديك عمل؟	☐	1. نعم 2. لا
V23	هل الام تعمل؟	☐	1. نعم 2. لا
V24	هل الأب يعمل؟	☐	1. نعم 2. لا
V25	هل الوالدين يعيشون مع بعض؟	☐	1. نعم 2. لا
V26	هل الوالدين يدعمون دراستك؟	☐	1. نعم 2. لا
V27	هل أنت ملتحق بدروس اضافية مدفوعة؟	☐	1. نعم 2. لا
V28	هل ترغب بإكمال دراستك العليا في المستقبل؟	☐	1. نعم 2. لا
V29	هل زاد التعليم الإلكتروني معدلك التراكمي؟	☐	1. نعم 2. لا
V30	هل تتلقى أي مساعدات مالية لدراستك؟	☐	1. نعم 2. لا
V31	هل غيرت تخصص الدراسة؟	☐	1. نعم 2. لا
V32	هل تعاني من مشاكل صحية؟	☐	1. نعم 2. لا

V33	هل لديك وقت فراغ بعد الجامعة؟	☐	1. نعم 2. لا
V34	المؤهل العلمي للوالدة؟	☐	1. ثانوية أو أقل. 2. كلية 3. بكالوريوس 4. ماجستير أو أعلى.
V35	المؤهل العلمي للوالد؟	☐	1. ثانوية أو أقل. 2. كلية 3. بكالوريوس 4. ماجستير أو أعلى.
V36	كم من الوقت تقضي للوصول من البيت إلى الجامعة؟	☐	1. أقل من 15 دقيقة. 2. من 15 إلى 30 دقيقة. 3. من 30 إلى ساعة. 4. أكثر من ساعة.
V37	كم من الوقت تقضي في الدراسة؟		1. فقط أدرس وقت الامتحان. 2. ساعة واحدة 3. بين 2 و0 ساعات. 4. أكثر من 4 ساعات.

يرجى ملء ما يلي بوضع (✓) في الفراغ التالي الذي يناسب إجابتك.

استخدم مقياس التصنيف:

5- جيد جداً 4- جيد 3- متوسط 2- ضعيف 1- ضعيف جداً

	السؤال	1	2	3	4	5
V39	المهارة في استخدام الحاسوب					
V40	الوضع المادي للأسرة					

يرجى ملء ما يلي بوضع (✓) في الفراغ التالي الذي يناسب إجابتك.

استخدم مقياس التصنيف:

5- راضي جداً 4- راضي 3- محايد 2- غير راضي 1- غير راضي جداً

	السؤال	1	2	3	4	5
SAT1	مدى الرضى عن التخصص					
SAT2	مدى الرضى عن الخدمات اللوجستية في الكلية					
SAT3	مدى الرضى عن الطاقم الأكاديمي					
SAT4	مدى الرضى عن المناهج والمقررات الدراسية					
SAT5	مدى الرضى عن التدريس الإلكتروني					

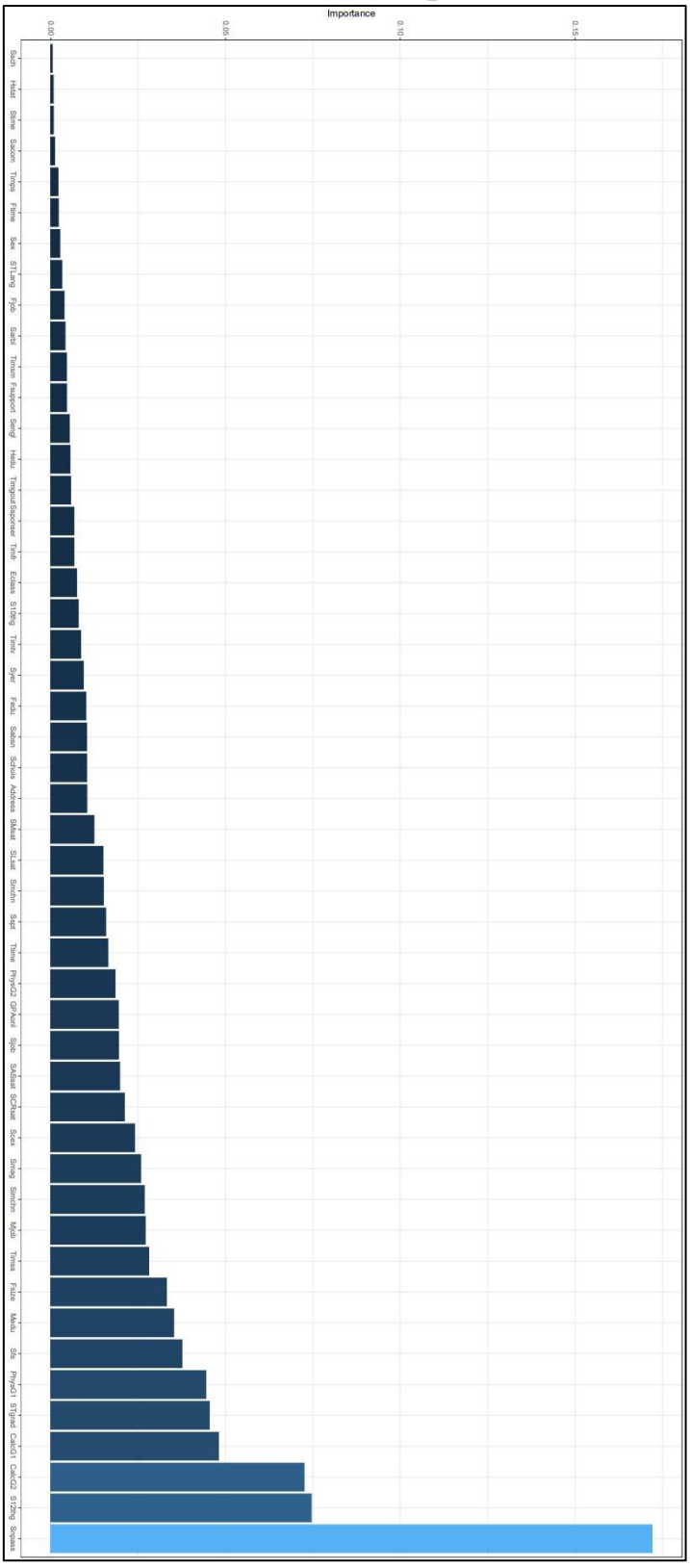
يرجى ملء ما يلي بوضع (✓) في الفراغ التالي الذي يناسب إجابتك.

استخدم مقياس التصنيف:

5- دائماً 4- غالباً 3- أحياناً 2- نادراً 1- مطلقاً

	السؤال	1	2	3	4	5
Time1	هل تقضي بعض الوقت للخروج مع الأصدقاء؟					
Time2	هل تقضي بعض الوقت على وسائل التواصل الاجتماعي؟					
Time3	هل تقضي بعض الوقت على التلفاز؟					
Time4	هل تقضي بعض الوقت في الخدمة المجتمعية التطوعية؟					
Time5	هل تقضي بعض الوقت في القراءة الحرة؟					
Time6	هل تقضي بعض الوقت لممارسة النشاط السياسي؟					
Time7	أقوم بالغياب عن المحاضرات					

APPENDIX (B): ANN Variable Importance



7 References

Acharya, A., & Sinha, D. (2014). Early prediction of students performance using machine learning techniques. *International Journal of Computer Applications*, *107*(1).

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

Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, *27*(1), 194-205.

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*, *9*(8), 119-136.

Anuradha, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian Journal of Science and Technology*, *8*(15), 1-12.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Buis, M. L. (2017). Logistic regression: When can we do what we think we can do. *Unpublished note*, *2*.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

Cunningham, P., & Delany, S. J. (2020). k-Nearest neighbour classifiers: (with Python examples). *arXiv preprint arXiv:2004.04523*.

Genders, T. S., Spronk, S., Stijnen, T., Steyerberg, E. W., Lesaffre, E., & Hunink, M. M. (2012). Methods for calculating sensitivity and specificity of clustered data: a tutorial. *Radiology*, *265*(3), 910-916.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hilbe, J. M. (2011). Logistic Regression. *International encyclopedia of statistical science*, *1*, 15-32.

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development Discussions*, 1-10.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, *15*(1), 41-51.

Khair, U., Fahmi, H., Al Hakim, S., & Rahim, R. (2017). Forecasting error calculation with mean absolute deviation and mean absolute percentage error. Journal of Physics: Conference Series,

Lee, Y.-C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, *33*(1), 67-74.

Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, 1-13.

Madhumitha S, V. R., Vinitha B, Vijayakumar S. (2018). Using Data Mining to Predict Students Performance. *International Research Journal of Engineering and Technology (IRJET)*, *05*(09).

Madnaik, S. S. (2020). *Predicting Students' Performance by Learning Analytics* [Master's Projects. 941., https://scholarworks.sjsu.edu/etd_projects/941

Maldonado, L. E. A. (2019). Los Rostros del Caño-An Exploration in Immersion Journalism from University. 4 th Pan-American Interdisciplinary Conference PIC 2019,

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Monroe, W. (2017). Logistic regression. *Recall*, *1*(1).

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nasser, I. M., & Abu-Naser, S. S. (2019). Predicting Tumor Category Using Artificial Neural Networks.

Otchere, D. A., Ganat, T. O. A., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal of Petroleum Science and Engineering*, *200*, 108182.

Pojon, M. (2017). *Using machine learning to predict student performance*

Resende, P. A. A., & Drummond, A. C. (2018). A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR)*, *51*(3), 1-36.

Shah, H. (2021). *A Full Overview of Artificial Neural Networks (ANN)*. https://learn.g2.com/artificial-neural-network

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. *Journal of Statistics and Management Systems*, *20*(4), 713-722.

Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science*, *184*, 835-840.

Tran, T.-O., Dang, H.-T., Dinh, V.-T., & Phan, X.-H. (2017). Performance prediction for students: A multi-strategy approach. *Cybernetics and Information Technologies*, *17*(2), 164-182.

Ünal, F. (2020). Data mining for student performance prediction in education. *Data Mining-Methods, Applications and Systems*.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67.

Yamane, T. (1967). *Statistics: An introductory analysis*.

Yassein, N. A., Helali, R. G. M., & Mohomad, S. B. (2017). Predicting student academic performance in KSA using data mining techniques. *Journal of Information Technology & Software Engineering*, *7*(05).

Yeşilkanat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons & Fractals*, *140*, 110210.