MACHINE LEARNING BASED STUDENT GRADE PERFORMANCE ANALYSIS

By

ASHI MEHMOOD

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

January, 2022

Machine Learning Based Student Grade Performance Analysis

By

ASHI MEHMOOD

BS (IT), PMAS Arid Agricultural University, Rawalpindi 2016

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In **Software Engineering**

To

FACULTY OF ENGINEERING & COMPUTER SCIENCES

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

NATIONAL UNIUVERSITY OF MODERN LANGUAGES FACULTY OF ENGINEERIING & COMPUTER SCIENCE

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Engineering and Computer Sciences for acceptance.

Thesis Title: Machine Learning Based Student Grade Performance Analysis

Submitted by: Ashi Mehmood **Registration #:** 30MSSE/lbd/F19

Master of Science in Software Engineering
Degree name in full

Software Engineering
Name of Discipline

Dr. Raheel Zafar	
Name of Research Supervisor	Signature of Research Supervisor
Dr. Javvad-ur-Rehman	
Name of Research Co-Supervisor	Signature of Research Co-Supervisor
Dr. Basit Shahzad	
Name of Dean (FE&CS)	Signature of Dean (FE&CS)
Prof. Dr. Muhammad Safeer Awan	
Name of Pro-Rector Academics	Signature of Pro-Rector Academics

Date

AUTHOR’S DECLARATION

I Ashi Mehmood

Daughter of Khalid Mehmood

Registration # 30MSSE/lbd/F19

Discipline Software Engineering

Candidate of **Master of Science in Software Engineering (MSSE)** at the National University of Modern Languages do hereby declare that the thesis **Machine Learning Based Student Grade Performance Analysis** submitted by me in partial fulfillment of MSSE degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution. I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

____
Signature of Candidate

Ashi Mehmood
Name of Candidate

____
Date

ABSTRACT

Machine Learning Based Student Grade Performance Analysis

Machine learning algorithms may be able to address the growing difficulty of integrating student-related data for the prediction of student performance in order to make better administrative decisions. Machine learning reviews data mining techniques and provides various models to predict students' performance. The study aims to identify the factors that can improve the students’ performance using machine learning techniques. Machine learning involves various features, and it needs statistical and classification algorithms for better prediction. This study indicates the key factors and predicts student performance with better accuracy based on identified factors. Currently, different studies are using various machine learning techniques to predict students' performance. This research presents a paradigm for evaluating academic achievement in students. In this research, the dataset is carefully chosen which includes demographics, previous academic records, and information related to family background. The data was collected from students of various universities. Due COVID-19, the data was collected through online questionnaires and twenty-four different attributes were selected which were taken from different previous studies after being pre-processed. The study is designed to determine the key attributes influencing the students' performance. Another important part of predication is based on different classifiers having various classification algorithms. The result of this study show that the Support Vector Machine is better than various other algorithms. As a result of this study, student’s performance can be improved by working on the specific features which can improve the quality of education. Furthermore CNN can be used to improve accuracy by collecting more data which can help for better utilization of school systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

`

LIST OF ABBREVIATION

ANSI	American National Standards Institute
CNN	Convolutional Neural Network
EDM	Educational Data Mining
GPU	Graphics Processing Unit
GPA	Grade Point Average
IHL	Institute of Higher Learning
ID3	Iterative Dichotomiser 3
K-NN	K- Nearest Neighbor
KNIME	Konstanz Information Miner
LMS	Learning Management System
MPL	Multilayer Perception
RB	Rule Base
SAM	Student Attribute Matrix
SPPN	Students Performance Prediction Network
TSQ	Time Structured Questionnaire
WEKA	Waikato Environment for Knowledge Analysis
DT	Decision Tree
ANN	Artificial Neural Network
RF	Random Forest
LG	Logistic Regression
VLE	Virtual Learning Environment
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative

ACKNOWLEDGEMENTS

All the praises and thanks to almighty Allah who bestowed his innumerable blessings upon mankind, one of which is knowledge a distinction for mankind. I offer my gratitude to the Holy Prophet Muhammad (S.A.W) who preached us to seek knowledge for the betterment of mankind in particular and other creatures in general. This dissertation is the end of my journey in obtaining my MS Degree. This dissertation has been kept on track and seen through to completion with the support and encouragement of numerous people including my all teachers, friends, colleagues and various institutions. At this moment of accomplishment, first of all, I pay homage to my supervisor Dr Raheel Zafar and co-supervisor Dr Javvad-ur-Rehman. This work would have not been possible without his guidance, support and encouragement. Under their guidance, I successfully overcame many difficulties and learned a lot. I am very much thankful for his valuable advice, constructive criticism and extensive discussion around my work.

I am indebted to all my colleagues and friends especially Muhammad Azhar, Malik Muhammad Shahid Ali for providing a stimulating and fun-filled environment and many rounds of discussion with them helped me. Last but not least, I would like to pay high regard to my Father, Mother, and Siblings, family for their prayers, sincere encouragement and inspiration throughout my research work and for lifting me uphill in this phase of life. I owe everything to them.

DEDICATION

This thesis work is dedicated to my parents and my teachers throughout my education career who have not only loved me unconditionally but whose good examples have taught me to work hard for the things that I aspire to achieve.

CHAPTER 1

INTRODUCTION

1.1 Overview

Predicting the outcome of any task completed by an individual is an extremely interesting issue that has grabbed the interest of the people in recent years. Predictive analysis is a strategy for analyzing previous occurrences in order to anticipate future outcomes. Predictive Analysis uses a variety of approaches to analyse [1] current data and create predictions about the future, including data mining, machine learning, statistical modelling, and artificial intelligence. We can forecast the outcome of the student's grade pretty accurately using these strategies. Analysis of educational data, such as learning analytics, academic analytics, educational data mining, predictive analytics, and learner analytics, has recently recently emerged as an important area of study.

Personal, socioeconomic, psychological, and other environmental variables all influence the academic performance of the student. Prediction models that include all of these variables are required for accurate prediction of student performance. Predicting student performance with high accuracy is useful for identifying students with low academic achievements early on. Educators can help identified students individually so that their performance improves in the future.

Educators, parents, and institutions want to know the answer to the question, "Is it possible to predict a student's performance in an educational institution?"For instance, will he finish his degree or not? However, the process of learning is now defined as an individual's effort. As a result, developing models for evaluating a student's learning efforts is a difficult task [1]. Machine learning techniques have recently been used to provide new insights into this problem. To evaluate the performance of the students, a variety of influential factors are used. These can be identified by employing machine learning techniques in the educational sector.

Every person's ability to attain success in life is dependent on their level of education. Education plays a significant role in achieving a brighter future. The government and educational institutions interact with a variety of activities to improve the value of education in the country [2]. Good education serves various functions in life, including personal improvement, social status, social health, economic progress, economic achievement, setting life goals, making us aware of many global problems, and providing answers to environmental problems and other connected issues. There are numerous strategies to increase educational levels today.

Data mining is the process of evaluating the extracted meaningful information from huge amounts of data. The primary purpose of data mining is to extract information from huge amounts of data that can be kept in large datasets and to convert raw data into meaningful information. Data mining facilitates the discovery of patterns in data. Data mining techniques are beneficial in a variety of industries, including games, business, medical diagnostics, research and engineering, and so on. Machine learning's key functions include pattern discovery, association, correlations, prediction, classification, clustering, trend analysis, and many others. For more accurate prediction, the data mining task also incorporates a decision support system. For decision-making and data mining, data mining technology employs a variety of data mining tools and processes.

1.2 Machine Learning

Machine learning is a branch of computer science that allows a computer to learn without the assistance of external applications. These machine learning approaches can be used to forecast the outcome of specific inputs. Machine Learning is a set of techniques that enable computers to learn without the need for human interaction. Medical diagnostics, stock market analysis, DNA sequence classification, games, robots, predictive analysis, and other applications have all benefited from machine learning. We're particularly interested in predictive analysis, where machine learning allows us to create complicated models that may be used to make predictions. People benefit significantly from these models since they provide relevant data that helps them make better decisions. Designing and developing algorithms for computers to predict behaviour based on a dataset collected that's what Machine Learning is all about. It's an artificial intelligence sub-discipline. In recent years, there has been a tremendous increase in the use of these algorithms in the field of education. Pattern recognition and decision making are the major goals of these machine learning systems [3]. First, patterns are recognized, and then rules are formed based on the supplied data. The behaviour is expected and decisions are made based on those guidelines.

Machine learning is a branch of computer science that is distinct from the basic computing methods that are used to solve problems [4]. The algorithms used in machine learning are developed in such a way that the system or computer can evaluate data inputs, create training sets, and produce the required range determined output using statistical estimation.

Machine learning is a branch of science that uses patterns and inferences to make decisions. It builds a system that requires less human intervention using a statistical model and algorithms. Supervised learning and unsupervised learning are the two primary kinds of machine learning algorithms. This [5] study focuses on supervised learning, which employs a classification algorithm to predict students' academic success. I explored several classification methods for studying and forecasting final year students' grades [5]. The algorithm and data are used to determine the outcome of machine learning in the field of education. It's important to set the appropriate statistical methods for forecasting students' success. The machine learning algorithm determines the efficiency of the result. According to current research, a student's academic success is influenced by his or her background and other characteristics. [6] Many studies show that, in addition to past academic performance, a student's background and other characteristics have a significant impact on their performance.

Most machine learning classifiers, such as the Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Naive Bayes, allow multiclass classification by default [7]. It's important to keep in mind that some research studies have introduced and compared various machine learning and data mining techniques. In 2014, several machine learning algorithms were used to predict student performance.

Previously, manual machine learning techniques were used to review data and provide business projections. Currently, the rapid growth in educational data and its utilization to improve the quality of managerial decisions are the biggest challenges faced by educational institutions [8]. New researchers are being conducted to discover innovative interventions to help the education sector better manage its educational data.

Currently, the Institutions of Higher Learning (IHL) database carry a lot of information about students [9]. The number of students and data is increased over time and there is no technique to convert the information in knowledgeable format. Machine learning techniques can be applied to the Institutions of Higher Learning (IHL) database to find new information about the students. Prediction of student performance is the most demanding application of ML.

Machine Learning (ML) application is concerned with the extraction of hidden patterns and recognition of the relationship among the parameters in a large amount of data. M. Dayalan et.al [10] machine Learning is subjected to two main objectives; prediction and data possible outcomes.

1.3 Prediction

Regression analysis is the most widely used prediction approach. It is made up of one or more predictor variables. Continuous and attribute variables can both be used in the regression. Prediction is based on the link between something that is known and something that needs to be predicted. For example, if a student's existing knowledge and communication skill are known, multiple regression can be used to forecast his or her placement possibilities. The dependent variable is often denoted by y, whereas the independent variables are generally denoted by x and domain knowledge and communication level are generally marked by x.

1.4 Prediction Student’s Performance

Predicting student performance is one of the most significant topics for learning contexts such as schools and universities because it contributes to the development of successful mechanisms that, among other things, improve academic results and decrease dropout. In higher education institutions, predicting students' academic progress is important. Understanding the aspects that influence student performance is a challenging academic project due to a variety of factors such as cultural, social, previous academic success, teacher interaction, and so on [11]. Several researchers have been studying these aspects, and their findings have been promising. Some researchers, for example, looked into the impact of socioeconomic level. Others investigated the relationship between student academic achievement and parental actions, while others investigated the effectiveness of teachers in improving student academic success. Predicting student performance entails identifying the significant traits, the relationship between the attributes, and other factors that influence student achievement. Many indicators assist in determining student achievement.

Machine learning techniques were recently used to solve the problem of predicting a student's total grade in a particular course. Based on previous performances of similar students, we forecast final grades. Because study-related records, such as age, gender, and field of study, are readily available in university computer systems, researchers frequently evaluate them [12]. They also try to find other traits that will help them better understand students' conduct, such as their habits or their parents' education. Questionnaires are the most common method of obtaining such information.

The majority of techniques are designed to boost students' academic performance and encourage their overall development. We focus on students' behaviours in this study, to anticipate their performance, as this allows us to identify students' learning challenges ahead of time [13]. At the same time, personal advice can be provided to help students improve completely. Furthermore, because students' behaviours are intuitive, we may more easily determine consequences directly and rapidly, rather than waiting until the end of the semester to discover students' learning and life problems. Many researchers have discussed using various technologies to examine this problem, such as statistical analysis, data mining, and questionnaire surveys, to determine students' performance based on their behaviour data.

Predicting student academic performance is critical for academic growth, but it can be difficult due to the influence of several factors on students' performance. The high accuracy prediction of student performance is useful because it allows students with low academic achievements to be identified early in their academic careers [7]. There are two types of machine learning algorithms. There are: • supervised learning and • unsupervised learning.

1.5 Supervised Learning

The most common application of supervised learning is in classifying issues. The primary purpose of this learning is to develop a classification model that allows the computer to learn about the input and forecast the outcome. Collecting data, finding an appropriate method, creating the model, and applying the model for prediction are the main steps in supervised learning. To develop a model in supervised learning, you need a set of inputs and known responses. New data is mapped to the required answers or outputs using the model that was created. The probability of all the given inputs is also provided by the supervised learning algorithms. In the dataset, there should be no missing values. It is impossible to predict the outcome if any values are missing. Nave Bayes, Logistic Regression, Decision Trees, and Neural Networks are examples of algorithms used in this form of learning.

1.6 Un-supervised Learning

Only inputs are used in unsupervised learning, and there are no outputs. The main objective of this form of learning is to model the structure of unlabeled data to get a better understanding of it. The data is left to the algorithm to analyze the relevant patterns in the data because there is no specific output. The relationships between the data are typically found here. The same collection of the dependent variable is used to obtain all of the outputs. However, in the case of supervised learning, the reason for an output set of data is the input set of data.

1.7 Problem Statement

Machine Learning (ML) is a growing and emerging field to investigate the unique type of data that come from an academic setting source. As it involves various features. It needs statistical and machine learning algorithms for better prediction. Since there are limited investigations on the different factors affecting student achievements, the main objective of this study is to identify the factors and predict student performance with better accuracy using machine learning. [20] [15].

1.8 Research Question

The main purpose of this study is to explore the existing research, critically examine the existing student performance prediction models and identify how different features can affect student performance.

RQ 1: Which attributes can help in improving the academic performance of the student?

RQ 2: Which machine learning algorithms can effectively predict the academic performance of the student with better accuracy?

1.9 Aims and Objectives

The objectives of our research are:

- To identify the attributes for the improvement of student performance.

- To identify a machine learning algorithm for better accuracy and model validation by comparing with existing systems.

1.10 Organizational Chart of Thesis

Figure 1 is representing the organization of the thesis.

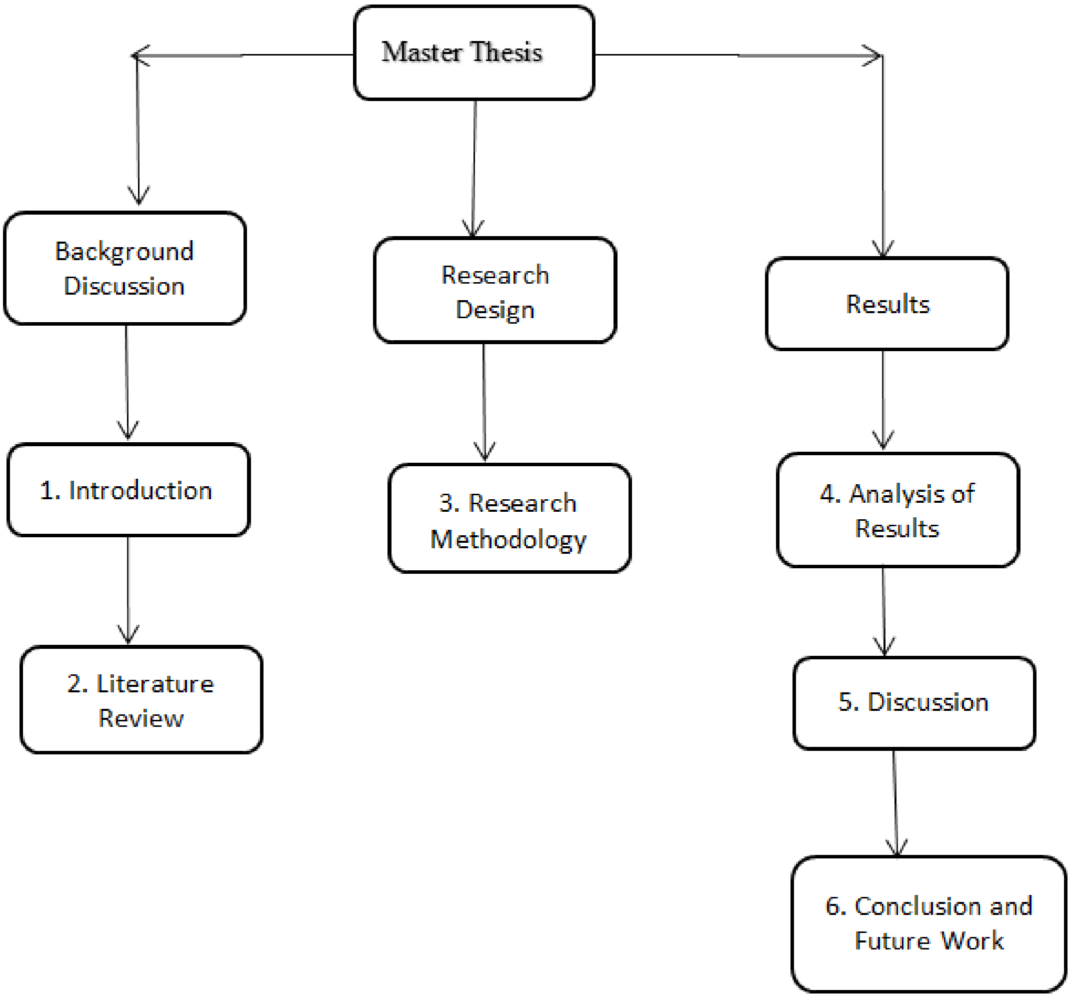


***Figure 1.1:** Organizational Representation of Thesis (This figure shows the flow of conducting the research)*

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Machine Learning is a set of techniques that enable computers to learn without the need for human interaction. Medical diagnostics, stock market analysis, DNA sequence classification, games, robots, predictive analysis, and other applications have all benefited from machine learning. Machine learning can be utilized in the educational field to boost the understanding of the learning process by specializing in discovering, extracting and analyzing factors associated with students' educational processes. Predicting a student's performance means detecting the different attributes, the relationship between the attributes and other aspects that affect a student's performance. The main target of this task is to assess the learner if he/she can accomplish a given task to attain a particular learning goal. Many studies have been conducted on the prediction of student performance.

F. Ahmad, et.al [9] presented the framework that was used to predict the first-year bachelor student's performances that were enrolled in the computer science course. The data was gathered for the duration of 8 years dated July 2003/2004 to July 2013/2014 containing the data about student's information e.g. student's previous academic record, family background and student's demographics. To produce the best student academic performance prediction model different classification techniques like Decision Tree, Naïve Bayes and Rule-Based (RB) were applied to the student's data. The finding of the study reported that the Rule-Based produce the best accuracy as compared to the other techniques. This model works efficiently for the poor and average grade students and allow the teachers to help and assist this category of students to improve their performance.

Ishwank Singh, et.al [14] proposed to offer a basic clustering analysis to look into the student's behaviour. A good benchmark has been produced by the data mining algorithms to comprehend if there is a consistent improvement in the student's performance. This analysis is extremely useful when the student admissions and placement process begins. This analysis includes certain parameters such as projects, internships, skill set, 10th, 12th, and B.Tech marks. The K-means algorithm is utilized in clustering because it is simple to implement and has high computational efficiency. Many other clustering strategies can be used in the future to boost efficiency levels. The ranking or classification of the post within the clusters will also be done to achieve a better student performance analysis.

Ihsan A. Abu Amra, et.al [15] presented that there is a large growth of information available in the academic database. Several classification algorithms are applied to the educational datasets to gather knowledge about student's performance. For prediction of the student performance, the study proposed the prediction model by focusing on the K nearest neighbour (KNN) and Naive Bayes data mining algorithms. K nearest neighbour (KNN) and Naïve Bayes were presented to predict the performance of the student. Evaluations were performed and the comparison was done against KNN and Naive Bayes in terms of certain achievement parameters. When the Naïve Bayes algorithms were applied 93.17% of accuracy was achieved. The study [15]stated that a strong relationship exists between the significant features which affect the student's performance.

Fan Yanga, et.al [16]analyzed the student's achievements, development and abilities using the different analysis techniques. Initially, Student Attribute Matrix (SAM) was used to develop the student model accompanying the performance and non-performance related attributes. When providing the student's performance estimation tool, the BP-NN algorithms were applied. The existing information of the student's performance and their achievements attributes were used to evaluate student's attributes. Through these indicators and predictors can be known the level at which level the factors would affect the student's performance. These estimation tools were used to check the real academic performance of the student's data which is collected from 60 high schools. As per the evaluation results, attain better and perfect results. Thus, a better understanding process was achieved here.

Raheela Asif, et.al [17] studied undergraduate student's performance by making use of data mining techniques. Initially, the academic achievement of the students was estimated after completing the four-year graduation program. Moreover, typical development was observed and all the prediction results were combined with them. The two most important groups were identified in this study which was low and high achieving students. This study provided a timely warning for low achieving students and supported that by focusing on fewer number of courses, the performance can be improved which help in indicating that the performance is good or not.

Febrianti Widyahastuti, et.al [18] presented to show a system that used two separate categorization algorithms: linear regression and multilayer perceptron. In addition, the algorithms were compared based on their mean value and absolute inaccuracy. In comparison to linear regression, the prediction results of a multilayer perceptron were found to be better. Students' educational experiences were enhanced through the utilization of online discussion forums.

Ms Tismy Devasia, et.al [19] presented the division of students can be anticipated using the provided categorization technique within the student's knowledge based on previously available information. Since many strategies for knowledge classification within the area unit were used, the naive theorems were used. Various types of knowledge were collected from the student's previous knowledge to forecast his or her performance for that semester. The lectures and students can profit from this study to motivate students of various categories to do effectively. Students that require further assistance can be identified using this research. Also, with the help of this approach, the failure rate can be decreased. Acceptable measures are frequently taken through this for next semester examinations.

P. Cortez, et.al [20] proposed the system that addressed the prediction of secondary school student grades by utilizing data mining techniques. Three data mining methods such as binary, regression and five-level classification were used along with four techniques i.e. Decision tree, Random Forests, Neural Networks and Support Vector Machine (SVM) were examined clear inputs were chosen e.g. with or without past academic performance was investigated. The outcomes revealed that only if the first and second school grades are known then it’s possible to attain excessive predictive accuracy. This study validated the assumption that student success is highly influenced by earlier performance. In this study [20] after the data was collected the data mining techniques were applied that’s why this study focuses on off-line learning. This will allow the gathering of information about extra attributes such as precious academic records and treasure feedback from the school.

M. M. Abu Tair, et.al [21] explored academic knowledge mining by deploying a case study on graduate students. For this study [21] the data was obtained from the Khanyounis College of Science and Technology over 15 years, started in 1993 and ending in 2005. The data contained 3360 trials and 18 features. To predict the grade of graduate students, it has used two classification methods which are association, Classification, Cluster and Outlier detection. Two Classification methods Rule Induction and Naïve Bayesian classifiers and the Rule Induction shows 71.25% and Naive Bayesian shows 67.50% which seems to be good accuracy. These two classification methods are used for estimating the low grade of students on time. Furthermore, the K-mean clustering algorithm was used to cluster the students into groups. To discover all outliers in the data, used two outlier approaches: the Distance-based Approach and the Density-Based Approach. Each of these tasks can be used to enhance a graduate student's performance. Outlier analysis in educational data mining can be used to determine students who may have difficulties.

M. S. Mythili, et.al [22] analyzed and estimated the school student's performance. WEKA tool was used to classify the different data mining algorithms. Data mining tool has been accepted as the decision-making tool for facilitating greater resource implementation. The factors were attendance, gender, results, economic status and parent's education. The classification techniques were used to gain high accuracy. To build the model, the classification rule was applied to the training data. Test data was used to examine the accuracy of the model in the second step of the classification. This model will be used to arrange the unknown tuples only if the accuracy of the model is suitable. The decision tree technique was put into the data set and the presumption appears that the attendance was significantly related to the student's performance. From the rule set, it was established that high potential variables like parent education, locality, gender; economic status affect the student's performance for getting good performance. To analyze the school student's performance, the outcomes of time and accuracy revealed that the Random Forests is the very best classifier according to this study [22] Random Forests is a well-organized and low error rate classification technique among the others.

P. Veeramuthu, et.al [23] the author proposed a system that can analyze the various attributes that influenced the students learning behaviour and performance using the clustering technique. Clustering is the technique that divides a data set into groups called clusters so that similar data points are grouped in one cluster. A significant collection of data was segmented into subsets through cluster analysis. Each cluster had a collection of data elements that were similar to those found in another area. Clustering was one of the earlier and simple techniques which were used to analyze the data set. In this study [23] according to the characteristics of the student, clustering analysis was used to make the segments of students into groups. Clustering will be appraised as the most important unsupervised learning technique. In the previous studies, various clustering techniques have been applied to Educational Data Mining (EDM). This study [23] valuated these clustering algorithms as applied to the EDM context. This study [23] has been conducted a survey on EDM that gives the wide-ranging resources of papers which are issues between 1995 to 2005. For student annotations, researchers have applied hierarchical clustering and statistical clustering method like K-means clustering. A cluster of students that's shared similar learning patterns used the k-mean clustering algorithms. For the innovations of attainable dependencies among the factors, it has been used classification

, clustering, association rule and regression. This study outcome revealed that the online learning platform was one of the factors which influenced their performance.

A.I, Adekitan, et.al [24] describes that in the Konstanz Information Miner (KNIME) workflow, the program and the entry year were used as the predictive inputs. Konstanz Information Miner (KNIME) workflow used the six independent data mining algorithms that were used separately for the modified performance analysis of the outcome of each algorithm. To improve the nature of the program, the model will allow the IT department to take the correct decisions to visualize and support the students. This study, only emphasizes predicting student's performance. The BS (IT) program has 4 years with a 2nd-semester study plan. The 1st year is the starting year, where students have to select some general courses like mathematics, English, religion and communication skills. Starting from the 3rd semester, students have to take some computer specialized courses with general courses and the final and last stage the graduation, students are needed to take 16 compulsory subjects and 7 elective courses according to their choice. For this case study, the classification method will be utilized to build the model to anticipate the low accomplishing under studies in the beginning phase. Furthermore, this model will be used to recognize the courses in the program which are considered to be the best indicators of student's performance. At the early stage, prediction of performance authorizes awarding the full support to the student's and remove the problems near graduation. The department will focus on the recognition of the courses to understand the problems faced by students.

B. K. Bhardwaj, et.al [25] built the model to predict the student's academic performances that were enrolled in the bachelor's program in the five colleges in Awadh University in India. The data set contained the 300 student's records including the 226 males and 74 females and the attributes were academic, social, and demographic and psychological. Naive Bayes classification techniques were used to build the model. It had been established that the factors which highly influence the student's performance were, the place of residence, instruction language and the secondary school grade. The study used the Bayesian classification method on the student database to predict the student performance on basis of the previous student databases. This model can be used to recognize the students who need special attention to overcome the failure and take the right decision at the right time. This model showed that the student's efforts are not always the key factor of the student's academic performance.

H. Hashim, et.al [26] proposed the classification model which was used to predict the student's performances that were enrolled in the mathematical science and statistics department. The data set was obtained from the 124 students of the mathematical science and statistics department. Iterative Dichotomiser 3 (ID3) decision tree algorithms were used to predict the performance of university students.ID3 algorithm because it may be a well-known commonly used data mining technique. Student information, such as their previous degree records, last year performance based on different attributes and the faculty results are collected to predict the performance of students throughout the study plan. WEKA software supported this application of study [26]. The ID3 algorithm attained high accuracy in testing the data. Despite other classification algorithms; the ID3 approach was considered as being the best approach of choices for primary analysis to predict the student's performance because of its simplicity and ease. The outcomes of this study disclosed that it is possible to predict the chances of obtaining the degree in the estimated time according to the degree of graduate in the performance attributes. In the last, concluded that the ID3 algorithms which were used in this study [26] can be helpful and efficient if implemented regarding the application of student's performance.

Amal Alhassan et.al [27] presented the model to evaluate the effect of evaluation and activity factors from the learning management system (LMS) on student's performance. The dataset contains only 241 rows and 19 attributes. Processes of selecting the features, various algorithms for feature selection were applied to recognize the features that were highly influenced the student's academic performance. Moreover, the reason for building the prediction models was to support evaluation and activity data jointly. Web-based learning emerged as a result of the increased usage of the internet in education. An educational institution can use an LMS to manage students, visualize their involvement, and evaluate their development. According to the findings of this study, evaluation grades such as assessment marks and final grades are the most essential factors influencing a student's academic achievement. When it comes to predicting student academic success, this study's findings show that both the model base and sub-model work well. When compared to other classifiers, Random Forest is rated as the best classifier for predicting student achievement. A decision tree is used after the basic model and sub-model that achieve the maximum accuracy degrees. When comparing the random forest and the decision tree techniques, the random forest does not produce as intelligible results as the decision tree. Furthermore, the factors that are included in evaluation grades are put alone or together with the activity data the prediction model performs better. Moreover, the random forest algorithm performs well when the student's performance needed to be predicted.

Leon Gerritson et.al [28] proposed a forecasting model to predict and analyze student's performance based on four classification models. This model was constructed using four data mining techniques, Naïve Bayes, Logistic Regression, Artificial Neural Network and Decision

Tree. The main purpose of this research is to identify those students who need extra attention and mentorship to enhance their performance. Recommended actions would be taken against average students such as taking extra sessions, providing lab facilities and other technical facilities. The Neural Network performed very well out of other classifiers with 61.15% accuracy.

Catarina et.al [29] implemented a survey strategy and gathered answers from the current university students. They have proposed the model based on random forest and decision trees, which are basic machine learning techniques. Supervised learning issues have been focused on by the researchers to develop a prediction model. The proposed model results in reducing the student dropout ratio, and enhancement of student's average performance in the semester. The model also suggests semester planning that will not only enhance student's motivation towards studies but will also increase their grade point average (GPA). The major drawback of this model is that it is implemented only one semester, not on the complete curriculum of the degree. Other machine learning techniques could also be implemented to receive better results.

Alyahyan, E. et.al [30] performed a survey analysis for university students to gather the information on which data mining algorithms were applied to predict the performance of students. To predict the student's performance is now a very hot topic for a country's higher institutions so that the authorities could design a plan to achieve the maximum output from the students. In this paper supervised learning methodology has been used. Three algorithms were implemented and their results were compared i.e. Naïve Bayes algorithms, Multilayer Perceptron and Decision Tree. The prediction accuracy of Naïve Bayes algorithms was 76.65% as compare to Multilayer Perceptron which has 71.2% and Decision Tree which has an accuracy of 73.93%. More over the correctly classified instances was 197/257 by the Naïve Bayes algorithm as compared to 183/74 by Multilayer Perceptron and 190/67 by Decision Tree. The authors concluded that a Naïve Bayes algorithm outperforms, Multilayer Perceptron and Decision Tree.

Rimadana et.al [31] suggested a novel approach to predict the student's performance. The author had used Time Management Skills data, which had been derived from Time Structured Questionnaire (TSQ) to predict student performance. Five different data mining algorithms were implemented using Time Structured Questionnaire (TSQ). Linear Support Vector Machine (SVM) proved to be more efficient as compare to other algorithms and predict the student's performance with 80% accuracy. In addition to overall accuracy, the algorithm also predicted the English performance of students with 84% accuracy.

Alana M.de Morais et.al [32] presented a novel approach in the virtual learning environment (VLE). The research is carried out to facilitate teachers to make decisions in absence of statistical tools in the virtual learning environment. Two steps were defined while implementing the model, clustering and prediction analysis. Clustering had been implemented by using K-means algorithms. On the other hand, regression methodology is being used for prediction. These two algorithms were used to support teacher decisions in the virtual learning environment when statistical analysis tools are not present.

Zeineddine H et.al [33] presented a novel approach in Educational Data Mining (EDM) for Portuguese university students by using four data mining that are Decision Tree, Support Vector Machines, Neural Networks and Random Forest. The author has forecasted the grades of students for two main classes which are statistics and mathematics by making previous grades the target class. It was concluded by the author that a higher accuracy can be achieved if the previous grades of the students are known.

Uday Kumar et.al [3] implemented clustering and classification techniques to measure the student's performance of K L University for the batch 2013-2017. They have performed clustering based on student's intermediate marks and B.Tech percentage of 3 semesters. . Three clusters have been made namely poor for students who have below than 85% in intermediate and B.Tech, good for student's having in between 85% - 90% marks in intermediate and B.Tech, and excellent for student's having intermediate and B.Tech marks above than 90%. K-means clustering algorithm is used to form clusters on B.Tech and intermediate marks of 200 students. Agglomerative clustering which comes under hierarchal clustering is also implemented on 200 students based on their intermediate and B.Tech marks. K-means outperforms hierarchal clustering as it achieved high performance on the larger datasets as compare to hierarchal which is moderate. Decision Tree which comes under supervised learning is also implemented on 200 students. The subjects on which classification is performed were DBMS, Operating Systems, Computer Networks and C. The dataset of 200 students was divided into 100 tuples for test and 100 tuples for training. The accuracy of the decision tree came out to be 91%. The second classification algorithm was Naïve Bayes that was implemented on the same dataset; the accuracy comes out to be 72%. So the decision tree outperforms the naïve Bayes algorithm for the given dataset.

Harvey et.al [34] implemented classification algorithms on students of 12th standard to predict the accuracy of math scores. The dataset collected for this study was collected from the Massachusetts State Department of Elementary and Secondary Education and the Census Bureau. Some of the features of this dataset were school finances, demographics, and teacher salaries, grades and so on. Initially, the size of the dataset was 1861 with 303 features. After cleaning the dataset the size was reduced to 403 with 27 features. The dataset was split into two parts, 70% for training and 30% for testing. Math score was set as the target class and 3 classification algorithms were implemented. Linear regression was implemented firstly on the given dataset and the accuracy comes out to be 52%. After that decision tree was implemented on the given dataset and the accuracy comes out to be 59.8%. Naïve Bayes outperforms the other classifiers and has an accuracy of 71.0% as compared to other classifiers.

Guo, B et.al [35] this study proposed a Deep Learning-based algorithm for predicting student performance. Use an unsupervised learning algorithm sparse auto-encoder to pre-train hidden layers of features layer by layer in advance. Students performance prediction network (SPPN) is designed to forecast student performance exploiting recent developments. [35] This system is especially proposed when huge data sets are available, the Deep Learning methodology is a particularly successful strategy for predicting outcomes with a high level of accuracy There are millions of parameters to train in SPPN, which involves huge amounts of computational capabilities graphics processing unit (GPU) is employed because of its parallel architecture for the quick processing of data. According to our understanding, SPPN is the first deep learning system for the educational prediction that is GPU implemented. Our deep learning approach is implemented using a 6-layer neural network in SPPN. Typically, a network has one input layer, four hidden layers, and we pre-train hidden layers of features one at a time. One hundred and twenty thousand students were used to test the proposed model. Various sources of educational data can be used to acquire data of various forms and variables. Hubei province junior high school students provided us with real-world information. In each school, 1200 students have sampled of the grade 9 students for the recent three years. The unlabeled data is then utilized to discover features using an unsupervised learning approach called sparse auto-encoder. To train SPPN, the recently introduced technique known as a dropout is applied. SPPN is written in ANSI C and uses Theano, a Python package that makes GPU utilization transparent. SPPN is trained on a dataset of around 120, 000 labelled students, with training parameters displayed. To compare results with SPPN on our dataset, we used three available classification algorithms: Naive Bayes, Multilayer Perception (MLP), and SVM. With an average accuracy of 77.2%, SPPN has the best accuracy of the algorithms. MLP suffers from significant overfitting in traditional neural networks. The other two shallow models, SVM and Naive Bayes are unable to differentiate as well as SPPN. The findings of the experiments suggest that our method can be used in educational settings to identify specified events, such as pre-warning for at-risk children. In addition, we examine the training efficiency of GPU and CPU. SPPN-g is a GPU-trained network that is around 9 times faster in the training process than SPPN-c, which is completely CPU-trained. Although the result of SPPN-c is considerably better than that of SPPN-g, SPPN-c took nearly two and a half days to train, whereas SPPN-g took only six hours. If the training set grows in size in the future, GPU parallel architecture will be required for resolving.

J.Dhilipan et.al [36] presented a model to predict the student's performance based on their 10th, 12th and previous semester grade. To predict the student's performance, the system has been assigned 3 tasks. Students who have below 50 % in 10th and 12th, the student who have not to clear the internal and student who is irregular. If any one of the cases is achieved by the student after running classification algorithms, a student may not be able to complete his/her degree. This study [36] has been implemented by using a Decision tree, Binomial logical regression and K-NN classifier. Out of the before mentioned classifiers, Binomial logical regression has the highest accuracy of 97.05%.

Tanuar, E et.al [37] this study experiments on the students of Bina Nusantara University's School of Computer Science. The data used in this study were from graduates in the years 2017-2018. The final GPA (Grade Point Average) average is used for the integration framework. These results were also divided into 2 trials. The first trial predicts directly based on true GPA, whereas the experiment includes GPA into six groups based on the rule GPA >=3.50, 3.50 – 3.00, 3.00 – 2.75, 2.75 – 2.00, 2.00 – 1.80, and 1.80. The classification is based on the graduation standards, which include High Distinction, Distinction, and a minimum GPA of 2.00 to graduate. In this study, Rapid Miner is used as the platform. This study will employ three models: the General Linear Model, Deep Learning, and Decision Tree. Preprocessing data requires gathering final results from first-semester subjects and combining them with the final GPA. The model was created and tested using Rapid Miner. The results of this experiment were classified into two categories: first, a GPA experiment, and second, a GPA group experiment. The outcome of the first type of data, "Result on GPA," in which 1.835 data was analyzed to determine the GPA result. The Deep Learning approach (Root Mean Square Error value of 0.276) was found to be more accurate than the others, while the Decision Tree method (run time of 11s) was the fastest. The tests also continued to decrease two data points that were less significant than the previous experiment's findings, but the outcome did not improve. The Root Square Mean Error for the Generalized Linear Model become 0.309, according to the findings of the second type of experiment, "Result on group GPA.", Deep learning now has a score of 0.295, while the Decision Tree has the same score of 0 because it has no effect. The less significant variables, English and Character Building, were also excluded for the experiment, but indicated little improvement, with the Generalized Linear model increasing to 59.9%, Deep learning increasing to 59.1%, and Decision Tree increasing to 52 %. In determining the success of the GPA experiment, another two programming subjects, Data Structure and Object-Oriented Programming, were added to the advanced experience. The results show that the Generalized Linear Model improved by 66.6 %, Deep Learning improved by 67.6 %, and Decision Tree improved by 60.6 %. This data mining research, which used Rapid Miner to predict Final Year GPA based on 1st-semester results, show that grouping GPA is more efficient in terms of run time than other types of data. By grouping the results, it is easier to forecast the outcome and for students to understand their abilities. The decision tree is the most straightforward model to explain to users. The experiment also reveals that the outcome of the algorithm, discrete mathematics, and calculus subjects are the most essential components. As a result, students who want to do well in computer science must achieve a higher grade or a deeper understanding of those courses, as they represent the basic knowledge required. Along with the results, various types of data and information, such as final examination marks or other psychological assessments, can be added to the data for future research. In addition, by combining the performance results from the second semester, the progressing prediction may be developed for future study. Finally, data mining from earlier student experiences can be used to provide knowledge to future students. The next generation of scholars will be prepared and able than previous generations since they will have access to and understand the material.

Tarik, A. et.al [38] this study proposes a system that using a recommendation system and Artificial Intelligence approaches, an intelligent way to evaluate the performance of Moroccan students in the Guelmim Oued Noun region during COVID-19. Algorithms allow computers to make decisions instead of people, which is the foundation of artificial intelligence. In a variety of ways, this technology improves the user experience. The Ministry of National Education has started the project "Set up an early guidance system and active effective" from the start of school in 2019-2020, which is one of seven programs aimed at enhancing the unique abilities that he holds. Through group lessons and individual interviews with learners, as well as providing educational support service and teaching space that institutionalizes the learner's project, the educational school counsellor maintains a teaching space that fertilizes the learner's project. Educational advising is a critical component of governance and a vital approach for improving the human element's qualification and simplifying decisions by placing the right person in the right position. The Moroccan educational system is organized into three levels: primary school, secondary school, and high school, the latter of which is a three-year cycle. Students in their ninth year of basic education who are serious about continuing their education in a general or technical education section are accepted. This cycle is for those between the ages of 16 and 18. The goal of the project is to create and execute a framework that will allow students in the common core to pursue one of the first baccalaureate's technical or scientific specialties. The method's goal is to provide a good orientation that will enable the student to achieve a good outcome based on an already existing system that includes a group of students who have already completed their baccalaureate in the Guelmim Ouad Noun region. The regional academy of education and training provided us with the grades of 142110 students between 2000 and 2015 after filtering the data (removing missing values and students who did not complete their baccalaureate). Our research's main purpose is to develop an efficient system for calculating the baccalaureate average connected with the existing model. Students' grades are stored in the Gestion Notes database at three levels: Common Core, first-year baccalaureate, and second-year baccalaureate. It's a matter of choosing from a variety of machine learning models that best model the variable to be explained (business problem). This research create three regression algorithms in the execution portion to produce a good prediction: Random Forest, Linear Regression, and Decision Tree are three types of decision trees. At the end of the first year of lycée (common core), students are advised to choose a study program depending on their preferences and talents, while keeping in mind the studies chosen for the bachelor. The goal of these systems is to provide a good orientation that will assist students in receiving a good grade based on an existing model that includes all students who have already completed their baccalaureate in the Guelmim Oued Noun region. According to this study, the random forest is the best model for forecasting the baccalaureate average using a regression method. The Random Forest approach has a better average score since it consists of several separate decision trees.

2.2 Literature Review Table

Table 2.1: Existing Studies related to research (this table shows the existing literature related to this research)

Sr #	Article Name	Year	Description	Outcome/Limitation
09	The prediction of students' academic performance using classification data mining techniques	2015	This research establish the best students' academic performance prediction model, Decision Tree, Naive Bayes, and Rule Based categorization approaches are applied to the data of the students. The prediction model's obtained knowledge will be utilized to identify and profile individuals in order to estimate their level of success in the first semester.	In comparison to the NB and DT, the experimental results suggest that the RB has the superior classification accuracy. The methodology will allow instructors to intervene early to help and assist students in the bad and average categories improve their grades. The low amount of the data in this study is a limitation due to incomplete and missing values in the obtained data.
14	Student Performance Analysis Using Clustering Algorithm	2016	A simple Clustering analysis is proposed through which the behavior of student is understood. To understand if there is a	Since the implementation is easy and the computational efficiency is high, K-means algorithm is used is clustering.

			regular improvement in the performance of student, a good benchmark has been set up by the data mining algorithms.	
15	Students Performance Prediction Using KNN and Naïve Bayesian	2017	Prediction model is proposed by focusing on the KNN and Native Bayes algorithms. Evaluations are performed here by making comparisons against KNN and Native Bayes in terms of certain performance parameters.	In comparison to KNN, the performance of Native Bayes is better which states that amongst to features affecting the performance of students; a strong relation is identified through which the performance of students can be predicted.
16	Analysis of educational data mining. Advances in Intelligent Systems and Computing	2019	This study analyzed the student’s achievements, development and abilities using the different analysis techniques. Initially, Student Attribute Matrix (SAM) was used to develop the student model accompanying the performance and non-performance related attributes. When providing the student’s performance estimation tool, the BP-NN algorithms were applied.	Teachers examine the results given by the prediction tool. Students' performance is improved, and the teacher assists them in passing the course by making appropriate decisions. The administrator or principal examines the students' results by class. As a result of these findings, notices are sent to teachers. This administrator can also decide on the course that needs to be offered to students.
17	Analyzing undergraduate student’s performance using educational data mining	2017	This research studied undergraduate student’s performance by making use of data mining techniques. Initially, the academic achievement of the students was estimated after completing the four-year graduation program. Moreover, typical development was observed and all the prediction results were combined with them.	This study provided a timely warning for low achieving students and supported that by focusing on fewer numbers courses the performance can be improved which help in indicating that the performance is good or not.
18	Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron	2017	This paper used WEKA to assess the accuracy, performance, and error rate of linear regression and multilayer perceptron in predicting students' final exam results. Extraction and analysis of e-learning logged-posts in discussion forums and attendance	Final examinations were predicted in this paper using two classification algorithms, linear regression and multilayer perceptron, and the magnificence of the two was compared using the value of mean absolute error difference. The results revealed that multilayer

			provided the foundation for the data. Furthermore, this forums and attendance provided the foundation for the data. Furthermore, this research not only predicts performance (on the final exam), but also compares the two models to evaluate	perceptron outperform than linear regression in terms of prediction accuracy. With lower mean absolute error and root mean squared error, multilayer perceptron gives better prediction results than linear regression.
19	Prediction of Students Performance using Educational Data Mining	2016	The suggested system is a web-based application that extracts important information by employing the Naive Bayesian mining technique. It also records all student entrance details, course details, subject details, student marks details, attendance details, and so on. It uses a student's academic background as input and predicts their future performances on a semester-by-semester basis.	This research can help students and professors in motivating students of all types to achieve effectively. This research helps to identify students who require extra attention, lowering the failure rate, and preparing for the future semester's examination. In the future, data processing approaches for extra usual qualities will be used to encourage accurate and cost-effective outputs.
20	using data mining to predict secondary school student performance	2003	The proposed system that addressed the prediction of secondary school student grades by utilizing data mining techniques. Three data mining methods such as binary, regression and five-level classification were used along with four techniques i.e. Decision tree, Random Forests, Neural Networks and Support Vector Machine (SVM) were examined.	The outcomes revealed that only if the first and second school grades are known then it’s possible to attain excessive predictive accuracy. This study validated the assumption that student success is highly influenced by earlier performance. In this study after the data was collected the data mining techniques were applied that’s why this study focuses on off-line learning.
21	Mining Educational Data to Improve Students Performance : A Case Study	2012	This paper presented a case study in educational data mining. This study used data mining techniques to discover knowledge, specifically discovered association rules, and used the K-means clustering algorithm to group the students into groups, as well as two classification methods, Rule Induction and Naive Bayesian classifier are used.	The model produced by applying the Naive Bayesian classifier has an accuracy of 67.50 %, which is satisfactory accuracy, and it is proposed that the Naive Bayesian approach be used to predict the student's grade. These two techniques have the advantage of being able to forecast lower grades ahead of time.

23	Analysis of Student Result Using Clustering Techniques	2014	This study using Clustering techniques of data mining, will examine how various factors influence a student's learning behaviour and performance. This study is an effort to provide motivation for using data mining technology to advance the traditional educational process. The model is also given as a tool for higher education institutions to use in order to improve their decision-making processes.	A data mining system that can give the needed knowledge and insights for decision makers in the higher educational system is required to accomplish quality improvement. As a result, this knowledge benefits the higher education system by increasing educational system efficiency, decreasing student dropout rates, increasing student promotion rates, increasing educational improvement ratio, increasing student success, increasing student learning outcomes, and lowering system costs.
24	The impact of engineering student’s performance in the first three years on their graduation result using educational data mining	2019	Predictive analysis was used in this study to see how well the fifth year and final (CGPA) of engineering students at a Nigerian university could be predicted using the first three years of study as inputs into a Konstanz Information Miner (KNIME) based data mining model. The program and the year of entry were used as predicted inputs in a KNIME workflow, which was evaluated using six different data mining methods.	The results of the KNIME-based predictive model show that the logistic regression predictor achieved the maximum accuracy of 89.15 %, random forest predictor had the fourth greatest accuracy of 87.70 %, decision tree predictor had the third best accuracy of 87.85 % and PNN predictor had the lowest accuracy of 85.89 %, while the Naive Bayes predictor had the highest accuracy of 86.438 %.
25	A prediction for performance improvement using classification	2011	The classification task is used on a student database in this paper to forecast student division. The decision tree method is utilized here because there are several approaches to data classification. The data such as attendance, class test, seminar, and assignment marks were gathered from the students' previous database.	It assists in detecting dropouts and students who require particular attention earlier in the school year, allowing the teacher to provide appropriate advice and counselling. This research will also look for students who require extra attention in order to lower the number of students who fail and to take necessary action for the next semester's exam.

26	Data mining methodologies to study student's academic performance using the C4. 5 algorithms	2015	The purpose of this research was to develop a predictive model that might be utilized to help students improve their academic performance. To accomplish this purpose, used data from previous students' academic records. Despite the availability of various classification algorithms, the C4.5 technique has been the primary data mining algorithm due to its simplicity and ease of implementation.	According to the data, recall and precision bring estimations closer together. Because FP replaced FN in the denominator of the precision matrices, the accuracy does not necessarily decrease linearly as the level of recall varies. The results were positive (above 70%), indicating that the C4.5 algorithm is an effective and reliable technique that should be promoted.
27	Predict Students ' Academic Performance based on their Assessment Grades and Online Activity Data	2020	The major goal of this study is to see how assessments and activity features from a learning management system affect students' academic performance. Five classification methods are used: decision tree (J48), random forest (RF), sequential minimum optimization (SMO), multilayer perceptron (MLP), and logistic regression (Logistic).	The result reported that assessment data has a considerable impact on student achievement, whereas activity data has a less impact. However, combining assessment and activity data improves the accuracy of the prediction model. In addition, the random forest algorithm outperforms the decision tree in terms of predicting student academic success.
28	Predicting Student Performance with Neural Networks	2017	The goal of this study is to see if Neural Networks are a good fit for predicting student performance. A Moodle log file comprising log information about undergraduate courses as the study's dataset. On this dataset, compared the predictive performance of Neural Networks against six different classifiers like Naive Bayes, k-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression are examples of classifiers.	Half of these course classifiers outperform usually trained classifiers, according to the results. Individual predictors were also studied for their importance in classification, with previously obtained grades contributing the most to successful predictions. It can be concluded that Neural Networks outperform the other six methods tested on this dataset and might be used to forecast student performance well.
29	Predicting students' performance using survey data	2020	One of the most significant challenges that computer science students confront is learning how to construct	The problem was approached using machine learning techniques (decision trees and random forest). When

			computer programs. Students drop out of subjects and occasionally entire courses as a result of their inability to learn the necessary skills. There is a need to investigate the causes of student success (or failure) in initial curricular units in order to look for determining behaviours or characteristics and thus try to mitigate and modify them.	compared to various baseline models, the findings indicate that machine learning can be employed for the task, as it produced models with excellent accuracies. Future studies will aim to expand the databases by include data from more semesters. Plan to employ a variety of machine learning approaches in order to create even more efficient models. The results help us to prepare the semester accordingly, anticipating how many students may require more assistance.
30	"Predicting academic success in higher education: literature review and best practices	2020	This study performed a survey analysis for university students to gather the information on which data mining algorithms were applied to predict the performance of students. In this paper supervised learning methodology has been used. Three algorithms were implemented and their results were compared i.e. Naïve Bayes algorithms, Multilayer Perceptron and Decision Tree.	The prediction accuracy of Naïve Bayes algorithms was 76.65% as compare to Multilayer Perceptron which has 71.2% and Decision Tree which has an accuracy of 73.93%. More over the correctly classified instances was 197/257 by the Naïve Bayes algorithm as compared to 183/74 by Multilayer Perceptron and 190/67 by Decision Tree. The authors concluded that a Naïve Bayes algorithm outperforms, Multilayer Perceptron and Decision Tree.
31	Predicting Student Academic Performance using Machine Learning and Time Management Skill Data	2019	This study uses different machine learning models to predict student academic achievement. Other data had previously been used as a feature in creating predictions, but the TSQ result had never been used as one. Five different classification methods were used: Linear Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Neural Network (NN), and Naive Bayes (NB).	In terms of classification, Linear SVM outperform other models, both in terms of predicting Academic Performance and English Performance. SVM predicts academic success with an accuracy of 80%. Meanwhile, SVM predicts English performance with an accuracy of 84 %. In future, more data will be used to test prediction accuracy. Other classification models can also be tested to see if they can provide better prediction results.
32	Monitoring Student Performance Using Data Clustering and Predictive Modelling	2014	This paper discusses how to deal with e-learning data using an analytical method. The major goals of this article were to establish	Several variables were discovered in this study that described the tutor in separate groups. As a result, the teacher's judgments (based on

			which criteria, in each group, are the most relevant for the tutor's assistance in order to direct a student to future learning activities. For e-learning challenges, there are several strategies that can be used.	the responses of all students) can be efficient on the VLE. For future work, this study will validate the student groups in various contexts and analyze the subgroups.
33	Enhancing prediction of student success: Automated machine learning approach	2021	The use of Automated Machine Learning to improve the accuracy of predicting student performance using data available before the starting of the academic program is proposed in this research. For a given decision-making situation, Automated Machine Learning (AutoML) is a technique for determining the appropriate classification model.	As a result of the increased accuracy in predicting students at risk, academic institutions can be more efficient in supporting those student. In the future, descriptive statistics may be used to investigate the role of various psychographic variables and their impact on the predictive model. Predicting failure of new-start students has a maximum accuracy rate of 70%. Auto-generated Ensemble Model correctly predicts failing students.
34	A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning	2019	The purpose of this study is to present a strategy for determining the most important characteristics of good student performance in grades K-12. This research will present a prediction model that can be applied to a range of data sets, as well as how the data can be used to develop solutions for improving K-12 school performance. To determine which classifier has the best prediction accuracy for the current dataset, three classifiers will be evaluated: linear regression, decision tree, and Naive Bayes.	When it came to forecasting SAT Math results for high school students, the Naive Bayes approaches had the best accuracy. Stakeholders in K-12 education can utilize the results of this evaluation of current research and the models described in this paper to make predictions about student performance. To produce better predictions of student performance, the linear regression model can be enhanced. The naive Bayes classifier model had a 71.0 percent accuracy. This was the most effective categorization strategy for predicting SAT math test scores in this study.
35	Predicting Students Performance in Educational Data Mining	2016	In this paper, describe a deep learning architecture for forecasting student performance that uses unlabeled data to learn various levels of representation automatically. This study used a sparse auto-encoder to layerwise learn hidden	According to the findings, SPPN has the best accuracy among the algorithms, with an average accuracy of 77.2 %. MLP suffers from significant overfitting in traditional neural networks. The other two shallow models, SVM and Naive Bayes, are not capable of discriminating as well as

			layers of features, and then used supervised training to delicate the parameters. Train the model on a large dataset of real-world students in this study.	SPPN. The findings of the experiments suggest that our method can be used in educational settings to identify certain events, such as pre-warning for at-risk children.
36	Prediction of Students Performance using Machine learning	2021	The primary purpose is to have a concept of the artificial intelligence systems that have been employed to forecast academic learning. This study also looks at how to utilize a prediction algorithm to determine the most important features in student data. This study uses four strategies and using these strategies find the accuracy.	The accuracy level of the student prediction system was determined in this study. Four strategies are used to analyze the data that has been gathered. Binomial logical regression has a 97.05 100 % accuracy, Decision Tree has an 88.23 100 % accuracy, Entropy has a 91.19 100 % accuracy, and K-NN has a 93.71 %. Additional features will be added to our dataset in the future to improve accuracy.
37	Using Machine Learning Techniques to Earlier Predict Student’s Performance	2019	The students of Bina Nusantara University's School of Computer Science were used in this investigation. The student's final year outcome (GPA) can be predicted using machine learning algorithms based on their first semester results. The Generalized Linear Model, Deep Learning, and Decision Tree approaches were used in this experiment.	By grouping the results, it is easier to forecast the outcome and for students to understand their abilities. The decision tree is the most straightforward model to explain to users. The results show that the Generalized Linear Model improved by 66.6 %, Deep Learning improved by 67.6 %, and Decision Tree improved by 60.6 %.
22	An Analysis of students performance using classification algorithms	2017	The paper's main purpose is to use the Weka tool to analyze and assess school students' performance using data mining classification methods. The paper's major goal is to research and analyze the performance of school students using data mining approaches.	As a result, it is possible to conclude that the Random Forest algorithm outperforms other algorithms. Based on the correctness and timeliness of school students' performance analysis based on their dataset it is proven that Random Forest is the best classifier because it takes less time and has a high level of accuracy.
38	Artificial intelligence and machine learning to predict student performance during the COVID-19	2021	This study proposes a system that using a recommendation system and Artificial Intelligence approaches, an intelligent way to evaluate the performance of Moroccan students during COVID-19. Algorithms allow	This research create three regression algorithms in the execution portion to produce a good prediction: Random Forest, Linear Regression, and Decision Tree are three types of decision trees. According to this study, the random forest is the best model for forecasting

			computers to make decisions instead of people, which is the foundation of artificial intelligence.	the baccalaureate average using a regression method.
29	Predicting students' performance using survey data	2020	One of the most significant challenges that computer science students confront is learning how to construct computer programs. Students drop out of subjects and occasionally entire courses as a result of their inability to learn the necessary skills. There is a need to investigate the causes of student success (or failure) in initial curricular units in order to look for determining behaviours or characteristics and thus try to mitigate and modify them.	The problem was approached using machine learning techniques (decision trees and random forest). When compared to various baseline models, the findings indicate that machine learning can be employed for the task, as it produced models with excellent accuracies. Future studies will aim to expand the databases by include data from more semesters. Plan to employ a variety of machine learning approaches in order to create even more efficient models. The results help us to prepare the semester accordingly, anticipating how many students may require more assistance.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Overview

Research methodology deals with the methods adopted by the researchers to obtain the outcomes from the research. Two main important research methodologies have been widely in use. These are

- Quantitative research
- Qualitative research

Below is a brief introduction to these methodologies.

3.1.1 Qualitative Research

Qualitative research is a market research technique that focuses on gathering information through open-ended and conversational interaction. Qualitative research takes a humanistic or idealistic approach to address a study question. The qualitative technique is used to gain a better understanding of people's beliefs, experiences, attitudes, behaviour, and interactions. It produces non-numerical data. It can be utilized to gain in-depth insights into a challenge or to generate research ideas. Qualitative research includes learning more about how people see the world. While there are numerous techniques to qualitative research, they all tend to be adaptable and focused on preserving rich meaning when evaluating data.

3.1.2 Quantitative Research

Quantitative research deals with the methodology of gathering information based on numerical data. This methodology is used to testify the hypotheses, assumptions and theories.

Statistical analysis is being deployed in the form of graphs for validation purposes. In quantitative research numerical analysis is used to forecast and measure variables of interest. Quantitative research helps the researchers to extract the idea from their research and then predict the particular results drawn from the conclusion.

As stated earlier, it works with numerical data and is then used to make predictions. Using survey methodology, we gathered data from the target audience. The acquired data is identified, and a target class for predictions is chosen. Applying the WEKA (Waikato Environment for Knowledge Analysis) tool, which is designed for data mining implementations, we conducted tests using several machine learning classification methods to determine the accuracy of different classifiers on our dataset.

3.1.3 Surveys

The researchers have used surveys to gather information from a bigger group of people. Regarding the collection of data, several data and quantitative analyses were carried out to provide relevant research results. The researcher uses a survey, which is a quantitative approach, to gather information on specific scenarios by asking multiple-choice questions. Before sending the survey to the target audience, researchers had to define the questions that revolved around his research interests.

We must collect and analyze actual data regarding prediction while we work on educational data mining. One more factor that came into the way while dealing with educational data mining is the geographical location. As mostly work done in educational data mining is from Europe and America. We can take the existing (available) dataset, but we want to see the effect of those features in our environment. So, the survey was the most suitable way to gather information regarding educational data mining from the large pool of target audiences. We have used Google forms for the creation of the survey questionnaire.

Out of different survey conducting methodologies we have adapted online survey conducting methodology. Due to the covid-19 pandemic, universities have adapted the policy of “learn from home”. So, to collect information from the target audience we have used social media platforms to distribute the survey.

3.1.4 Other methods for Conducting Survey

There are some other methodologies other than online for conducting a survey which is as follows.

1. **Telephone**

The telephonic survey was once the popular form of surveying a particular topic or product. But with the advancement of technology and availability of the internet, this methodology has lost its importance. Moreover, it is very difficult to conduct a telephonic survey when you have a large size of the target audience. It is costly and requires much time to conduct.

2. **Face to Face**

Face to face methodology is an effective way of conducting the survey. But it requires much time and cost order to perform face to face survey. The major advantage of conducting this type is that the resulting dataset will not have any missing values or outliers.

3. **Print or Paper**

Print or paper methodology is also another form of conducting the survey. But this form of methodology requires time and expense. Especially when you have a large size of target audience it cost you much high to perform this methodology.

3.2 Analysis

The data analysis and findings of our research investigation are discussed in this chapter. The questionnaires used in our research study were usually required to ensure that the information gained was presented clearly, with percentages and graphs where appropriate. The main objective of our research project was to identify the characteristics that influence a student's achievement. To make the data presentation understandable, the data acquired from students studying at various universities were subjected to analyzing the data gathered, converted to percentages, and gathered in the form of tables, graphs, and figures. Our research questions were used to analyze the information.

3.3 Flow Chart of Dataset Processing

This flow chart shows the workflow of processing of dataset. The workflow for processing of dataset is depicted in Figure 2.

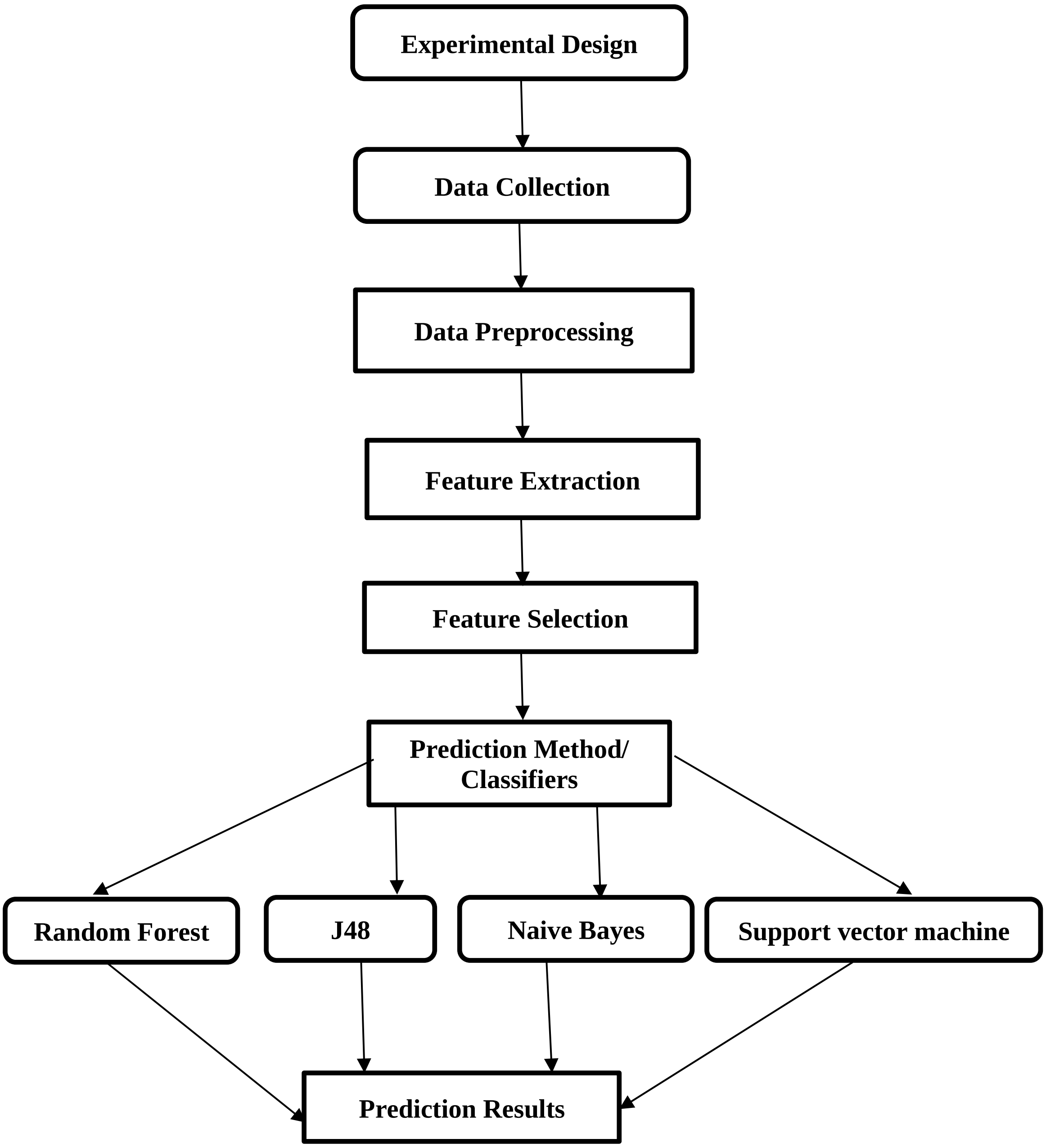


Figure 3.1: Dataset processing workflow (This flow chart shows the workflow of processing the dataset.)

3.4 Predictive Analysis

Predictive analytics is the process of applying a model developed from similar past data to predict future occurrences and behaviours in previously unknown data. Finance, education, healthcare, and law are just a few of the areas where they might be used. In all of these fields, education is the field where the most of the researchers were used the predictive analysis. This research used the predictive analysis technique to predict the student grades on the basic of their previous grades. This research used different machine learning techniques to predict the student grades. A machine learning system establishes relationships between unique features of student data using previously obtained data. The generated model can identify one of the database's attributes. In this research, to forecast the future trends and behavior patterns, used the predictive analysis technique and this technique deal with the gathering of data. In the next section discussed about how and where to collect the data for applying the predictive analysis technique.

3.5 Data Collection

Data collection methods refer to the procedures used to gather information about variables. In this research, we applied a sample survey methodology to gather information required from the students while using the questionnaire technique to gather the data from the students. Our research is statistical research which is based on data collection, where primary data can be collected. Due to the COVID-19 pandemic, we have implemented the questionnaire methodology to gather data from students of different universities. Google forms have been used for the creation questionnaire and after that mailed to target audiences for collecting their responses. In the next section, discussed what the sample size of the collected data is and how to find the sample size of the collected data.

3.6 Sample Size

The number of completed responses to the research survey is referred to as the sample size. It reflects a sample of the people (or target population) whose opinions or actions are engaged in. When the research survey based on the large group of people, you need to collect responses or results from everyone. Therefore, it is unlikely to get answers or results from everyone. So, this research choose a random sample of students who represents the entire population. The sample size is critical for obtaining accurate, statistically meaningful results and completing your project. The goal of this study [20] is to forecast student progress and, if possible, to identify the fundamental factors that influence educational success or failure. School reports and questionnaires were used to acquire recent real-world data (e.g., student grades, demographic, social, and school-related characteristics). Under binary/five-level classification and regression tasks, the two core classes (i.e. Mathematics and Portuguese) were represented. Four DM models (Decision Trees, Random Forest, Neural Networks, and Support Vector Machines) were also examined, as well as three input selections (with and without past grades).In this study, [7] the information was combined into two datasets: one for Mathematics (395 cases) and another for the Portuguese language (649 records) and 23 attributes. This research [7] aims to analyses the various resampling approaches for dealing with the unbalanced data problem in order to determine the best methodology and classifier for forecasting student performance. This study will also look into the differences between multiclass and binary classification, as well as the importance of feature structure. In this study, used two separate educational datasets from Iranian and Portuguese educational institutions. All relevant information about postgraduate students from Iran University of Science and Technology was manually collected and registered in the Iran dataset between the academic years 1992-93 and 2014-15. This dataset contains data on 650 students over 19 different variables. In addition, all of the data in the Portugal dataset is associated to student accomplishment in two Portuguese secondary schools. This dataset contains data on 394 students who have 19 different qualities. The Final GPA is the study's output variable. This study [2] proposes an effective automated system for classifying students into various groups or grades based on their exam results and predicting their results. The suggested model has two benefits for educational institutions: first, it may assist teachers in revising their instructional practices, hence enhancing student performance, and second, it can help students be classified. In this study there are 395 tuples and 34 attributes in the data set under consideration. Each tuple represents a student's attribute values or discusses the student's academic performance and social behaviour in depth. This study [39] explains how data mining tasks such as classification, prediction, and clustering may be applied to data from an e-learning system. The performance of sixth grade school students is used to analyse and show the findings obtained with the WEKA tool. This study involved sixty 6th students. As a result, a total of 3600 entries are kept, along with the student's current status (such as whether or not the learning material has been completed. As a result, the data is large, demanding the use of ML techniques for analysis. In this study [12], two different techniques to achieving goals are discussed. The purpose of this study is to forecast students' grades, with a particular focus on identifying students who would fail to achieve course requirements. As a result, this research focuses on two key objectives: predicting students' success or failure, and predicting students' final grades. The first approach employed in this study is based on state-of-the-art educational data mining techniques such as classification and regression analysis. In this study a total of 3,584 students were included in the study. The two separate data sets were employed in this study. The training set was comprised of data collected between 2010 and 2012 (37,005 instances) and was used to determine the best appropriate procedures and their settings. The test set included data from the year 2013 (11,026 instances) and was used to validate the procedures on a variety of data sets. The proposed system [40] is a web-based application that extracts important information by implementing the Naive Bayesian mining technique. Using Naive Bayesian, the system wants to improve students' success graphs. It also records all student enrollment details, course details, subject details, student marks details, attendance details, and so on. It uses a student's academic background as input and predicts their future performances on a semester-by-semester basis. In this study the experiment is performed on 700 students with 19 attributes. From the 2013 to 2016 academic year, student data was collected from the college Amrita School of Arts and Sciences, using the sampling method of the computer science department. In this step, data from many tables was merged into a single set. In this study [41], an effort is made to analyses the specified feature sets by gathering scholarship data from various Pakistani colleges and the specific feature which are related to student’s family, academic performance and family assets these features are mostly used in predicting the student performance. To forecast whether a student will be able to complete his degree, learning analytics, discriminative, and generative classification models are used. In this study around 3000 student records were first collected. To extract the most relevant attributes of students, pre-processing is used. We considered 776 student instances for experiments after removing inconsistencies and duplications from the dataset.

In this study, [42] various categorization algorithms were utilized to build a student SGPA prediction model based on the student's socioeconomic circumstances and previous academic performance. The goal of this research is to forecast the SGPA of B. Tech computer engineering third-semester students. The reason for using the third semester to predict SGPA is that it has been observed that some students leave out after the first year, others change their major, and some begin learning all computer-related subjects in the third semester. In this study gathered data from students doing B.Tech Computer Engineering at Punjabi University, Patiala's Department of Computer Engineering. A standardized questionnaire was used to collect data. A total of 260 students were gathered, with 17 variables including social parameters and previous academic performance.

This research dataset contains data on 394 students who have 24 unique attributes. We target the three different universities for collection of data, collect the data from 450 students, after applying pre-processing techniques we have left with 318. We use the pre-processing technique to clean the data and we use the data cleaning pre-processing technique to clean the data. In these existing studies which did use these approaches, their sample size is the same as our sample size is. The major reason for selecting this sample size is that four to five existing studies used the same sample size that’s why we use this sample size of data. In the next section, discussed about the experimental design of the study and discussed about which machine learning techniques are used.

3.7 Experimental Design

The goal of this research is to assess the effectiveness of existing popular machine learning algorithms for early detection of students who are likely to fail, as well as to look into the impact of process mining variables on the techniques' performance. Methods of classification, this research used classification algorithms that have been used in the field of education and are appropriate for data sets with imbalances. The following are the machine learning algorithms applied in the experiment: RF, LR, naive Bayes and KNN are all examples of machine learning algorithms. In the next section, discussed about dataset of the given data of the students and what are these features used to make the dataset.

3.8 Dataset

In machine learning, datasets are taken as pieces of data that is used for making predictions. As we know that this research worked on machine learning, the dataset consists of features that will cover university students to a maximum extent. In this research, applied Sampling methods which is a quantitative approach, to obtain the data. Due to the geographical location and difference of studies and facilities, we have used the survey methodology to collect the data concerning our target audience i.e. university students. This dataset contains data on 394 students who have 24 unique attributes. The final GPA is the study's output variable. Based on the grade point average (GPA) of the students, the information about the output attribute for datasets is classified into two categories. Students are divided into two groups: those who are good students and those who are average students. Our sample size is 450, after applying pre-processing techniques we have left with 318. In the next section, discussed about the detailed description of the features of this research dataset.

3.9 Data Description

In this dataset, we analyzed 24 different attributes. Student gender, living address, family size, family order, parental status, mother and father's job status, student's guardian, daily travel time, weekly time committed to studies, educational support provided by family members, co-curricular activities, internet connectivity at home, frequency of going out, health status Grade, teacher encouragement, increase/decrease in marks during COVID, suitable system, analytical skills, market-oriented coursework, and coursework that provides significant information. The data were collected from the students, find the sample size, and make dataset then in the next section, discussed about the preprocessing of the data that were collected form the students.

3.10 Data Pre-processing

Data preparation is one of the most important phases in machine learning. This process converts the raw data into a format that can be understood. In this research, collected the data from the students and this collected data contains several problems; as a result this phase can eliminate the errors/problems to make this research database easier to understand and manage. The database gathered information was saved on excel spreadsheet. Data preprocessing is the technique of machine learning used to transform the raw data into a useful and efficient format. The cleaning process entailed removing data with missing values and other problems which are further discussed in the data cleaning phase/section. In our research, data cleaning technique are used for our dataset. There are three steps involved in the data preprocessing technique which shows in Figure 3.

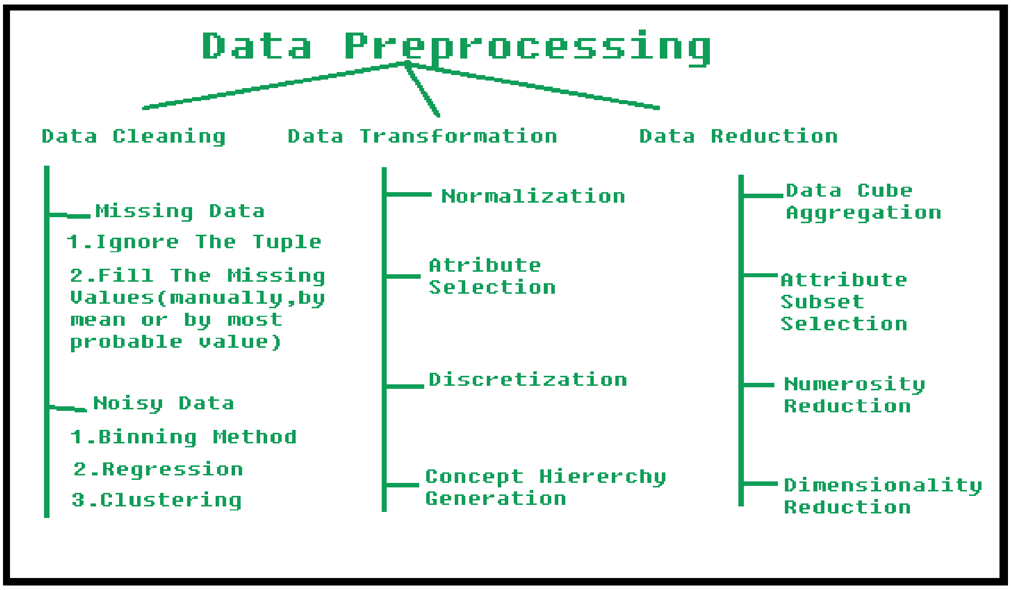


Figure 3.3: Data Pre-processing Methods (This Figure shows the methods that are used to preprocess the data. Data

preprocessing is the one of the most important phase in machine learning. This process converts the raw data into a format that can be understood. There are three methods used to preprocess the data 1. Data cleaning, 2. Data transformation and 3. Data reduction.)

3.10.1 Data Cleaning

Integrating educational data, for example, may result in inconsistencies in attributes since some institutions keep grades as letters and others as numeric values. Same like this research dataset, data may have values that are inconsistent or incorrect. These inconsistencies creates the problems and these problems effect the quality of the data and can reveal the interesting patterns. As a result, to obtain the correct and appropriate DM findings, the data must be cleaned, which may include correcting missing data, reducing noisy values, and resolving ambiguities. There are a variety of techniques for dealing with missing data, including ignoring the row when the class attribute is absent. Another option is to manually fill in the missing data using the attribute mean or a global single value like NULL. The data can contain many irrelevant and missing parts. To handle this part, data cleaning is performed. It includes dealing with the missing values and noisy data etc. In our dataset, data cleaning is used to correct or fill the missing values. To handle the missing values, there are the various way to handle that and these ways are discussed in the next.

1) Missing Values

In our dataset, some data is missing. To handle this situation there are many ways but in this research used two.

a) Ignore the tuple

For this research, used this approach because our dataset, have multiple missing values in the tuple. Using this approach, ignored the tuple which has the missing data in it.

b) Fill in the missing values

Using this approach, fill the missing values in the tuples manually.

3.11 Feature Extraction

Feature Extraction is a technique for reducing the number of features in a dataset by generating new ones from existing ones (and then discarding the original features). The original set of features should then be able to summarize the majority of the information in the new reduced set of features. From a combination of the original set, a summarized version of the original features can be generated.

3.12 Features Selection

Features are building blocks of a dataset where they have a major effect on the quality of prediction while one implement that dataset in machine learning. Feature selection is applied for the removal of irrelevant, unnecessary features from the dataset, which can greatly affect the accuracy of results while predicting. Furthermore, the feature selection method is applied to improve the accuracy of the models. Feature selection were used to examine research data attentively in exploring attributes that have a stronger impact on our output variable. Even though this research dataset do not have a huge number of features, some of them are unrelated to student achievement. WEKA comes with several feature selection algorithms. Features are considered very important in machine learning because It is a basic strategy for directing the use of variables to what is most efficient and effective for a particular machine learning system. This research used the WEKA tool to select the features because WEKA provides an automated tool for feature selection. To choose relevant attributes, classification method were employed. In the dataset, each one column represents the measurable piece of facts that can be used for analysis like Name, Age, and gender and so on. Gender, previous degree, CGPA and so on are the features/attributes of the dataset. Sometimes features are also referred to as the “Variable” or “Attributes “depending on what are you trying to analyze. There will be a few attributes in the database that do not become significant in the analysis. As a result, the feature selection method was used to remove the unwanted attributes from the dataset. There are the following features in the research dataset and discussed in the next section.

3.13 Features of Dataset

There are the following features of our dataset. These features are extracted from the features selection method.

3.14 Demographic Features

Demographic data refers to information on groups of students based on particular characteristics such as

- Gender
- Permanent Residence
- Current health status

This feature tells the gender of a student i.e. male or female. This feature tells about the audience of our dataset. Our audience belongs to two categories i.e. Male or female. These categories are related to gender class.

This feature tells about student's residence. This feature tells the permanent residential address of students. As it has a great impact on student behaviour. This feature tells that student lives in urban areas or lives in rural areas.

In this feature "current health status" deals with the current health status of the student. Either the student is handicapped or suffering from another disease or both.

3.15 Learning Environment Feature

Learning environment factor data refers to learning environment of the students where they learn from and the features are:

- Previous degree
- Level of study
- Current semester
- Marks increase or decrease during COVID-19
- Suitable system
- Encouraged by teacher
- Weekly study time
- Travel time from home to the university

This feature tells us the academic background of the student. The student can hold the degree of FSc, Bachelor or Masters. The previous degree will not be less than the FSc level and higher than the PhD level.

This feature tells the current study level of the student in the university. This feature tells the current study level of students at university whether the student is studying at Bachelor's level or a Master's level.

This feature tells the current semester of students. This feature tells about the current academic position/semester a student has. This feature tells about the student current semester.

This feature will have information about the final grades of the students before COVID-19. It tells the student's CGPA before COVID-19. This feature mostly effect the student performance because grading also gives teachers feedback on their students' progress, which can help them make better teaching decisions in the future.

This feature tells the student's performance during Covid-19. Does the covid-19 impact the student's marks? Either his/her marks increased or decreased due to study from home policy.

This feature "suitable system" deals with the education system of the university. Either the student support annual system or semester system. In universities mostly students prefer the semester system.

This feature "encourage by the teacher" describes how a teacher encourages students to improve themselves. Does it deal with the situation where teachers encourage the students during their education or not?

In this feature, we will discuss the study time of the student. How much the weekly time the students given to your studies. Students dedicate 5 to 10 hours each week to their studies or not.

This feature tells about the students travel time from home to university. The travel time is required by the student to reach at university from the place of his/her residence. This feature tells that how long it takes a student to reach the university from home.

3.16 Student's Socioeconomics Features

Socioeconomic factor data refers to information of students based on particular characteristics such as:

- Family size
- Order in family
- Parents relationship status
- Occupation of father
- Mother occupation
- Guardian
- Education expense support
- Internet facility

Family size contains the number of people living together particularly it adds siblings and parents in a family. This feature tells the family size of the students. This feature deals with the siblings with the addition of parents of the student.

This feature tells about the student's order in your family. This feature “Order in the family” has a great impact on a student’s personality. This feature tells about either the student is older in his/her family, middle one or the youngest of all.

This feature tells about the students’ parents’ relationship status. The relationship of parents has a strong impact on children psyche. This feature will help a lot of knowledge about the student. If parents are having good terms with each other it will have a positive impact on students and if their relationship has struggled it will have a negative impact. This feature tells that the students’ parents are happily living together or separated.

This feature tells about the father’s occupation. It deals with the occupation of father that either the student’s father is a serviceman (government or civilian), owns a business, farmer or a daily wager.

This feature tells about the mother’s occupation. It deals with the occupation of a mother that either the student’s mother is a Professional/working lady or a housewife.

This feature tells the guardian of the student in absence of mother and father both or any one of them. They can be maternal aunt or uncle or paternal.

In this feature "Educational expense support" deals with the educational expenses of the student. Either the student is studying on scholarship/loan, fees are paid by parents/guardians or him/her.

This feature indicates the network facility available to the students. This feature is about the internet facility used by the student in residence. Its connectivity and signal strength.

3.17 Social Features

Social data refers to information of students based on their social characteristics such as:

- Extra-curricular activities
- Go out with friends

This feature tells the extra-curricular activities of the students. Either the student takes part in extracurricular activities organized by the university or not.

This feature tells about the students outing with friends. How often does the student go on an outing with his/her friends? Either he/she goes twice a year, weekly or on monthly basis for an outing.

3.18 Course Related Features

Course related data refers to information about the course of the students which are currently studies in the university such as

- Course work provides sufficient information
- Improve analytical skill
- Course work market-oriented

This feature tells the university coursework. It deals with the coursework taught in university. Either the course work provides sufficient information to the student or not. This feature tells that students' knowledge and skills in the university course work improved as a result of the teaching materials.

This feature “course work market-oriented” is highly imported. It deals with the coursework. Either your course work is market-oriented or not.

The capability to deconstruct information into smaller categories to derive conclusions is known as analytical skills. Logical reasoning, critical thinking, communication, research, data analysis, and creativity are all examples of analytical skills. This feature is all about your analytic skills. Your analytical skills come about as a result of the study material you learned in the university.

3.19 Confusion Matrix

The confusion matrix is generated by the WEKA tool. The confusion matrix gives you a brief overview of the model's performance and accuracy level. The matrix contains detailed information about the model's predictions. The confusion matrix table, which comprises information on actual and expected classifications, is then generated. For the pass and fail categories, the prediction is correct, as seen in the matrix. The cross-tabulation of true labels from the labelled dataset and predicted labels from a classifier is called a confusion matrix.

- **Accuracy Rate**

Accuracy is a measurement that sums up how well a model performs across all classes. It's helpful when all of the classes are equally important. The ratio between the number of right guesses and the total number of predictions is used to evaluate it.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

The proportion of true negatives that are labelled as positive is known as the false positive rate (FP rate). If the forecast result class label is PASS but the true class label is FAIL, this is referred to as the FP rate. Similarly, the False Negative Rate (FN rate) is the percentage of true positives that are incorrectly classified as negative. If the forecast result class label is FAIL and the true class label is PASS, this is referred to as the FN rate.

- **Sensitivity**

The proportion of true positives that are identified as positive is known as the true positive rate (TP rate), also known as sensitivity or recall. The TP rate is calculated when the prediction result class label is PASS and the true class label is also PASS.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True positive+ False Negative}}$$

- **Specificity**

The proportion of true negatives that are detected as negative is known as the True Negative Rate (TN rate), also known as specificity. The TN rate occurs when the predicted result class label is FAIL and the true class label is also FAIL.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative + False Positive}}$$

- **False Positive**

The FP rate is a measure that is used to evaluate a model while categorizing data. The FP rate is the proportion of negative events incorrectly classified as positive to the total number of negative events.

$$\text{False Positive} = \frac{\text{False Positive}}{\text{False Positive + True Negative}}$$

3.20 Tabular Representation of Attributes

Table 2 represents the tabular form of attributes that have been used by proposed study dataset.

Table 3.1: Tabular Representation of Attributes (this table show the preprocessed student related variables. These e variables are used in the research. This table shows the student related attributes with their detail description.)

Attribute	**Description(Domain)**
Gender	Student’s Gender(binary: female or male)

Previous Degree	Student’s Previous Degree(binary: Fsc or Bachelor)
Level of study	Student’s level of study in university (binary: BS or MS)
Current Semester	Current Semester (numeric : 1 to 8)
Your CGPA	Grades(numeric : 1 to 4)
Address	Student’s home address type(binary: Urban or Rural)
Family size	Family size(binary: <3 to 5 or > 5)
Order in Family	In family student’s Order (nominal: Older, Middle, Younger)
Parents Relationship Status	Parent’s co-habitation status(binary: living together or apart)
Occupation of Father	Father’s Occupation(nominal: Government servant, businessman or Private sector)
Occupation of Mother	Mother’s Occupation(binary: yes or no)
Guardian	Student’s Guardian(nominal: father, mother or other)
Travel time	Home to school travel time(numeric: < 45 minutes, 45 to 1.5 hours, > 1.5 hours)
Weekly study hours	Weekly study hours (numeric: 5 hours, 10 hours, >10 hours)
Extracurricular activities	Extracurricular activities(binary: yes or no)
Internet Facility	Internet Facility (binary: yes or no)
Go outing with friends	Go outing with friends(nominal: Twice a year, Monthly, weekly)
Educational Expense	Educational Expense (nominal: parents, guardian or self)
Health Status	Current health status(nominal: Good, average or Bad)
Encourage your teacher	Encourage your teacher(nominal: yes, no or only sometimes)
Marks increase or decrease during Covid-19	Marks increase or decrease during Covid-19(binary: yes or no)
Suitable System	Suitable System(binary: Annual or Semester)
Coursework provides sufficient Information	Coursework provides sufficient Information (nominal: yes, no or maybe)

Analytical skills	Analytical skills (nominal: yes, no or maybe)
Coursework market Oriented	Coursework market Oriented (nominal: yes, no or improved)

3.21 10 Fold Cross-Validation

Our research used the ten-fold cross validation technique to measure the capacity of a model to generalize to new data sets. Using this technique, our research dataset divides into ten equal sized subsets, keeping the original minority-majority class ratio. One subset is kept for validation.

3.22 Over-fitting

When a model learns the information and noise in the training data to the level where it affects the model's performance on new data, this is known as over-fitting. This means that the model picks up on noise or random fluctuations in the training data and learns them as ideas. The issue is that these concepts do not apply to fresh data, limiting the models' ability to generalize. Over-fitting is a term used in data science to describe when a statistical model fits its training data perfectly. When this happens, the algorithm is unable to perform accurately against unknown data, defeating the goal of the method. The ability to apply machine learning algorithms every day to make predictions and classify data is ultimately due to the generalization of a model to new data.

In our research, machine learning algorithms are developed, a sample dataset is used to train the model. However, if the model is trained for too long on sample data or is too complicated, it may begin to learn the dataset's "noise," or irrelevant information. The model gets "over-fitted" when it memorizes the noise and fits too closely to the training set, and it is unable to generalize adequately to new data. If a model is unable to generalize adequately to new data, it will be unable to accomplish the classification or prediction tasks for which it was designed.

3.23 Training Data

Tenfold cross-validation were performed in this research, which means the dataset was divided into ten equal-sized sections at random. Percentages and fold cross-validation were the two forms of data splitting used. For the percentages 10:90, the training data set is 90% of the total data, while the testing data set is 10% while the rest is utilized to train the model.

This procedure is performed ten times, each time with a different subset for training and validation. At the end of each iteration, the average results are calculated. These strategies are used to predict whether a student will pass or fail a course based on data sets. The predictions was based on the demographics data of the students as well as dynamic data from the previous semester. Expect the forecast accuracy to improve in subsequent semesters as more data becomes available. For example, prediction after the semester indicates that all available data from before this semester was used, which could include assessment and quiz results, which are part of the final score and thus improve accuracy. Early prediction accuracy is necessary so that timely solutions can be taken to encourage students.

3.24 Testing Data

The test set is a collection of data used to analyze the performance of the model using a performance metric. The testing data set is 10% while the rest is utilized to train the model. No observations from the training set must make the test set. Whereas test data is simply used to evaluate the performance of the model, testing data is the unknown data for which predictions must be produced. It will be difficult to tell if the algorithm has learned to generalize from the training set or has just memorized it.

3.25 Classification Algorithms used for Prediction

In machine learning, classification is an important factor. It falls under the category of supervised learning, i.e. while training the input variables we can map the output variables. When the data get trained it can easily distinguish the incoming data and will correctly classify it. Classification can be performed on both unstructured and structured data. The following are the main classification algorithms that are used in machine learning. To predict the performance of the student we have four classification algorithms which are as follows.

- **Naïve Bayes Classifier**

The Bayes Theorem is the base of the Naive Bayes Algorithm. It is a machine learning algorithm that works on probabilities. This algorithm works on a supposition that all attributes under observation are independent. However, the independent attribute supposition is usually not applicable in real-world scenarios. Despite the fact, this algorithm performs well and produce classified results with high accuracy.

It is used for solving classification issues. It is a supervised learning algorithm based on Bayes Theorem. Text classification issues are mostly solved by using the Naïve Bayes algorithm. It works based on probability. Naive Bayes algorithm are used to soled the text classification issues because it allows us to determine the conditional probabilities of occurrence of two events based on the probabilities of occurrence of each individual event. Encoding those probabilities is incredibly useful. Analyses of sentiments are usually implemented by using the naïve Bayes algorithm.

- **Random Forest Classifier**

Random Forest comes under supervised learning. Its work is based on ensemble learning. Multiple decision trees combine to form the random forest. The predictions of the decision trees form the outcome of the random forest. Missing data can be effectively handled by sing random forest classifier. It is not only a more efficient algorithm than decision trees but also addresses the problem of over-fitting in decision trees.

- **J48**

For generating decision trees J48 algorithm is used. It is an extended version of ID3 algorithms. For classification purposes, decision trees generated from J48 can be used. J48 is a statistical classifier.

- **Support Vector Machine**

Support Vector Machine is a popular and widely used algorithm that comes under the banner of supervised learning. The main concept of the SVM algorithm is that it distinctly classifies the data points with the help of the hyperplane. It performs well when the number of input features is less than or equal to 3. It creates a hyperplane to distinguish the data points.

The points falling on either side of the boundary can be attributed to different classes.

Support Vector Machine is a supervised learning algorithm. This algorithm is used to perform analysis on data for classification purposes. Support vector machine creates a hyperplane that separates n-dimensional space into classes so that one can easily insert new data points in the right category.

- **Decision Tree**

A Decision Tree is a supervised learning algorithm i.e. the target value is already known. The main edge of using Decision Trees is that they can be used for solving problems relating to classification and regression. Decision Trees can be used with both numerical i.e. height, width and categorical i.e. grades, location data.

- **Neural Network**

Neural Network as its name interprets is a set of algorithms that copy the behaviour in which the human brain thinks and operate. While practicing a neural network, does not need to change the output variables with the variation of input variables as neural networks are adaptive. This algorithm has a wide range of implementation in risk analysis, detection of fraudulent activities and stock market.

- **K- Nearest Neighbor**

This algorithm can be used for both regression and classification. K-NN is another classification algorithm that comes under supervised learning. This algorithm works based on similarity i.e. it assumes that similar items would be near to each other. To predict with high accuracy this algorithm depends upon the value of K. To select the right value for K, this algorithm runs several times by replacing the value of k. The value on which the errors are minimum is being selected as the value of k.

3.26 Tabular Representation of Existing Dataset

In this table, show the student related variables/features of the existing dataset with their brief description. A tabular representation of the existing dataset is shown in Table 2.

Table 1.2: Tabular Representation of Attributes of Existing dataset (this table shoe the student related variables/features of the existing dataset with their description)

Attribute	**Description(Domain)**
Gender	Student’s Gender(binary: female or male)
Age	student’s age (numeric: from 15 to 22)
School	student’s school (binary: Gabriel Pereira or Mousinho da Silveira)
Father Education	Father’s education (numeric: from 0 to 4[a])
Mother Education	mother’s education (numeric: from 0 to 4[a])
Address	Student’s home address type (binary: urban or rural)
Family size	Family size (binary: ≤ 3 or > 3)
Quality of Family relationships	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
Parents Relationship Status	parent’s cohabitation status (binary: living together or apart)
Father’s Job	father’s job (nominal[b])
Mother’s Job	mother’s job (nominal[b])
Guardian	Student’s guardian (nominal: mother, father or other)
Travel time	Home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
Weekly study hours	Weekly study hours (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
Extracurricular activities	Extracurricular activities(binary: yes or no)
Internet Facility	Internet access to the home (binary: yes or no)
Go outing with friends	Go outing with friends(numeric: from 1- very low to 5- very high)
Extra educational support	Extra educational school support (binary: yes or no
Health Status	Current health status(numeric: from 1- very bad to 5- very good)
reason	Reason for choosing this school(nominal: close to home, school reputation, course preference or others)
Failures	Number of past class failures (numeric: n if 1 <= n < 3, else 4)

Extra paid class	Paid extra class (binary: yes or no)
Nursery	Attended nursery school (binary: yes or no)
Higher	Wants to take higher education (binary: yes or no)
Romantic	With a romantic relationship (binary: yes or no)
Free time	Free time after school(numeric: from 1- very low to 5- very high)
Weekend alcohol consumption	Weekend alcohol consumption(numeric: from 1- very low to 5- very high)
Workday alcohol consumption	Workday alcohol consumption(numeric: from 1- very low to 5- very high)
absences	Number of school absences (numeric: from 0 to 93)
G1	first grade period(numeric: from 0 to 20)
G2	Second grade period (numeric: from 0 to 20)
G3	Final grade(numeric: from 0 to 20)

The existing dataset contains information on 369 students with 33 different features. The study's output variable is previous grades (G3). The existing dataset includes information such as gender, father and mother's occupations, and so on. The classification accuracy rate of three common data mining techniques (decision tree, random forest, and Naive Bayes) was analyzed. The effects of two distinct grade categorizations on data mining are also investigated in this study: five-level grade categorization and binary grade categorization. Two datasets are available in this study the name of the datasets are mathematics and Portugal.

For the Portugal dataset, the random forest algorithm produced the best results for the five-level grading version of this dataset, with an accuracy rate of 73.50 %. However, with the binary grading version of this dataset, the accuracy rate was improved. The accuracy rate was improved to 93.07 % in the dataset where the final grade is categorized in binary form (passing or failing). For the mathematics dataset, the decision tree (J48) algorithm produced the best results for the five-level grading version, with an accuracy rate of 73.42 %. The random forest algorithm produced the best accuracy rate of 91.39 % for the binary dataset version.

3.27 Combine Features

The feature from the existing dataset are based on the demographic, social, student socioeconomic and course related features. Common features of this research dataset are shown in Table 3. The results reproduce based on common features. The major reason for selecting these features is that they are already being used in four to five existing studies. Another reason for selecting these features are that this research is used to make the comparison with the existing studies.

Table 3.3: Combine features (this table shows the features which are picked up from the existing dataset and the common features of this research dataset. The results reproduced based on common features. The major reason for selecting the features is that they are already being used in 4 to 5 existing studies)

Attribute	**Description(Domain)**
Gender	Student's Gender(binary : female or male)
Address	Student's home address type (binary: urban or rural)
Family size	Family size (binary: ≤ 3 or > 3)
Parents Relationship Status	parent's cohabitation status (binary: living together or apart)
Father's Job	father's job (nominal[b])
Mother's Job	mother's job (nominal[b])
Guardian	Student's guardian (nominal: mother, father or other)
Travel time	Home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
Weekly study hours	Weekly study hours (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
Extracurricular activities	Extracurricular activities(binary: yes or no)
Internet Facility	Internet access to the home (binary: yes or no)

Go outing with friends	Go outing with friends(numeric: from 1- very low to 5- very high)
Extra educational support	Extra educational school support (binary: yes or no
Health Status	Current health status(numeric: from 1- very bad to 5- very good)

In the existing study, we have included some features and some are excluded because of some reason. The exclusion/inclusion criteria was based on the geographical location and social culture. Some of the features were not according to our society and some features help to meet the goal of study. The exclusion/inclusion criteria was based on the feature that help to meet the goal of the study. The features which are excluded in proposed study dataset are “Alcohol Consumption weekly”, “Alcohol Consumption daily”, and “Romantic Relationship” and “Nursery” because these two features (Alcohol Consumption and Romantic Relationship) are prohibited in our Islamic culture. And the feature” Nursery” excluded in proposed study dataset because we collect the data from the university level students and at university level don’t need the Nursery class grades.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Overview

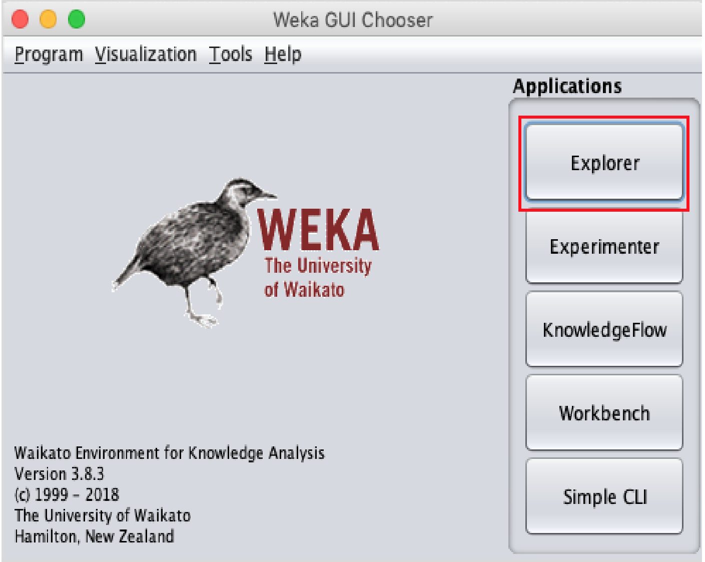
This chapter will provide detailed view of the outcomes obtained, the data used, and the experiment process used to answer the research questions. After analysis performed on research dataset, have to reproduce the results. A brief description and graphical representation are given below.

4.2 WEKA Tool (Waikato Environment for Knowledge Analysis)

WEKA is open-source software. WEKA is a set of machine learning algorithms designed specifically for data mining jobs. Data preprocessing, categorization, regression, visualization, and learning algorithms. For using prediction algorithms, the great majority of researchers use WEKA. WEKA is a simple piece of software that allows us to implement a variety of algorithms, but it does not support large datasets that require programming. WEKA (Waikato Environment for Knowledge Analysis) software, developed at the University of Waikato in New Zealand, was chosen for this investigation. It is portable and platforms-independent because it is written entirely in Java and so runs on practically any modern computing platform. It is presently applied in a wide range of applications, including education and research.

4.3 WEKA Tool Interface

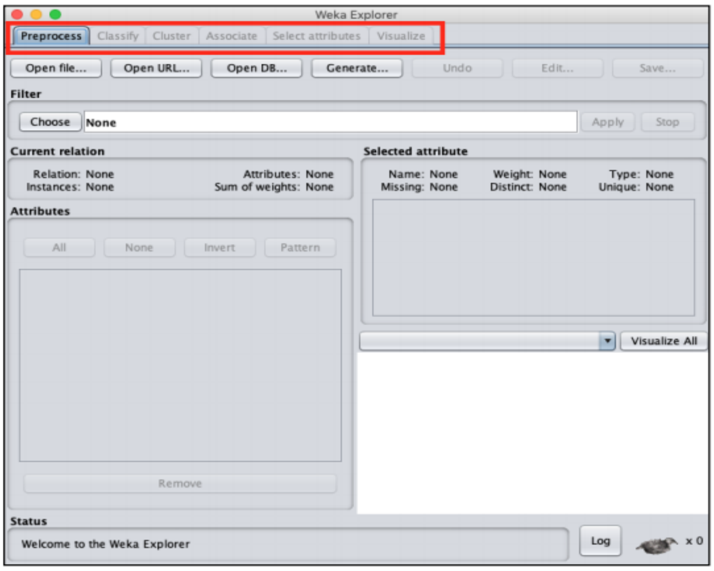
The WEKA GUI Chooser application will open, we used Explorer for the research. The following screen will appear:



Waikato Environment for Knowledge Analysis
Version 3.8.3
(c) 1999 - 2018
The University of Waikato
Hamilton, New Zealand

Figure 4.1: Weka GUI Chooser (choose the appropriate application from the different applications of WEKA tool. Select the Explorer application for the research)

The explorer does have a variety of functionalities for working with large amounts of data. The following screen appears when you click the Explorer button.



Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

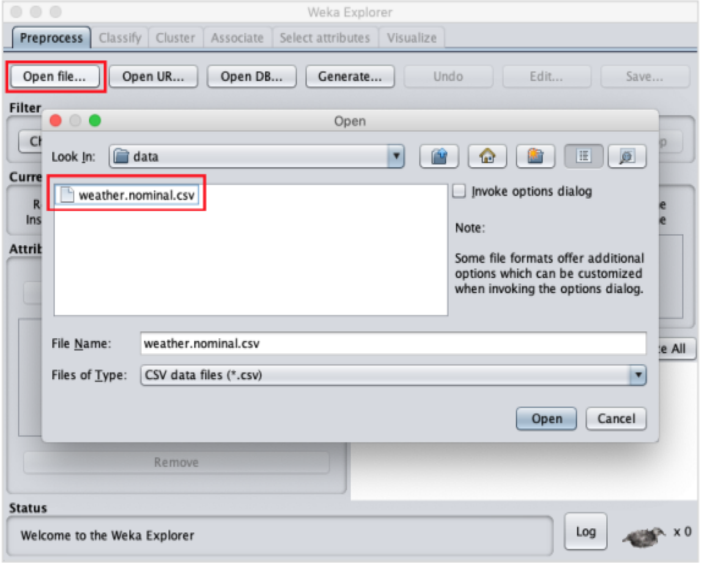
Figure 4.2: Weka Explorer (When choose the Explorer application from the Weka GUI chooser, Weka explorer shows the different tabs. The Explorer does have a variety of functionalities for working with large amount of data.)

Only the Preprocess tab is enabled when you first start the explorer. Preprocessing the data is the initial stage in machine learning. Thus, you will select the data file, process it, and prepare it for use with the various machine learning algorithms in the Preprocess option.

- **Loading Data**

You'll first learn how to open the data file in the WEKA Explorer. The local file system can be used to load the data:

Select the Open file option. As shown on the screen, a directory navigator window appears:



Look In: data

weather.nominal.csv

□ Invoke options dialog

Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: weather.nominal.csv

Files of Type: CSV data files (*.csv)

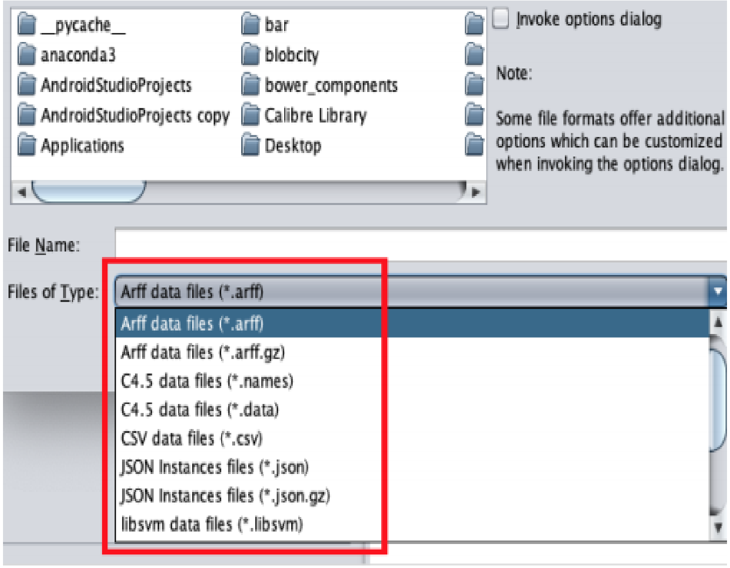
Status

Welcome to the Weka Explorer

Figure 4.3: Loading data file (After open the Weka explorer then need to load the file in Weka. click the "Open file" button to load the local file which is saved in the computer.)

- **File Format**

WEKA can read and write data in many different file formats. Here is the complete list.



□ Invoke options dialog

Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name:

Files of Type: Arff data files (*.arff)

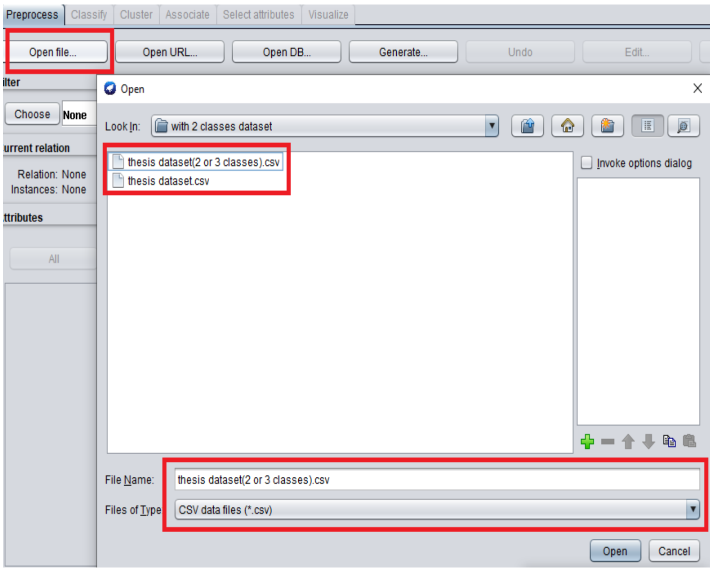
Arff data files (*.arff)
Arff data files (*.arff.gz)
C4.5 data files (*.names)
C4.5 data files (*.data)
CSV data files (*.csv)
JSON Instances files (*.json)
JSON Instances files (*.json.gz)
libsvm data files (*.libsvm)

Figure 4.4: File Format (Weka can read or write data in many different file formats)

- **Preprocessing data**

We'll utilize the Final dataset database included in the installation to demonstrating the possible features in preprocessing.

Select the final dataset. CSV file using the Preprocess tag's Open file option.



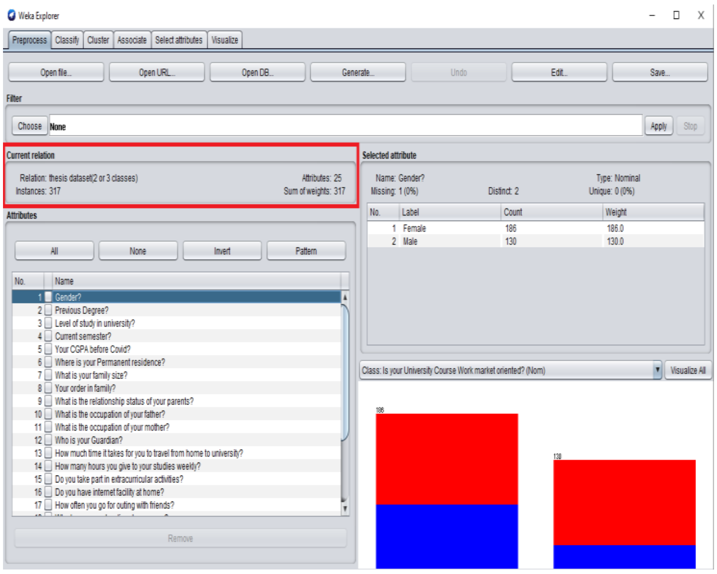
thesis dataset(2 or 3 classes).csv
thesis dataset.csv

File Name: thesis dataset(2 or 3 classes).csv

Files of Type: CSV data files (*.csv)

Figure 4.5: Preprocessing data (To preprocess the data, need to open the database file which is saved in the computer. Select the "thesis dataset (2 or 3 classes)" file in the database with "CSV" file format).

When you open the file, your screen should appear like this:



Filter: None

Current relation

Relation: thesis dataset(2 or 3 classes) Attributes: 25
Instances: 317 Sum of weights: 317

Selected attribute

Name: Gender? Type: Nominal
Missing: 1 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	Female	186	186.0
2	Male	130	130.0

Attributes

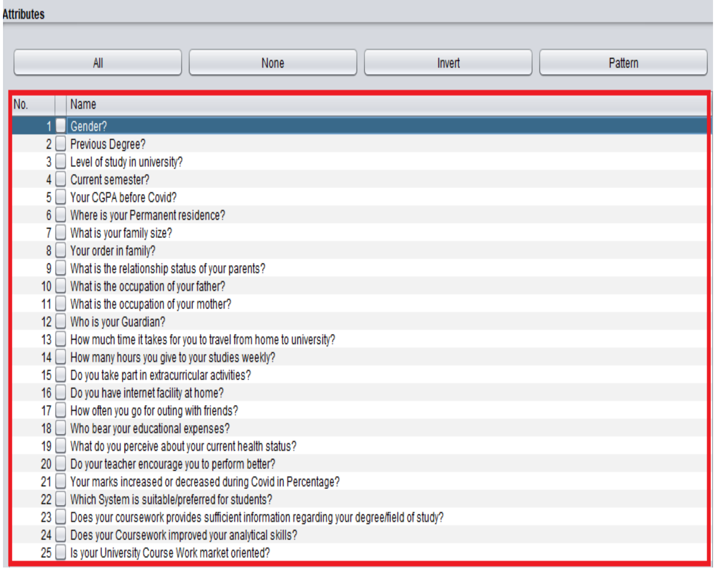
No.	Name
1	Gender?
2	Previous Degree?
3	Level of study in university?
4	Current semester?
5	Your CGPA before Covid?
6	Where is your Permanent residence?
7	What is your family size?
8	Your order in family?
9	What is the relationship status of your parents?
10	What is the occupation of your father?
11	What is the occupation of your mother?
12	Who is your Guardian?
13	How much time it takes for you to travel from home to university?
14	How many hours you give to your studies weekly?
15	Do you take part in extracurricular activities?
16	Do you have internet facility at home?
17	How often you go for outing with friends?

Class: Is your University Course Work market oriented? (Nom)

Figure 4.6: Attribute (the data file is loaded in the Weka, the following screen appears. This screen shows the information about the data file like attributes, instances, name of the data file and sum of weights)

- **Data Description**

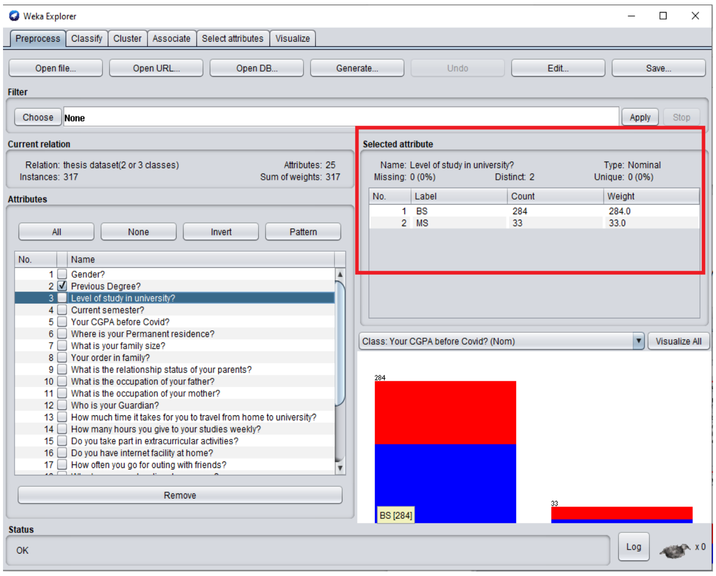
Let's start with the noted Current related sub-window. It displays the name of the currently loaded database. This sub-window teaches us two things: There are 317 occurrences, which correspond to the number of rows in the table. The table has twenty-five attributes - the fields that will be explained in the next sections.



No.		Name
1	□	Gender?
2	□	Previous Degree?
3	□	Level of study in university?
4	□	Current semester?
5	□	Your CGPA before Covid?
6	□	Where is your Permanent residence?
7	□	What is your family size?
8	□	Your order in family?
9	□	What is the relationship status of your parents?
10	□	What is the occupation of your father?
11	□	What is the occupation of your mother?
12	□	Who is your Guardian?
13	□	How much time it takes for you to travel from home to university?
14	□	How many hours you give to your studies weekly?
15	□	Do you take part in extracurricular activities?
16	□	Do you have internet facility at home?
17	□	How often you go for outing with friends?
18	□	Who bear your educational expenses?
19	□	What do you perceive about your current health status?
20	□	Do your teacher encourage you to perform better?
21	□	Your marks increased or decreased during Covid in Percentage?
22	□	Which System is suitable/preferred for students?
23	□	Does your coursework provides sufficient information regarding your degree/field of study?
24	□	Does your Coursework improved your analytical skills?
25	□	Is your University Course Work market oriented?

Figure 4.7: Data Description (This window shows the attributes of the data file which is loaded on the Weka)

Take note of the Attributes sub-window on the left, which displays the database's different fields.



Filter: None

Relation: thesis dataset(2 or 3 classes) Attributes: 25
Instances: 317 Sum of weights: 317

No.		Name
1	□	Gender?
2	☑	Previous Degree?
3	□	Level of study in university?
4	□	Current semester?
5	□	Your CGPA before Covid?
6	□	Where is your Permanent residence?
7	□	What is your family size?
8	□	Your order in family?
9	□	What is the relationship status of your parents?
10	□	What is the occupation of your father?
11	□	What is the occupation of your mother?
12	□	Who is your Guardian?
13	□	How much time it takes for you to travel from home to university?
14	□	How many hours you give to your studies weekly?
15	□	Do you take part in extracurricular activities?
16	□	Do you have internet facility at home?
17	□	How often you go for outing with friends?

Name: Level of study in university? Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	BS	284	284.0
2	MS	33	33.0

Class: Your CGPA before Covid? (Nom)

Status: OK

Figure 4.8: Selecting Attributes (The highlighted section of this window shows the information about the specific or individual feature/attribute of the dataset)

There are twenty-five separate fields in the thesis dataset database. When you click on an attribute in this list, further information about that attribute is displayed on the right side.

- **Visualization of data**

When you click the Visualize All button, you will be able to see all of the characteristics in a single window, as illustrated here:

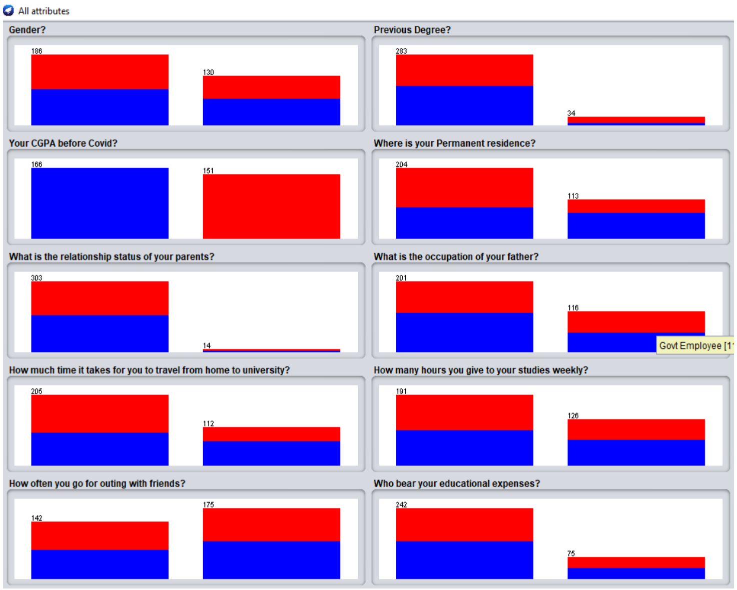

All attributes
Gender?
186
130
Previous Degree?
283
34
Your CGPA before Covid?
166
151
Where is your Permanent residence?
204
113
What is the relationship status of your parents?
303
14
What is the occupation of your father?
201
116
Govt Employee [1
How much time it takes for you to travel from home to university?
205
112
How many hours you give to your studies weekly?
191
126
How often you go for outing with friends?
142
175
Who bear your educational expenses?
242
75

Figure 4.9: Visualize the data (This window shows the visual representation of attributes with the target class)

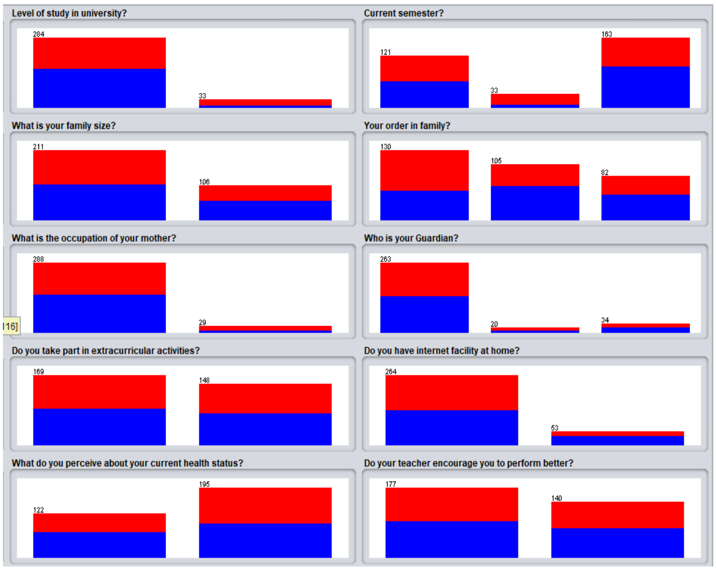
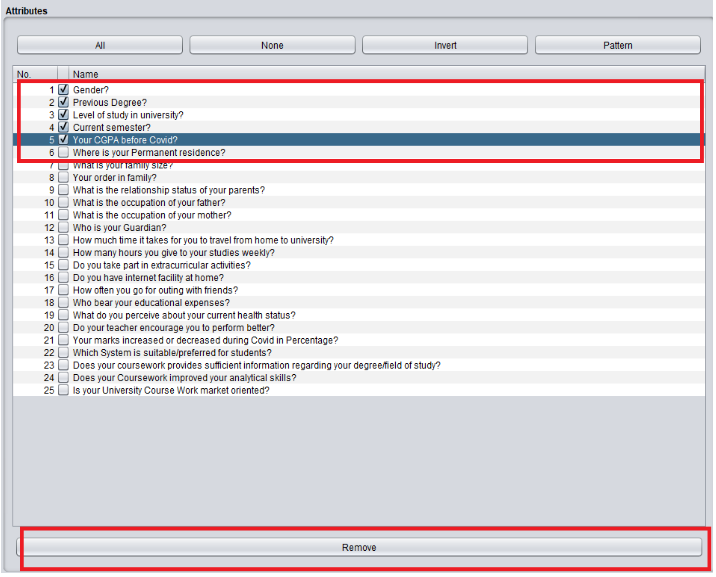

Level of study in university?
284
33
Current semester?
121
33
163
What is your family size?
211
106
Your order in family?
130
105
82
What is the occupation of your mother?
288
29
16]
Who is your Guardian?
263
20
34
Do you take part in extracurricular activities?
169
148
Do you have internet facility at home?
264
53
What do you perceive about your current health status?
122
195
Do your teacher encourage you to perform better?
177
140

Figure 4.10: Visualize the data (This window shows the visual representation of attributes with the target class)

- **Feature Selection**

Many times, the data you wish to use for model creation includes a lot of unnecessary fields. Select the Attribute/s you want to delete and then click the Remove button at the bottom. The database would be cleansed of the selected properties. You can save the data for model development after it has been thoroughly preprocessed. When you removed the attributes the following window appears:

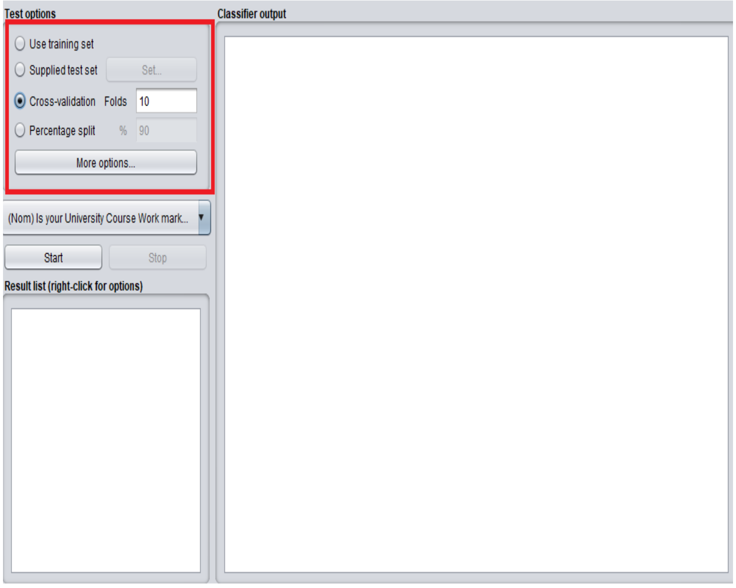


No.		Name
1	☑	Gender?
2	☑	Previous Degree?
3	☑	Level of study in university?
4	☑	Current semester?
5	☑	Your CGPA before Covid?
6	☐	Where is your Permanent residence?
7	☐	What is your family size?
8	☐	Your order in family?
9	☐	What is the relationship status of your parents?
10	☐	What is the occupation of your father?
11	☐	What is the occupation of your mother?
12	☐	Who is your Guardian?
13	☐	How much time it takes for you to travel from home to university?
14	☐	How many hours you give to your studies weekly?
15	☐	Do you take part in extracurricular activities?
16	☐	Do you have internet facility at home?
17	☐	How often you go for outing with friends?
18	☐	Who bear your educational expenses?
19	☐	What do you perceive about your current health status?
20	☐	Do your teacher encourage you to perform better?
21	☐	Your marks increased or decreased during Covid in Percentage?
22	☐	Which System is suitable/preferred for students?
23	☐	Does your coursework provides sufficient information regarding your degree/field of study?
24	☐	Does your Coursework improved your analytical skills?
25	☐	Is your University Course Work market oriented?

Figure 4.11: Feature Selection (If you want to remove any attributes in the list, select the attributes that want to remove and then click the remove button the selected attributes are removed in the list.)

- **Splitting of data**

Open the saved file by using the Open file under the Preprocess tab, click on the Classify tab, and the following screen appears. You'd use cross-validation or percentage split choices unless you had your own training set or a client-supplied test set. You can set the number of folds in which the complete data is split and used during each iteration of training using cross-validation. The data will be split between training and testing using the set split percentage in the percentage split.



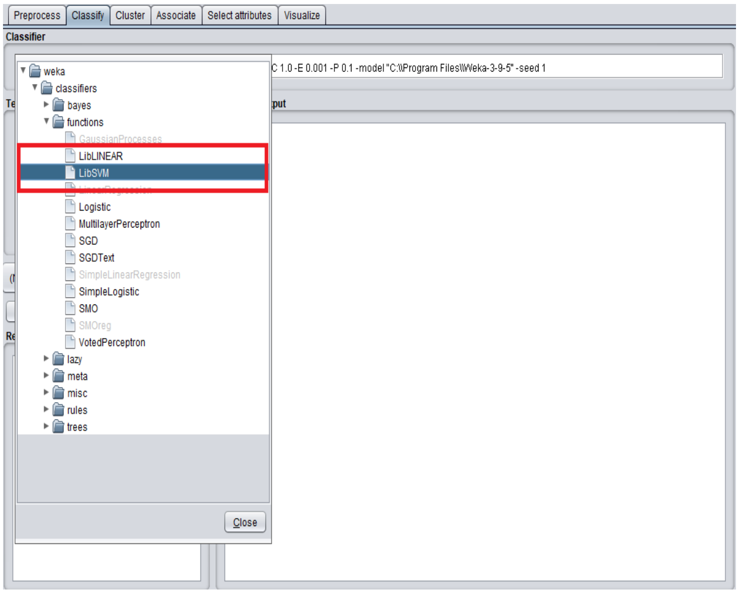
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 90

(Nom) Is your University Course Work mark...

Figure 4.12: Splitting of data (this figure shows the test option which are used in the testing the data file)

- **Selecting Classifier**

Click on the Choose button and select the following classifier:



C 1.0 -E 0.001 -P 0.1 -model "C:\\Program Files\\Weka-3-9-5" -seed 1

Figure 4.13: Selecting classifier (this figure shows that where and how to select the classifier to find the accuracy. Weka provides the different classifiers to find the accuracy. Some of the classifiers are J48, Naïve bayes, random forest Zero etc.)

4.4 Results

- **Demographic Features**

Demographic data refers to information on groups of students based on particular characteristics

such as:

- Gender
- Permanent Residence
- Current health status

One of the features in our dataset is “gender” which belongs to two categories Male and Female. The female ratio came around according to our chart is 186 and according to data received Male ratio in our chart is 130. The Accuracy of this feature against the Target class “CGPA” is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine.

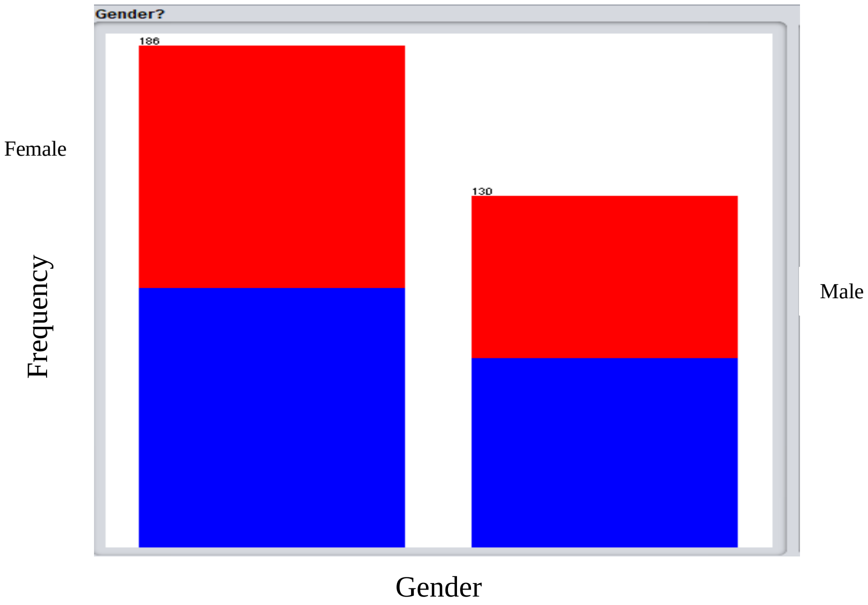


Figure 4.14: Bar chart of Gender (This bar chart shows the feature "gender" which belongs to two categories Male and Female. This chart show the Female ratio that is 186 and the Male ratio is 130.)

The next feature of our dataset is "Permanent Residence" which belongs to two categories Urban and Rural. The urban ratio came according to our chart is 204 and the data received the Rural, the ratio is 113 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 46.88%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students live in big cities.

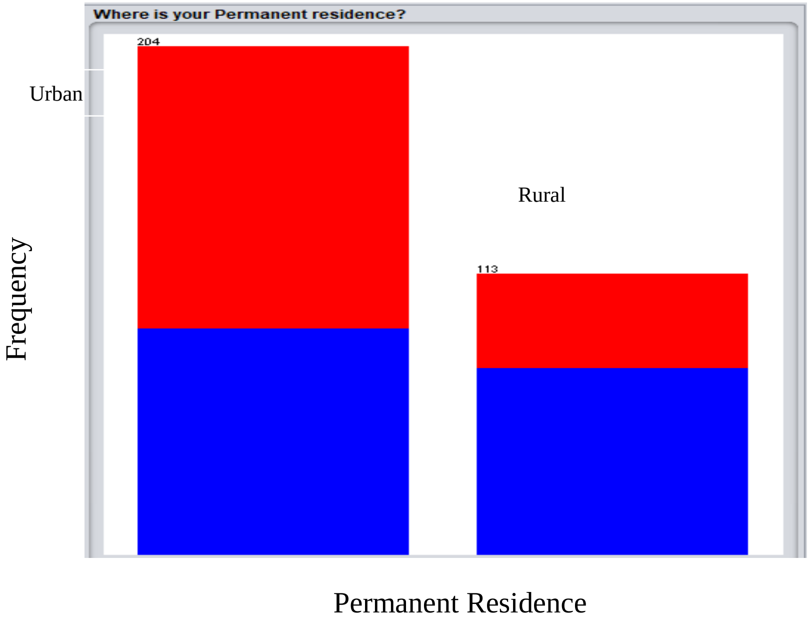


Figure 4.15: Bar chart of Permanent Residence This bar chart shows the feature "Permanent residence" which belongs to two categories Urban and Rural. This chart show the urban ratio that is 204 and the rural ratio is 113.)

The next feature of our dataset is "Current Health Status" which belongs to two categories good and bad. The good ratio came according to our chart is 195 and the data received for the

Bad ratio is 122 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of the students are in good health.

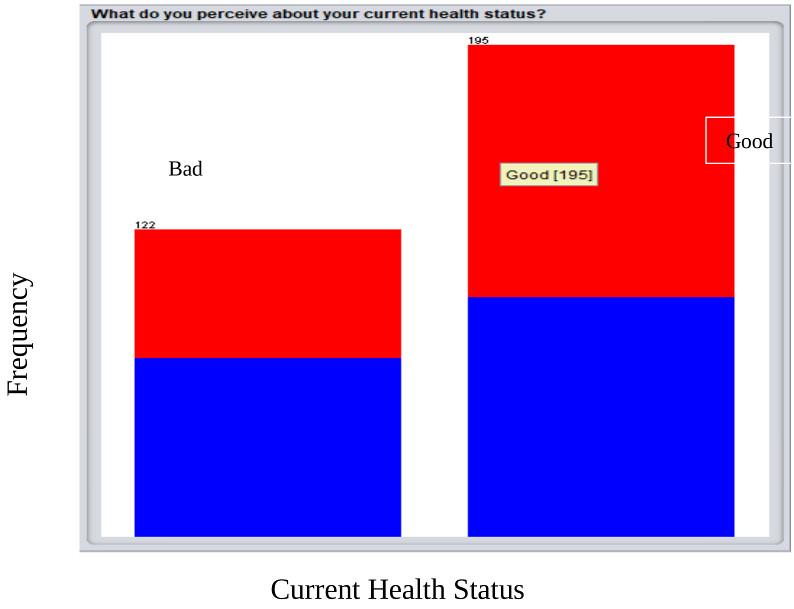


Figure 4.16: Bar chart of Current Health Status This bar chart shows the feature "current health status" which belongs to two categories Good and Bad. This chart show the Good ratio that is 195 and the Bad ratio is 122.)

- **Learning Environment Feature**

Learning environment factor data refers to learning environment of the students where they learn from and the features are

- Previous degree
- Level of study
- Current semester
- Marks increase or decrease during COVID-19
- Suitable system
- Encouraged by teacher
- Weekly study time
- Travel time from home to the university

The next feature of our dataset is "previous degree" which belongs to two categories FSc and BS. FSc ratio came according to our chart is 283 and the BS ratio is 84 according to our chart. It means the 284 students hold the FSc degree as the previous degree and 84 students holds the BS degree as the previous degree. The Accuracy of this feature against the target class

“CGPA” is 68.75%. This accuracy is reported with the best classifier that is the Support vector machine.

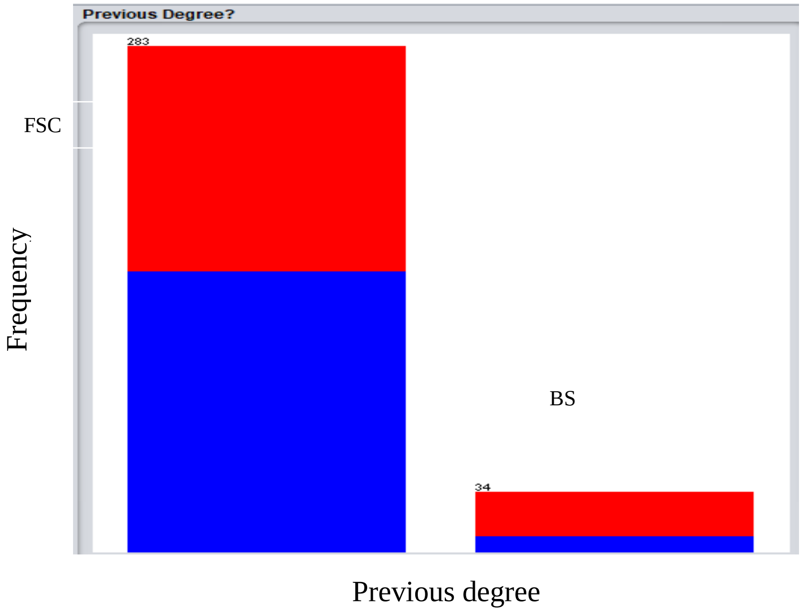


Figure 4.17: Bar chart of the Previous Degree (This bar chart shows the feature “previous degree” which belongs to two categories FSC and BS. This chart show the FSC ratio that is 283 and the male ratio is 84.)

The next feature of our dataset is “Level of Study” which belongs to two categories BS and MS. BS ratio came according to our chart is 284 and the data received the MS ratio is 33 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 68.75%. This accuracy is reported with the best classifier that is the Support vector machine. This graph illustrates that the majority of students are currently seeking a BS degree at a university.

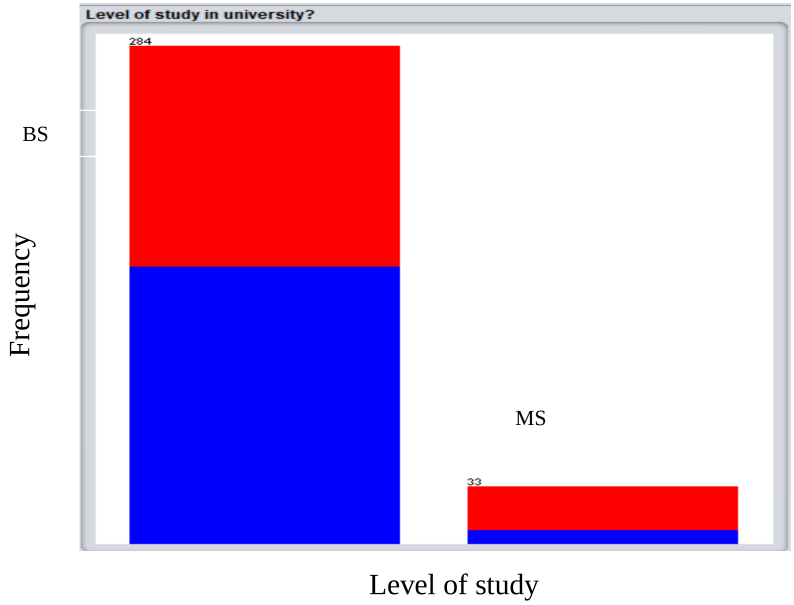


Figure 4.18: Bar chart of Level of study (This bar chart shows the feature “level of study” which belongs to two categories BS and MS. This chart show the BS ratio that is 284 and the MS ratio is 33.)

The next feature of our dataset is “Current Semester” which belongs to three categories

A, B and C. A contains the students from semester 1 to 4, B contains the students from 5 semesters to onward more specific to 5 to 8 and C will take the students from the Master’s degree because the master’s degree consists of 4 semesters if we don’t mention the category “C” we would fall in category A to make it more specific we have to make definite category C. A ratio came according to our chart is 121, the data received the B, the ratio is 163 and the data received the C, the ratio is 33 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 68.75%. This accuracy is reported with the best classifier that is the Support vector machine. This graph illustrates that the majority of students belongs to category A that means the majority of students are currently enrolled in one of four semesters.

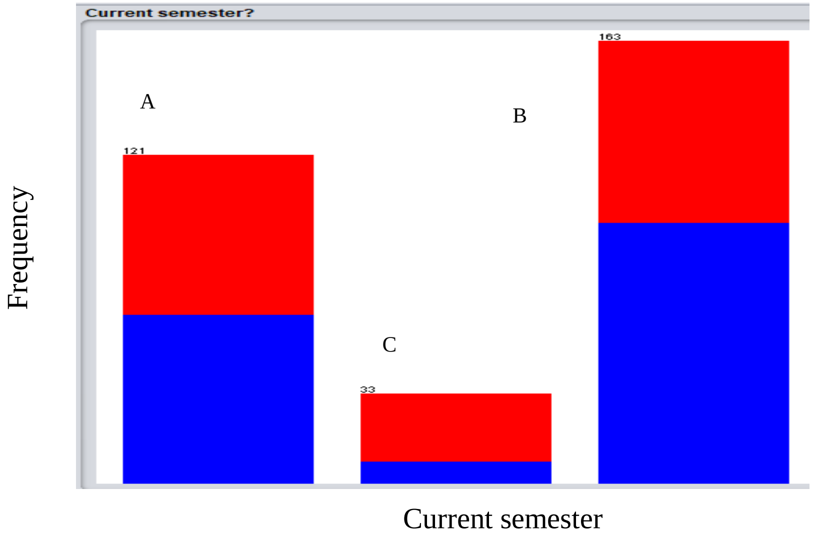


Figure 4.19: Bar chart of Current Semester (This bar chart shows the feature “Current semester” which belongs to three categories A, B and C. A contains the students from semesterv1 to 4 ,B contains the students from semester 5 to 8 and C contains the students from MS degree This chart show the A ratio that is 121, B ratio that is 163 and the C ratio is 33.)

The next feature of our dataset is “Marks Increase or Decrease during COVID-19” which belongs to two categories Yes and No. Yes ratio came according to our chart is 244 and the data received for the No ratio is 73 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 56.25%. This accuracy is reported with the best classifier that is the Support vector machine. This chart illustrates that during the COVID-19 pandemic, the majority of students achieve better grades while only a few students received lower grades.

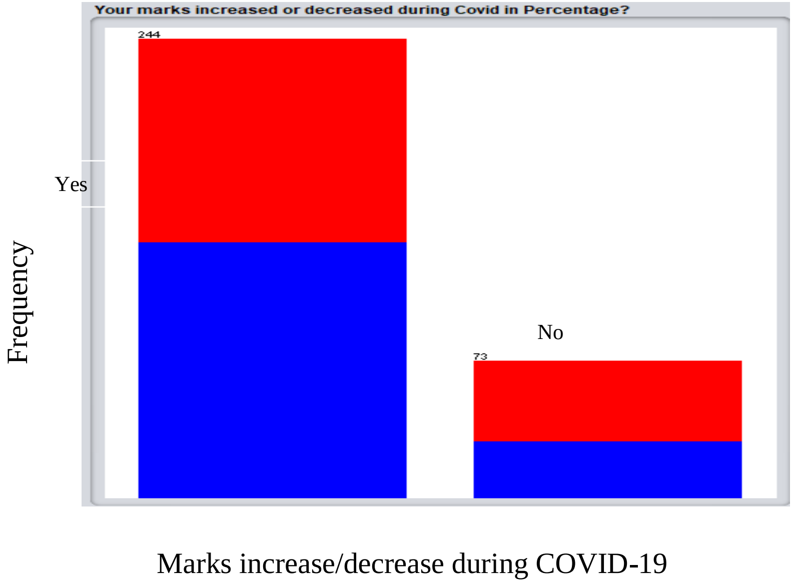


Figure 4.20: Bar chart of Marks Increase/Decrease during COVID-19 (This bar chart shows the feature "Marks increase/decrease during COVID-19" which belongs to two categories Yes and No. This chart show the Yes ratio that is 244 and the No ratio is 73.)

The next feature of our dataset is the "Suitable system" which belongs to two categories annual system and the semester system. The annual system ratio came according to our chart is 30 and the data received for the Semester System ratio is 287 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that most students prefer the semester system to the annual system.

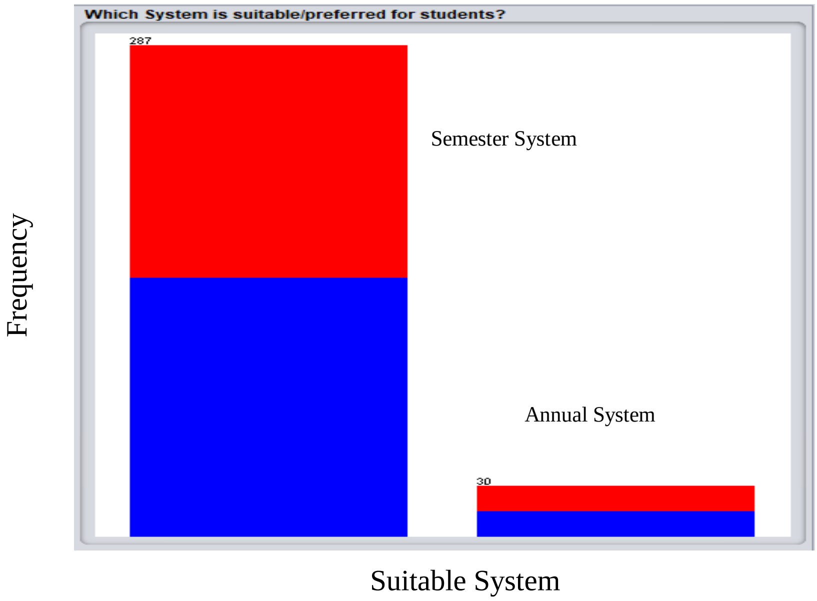


Figure 4.21: Bar chart of Suitable System (This bar chart shows the feature "Suitable system" which belongs to two categories Annual system and Semester system. This chart show the Annual system ratio that is 30 and the Semester system ratio is 287.)

The next feature of our dataset is "Encourage by Teacher" which belongs to two categories Yes and No. Yes ratio came according to our chart is 177 and the data received for "CGPA" is

62.50%. This accuracy is reported with the best classifier that is the Support vector machinemachine. This graph illustrates that the majority of teachers encourage students to achieve higher performance.

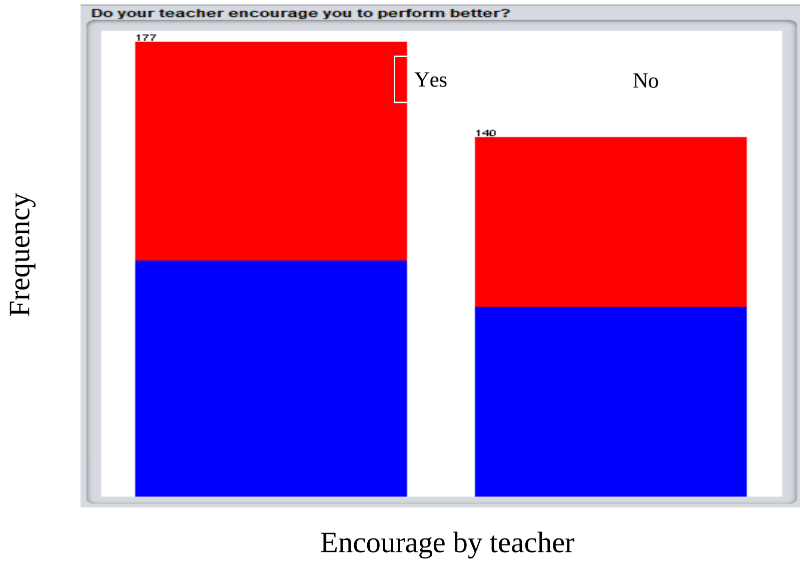


Figure 4.22: Bar chart of Encourage by Teacher (This bar chart shows the feature "encourage by your teacher" which belongs to two categories Yes and No. This chart show the Yes ratio that is 177 and the No ratio is 140.)

The next feature of our dataset is "Weekly time given to the Studies" which belongs to three categories 5 hours, 10 hours and 15 hours. Five hours ratio came according to our chart is 121, 10 hours ratio is 126 and the data received for 15 hours ratio is 70 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students dedicate 5 to 10 hours each week to their studies.

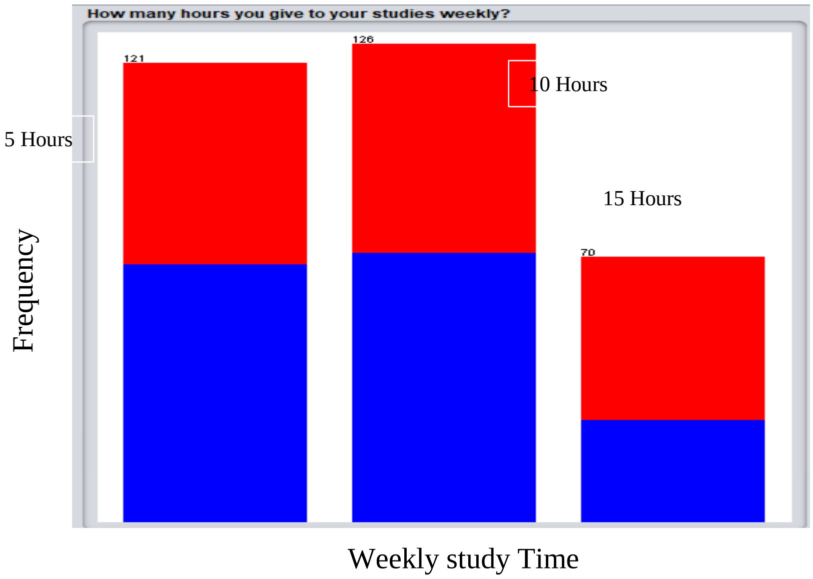


Figure 4.23: Weekly Study Time (This bar chart shows the feature "Weekly study time" which belongs to three categories 5 hours, 10 hours s and 15 hours. This chart show the 5 hours ratio that is 121, 10 hours ratio is 126 and the 15 hours ratio is 70.)

The next feature of our dataset is “Travel Time from home to University” which belongs to three categories less than 45 minutes, more than 1.5 hours and 45 minutes to 1.5 hours. Less than 45 mints ratio came according to our chart is 109, for more than 1.5 hours ratio is 112, and for 45 mins to 1.5 hours, the ratio is 96 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 40.63%. This accuracy is reported with the best classifier that is the Support vector machine. This graph indicates that the majority of students had to come to university for more than 1.5 hours from home.

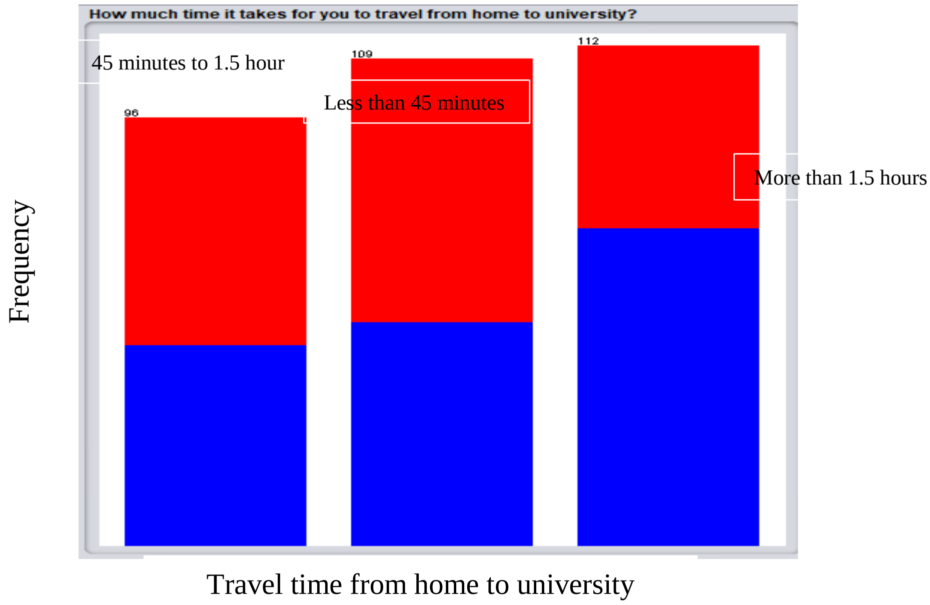


Figure 4.24: Bar chart of Travel time from University to Time (This bar chart shows the feature “travel time from home to university” which belongs to three categories less than 45 minutes, more than 1.5 hours and 45 minutes to 1.5 hour. This chart show the less than 45 minutes ratio that is 109, more than 1.5 hours ratio is 112 and the 45 minutes to 1.5 hours ratio is 96.)

- **Student’s Socioeconomics Features**

Socioeconomic factor data refers to information of students based on particular characteristics such as

- Family size
- Order in family
- Parents relationship status
- Occupation of father
- Mother Occupation
- Guardian

- Education expense support
- Internet facility

The next feature of our dataset is “Family Size” which belongs to two categories Less than 3 to 5 and Greater than 5. Less than 3 to 5 ratio came according to our chart is 211 and the data received the Greater than 5 ratio is 106 according to our chart. The Accuracy of this feature against the target class “CGPA” is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students belongs to a small family.

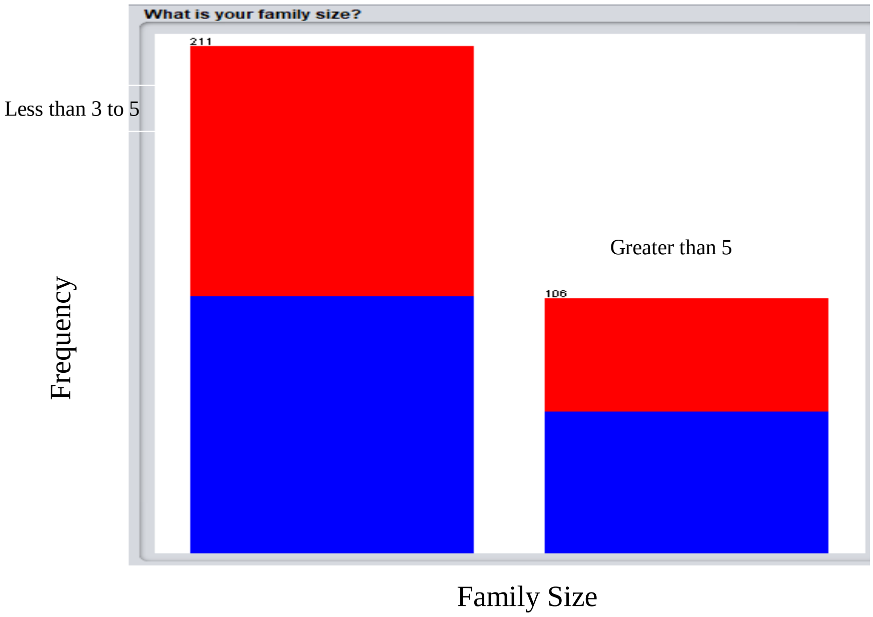


Figure 4.25: Bar chart of Family Size (This bar chart shows the feature “family size” which belongs to two categories less than 3 to 5 and greater than 5. This chart show the less than 3 to 5 ratio that is 211 and the greater than 5 ratio is 106.)

The next feature of our dataset is “Order in Family” which belongs to three categories Elder, Middle and Younger. The elder ratio came according to our chart is 105, the data received the Middle, the ratio is 130 and the data received the Younger, the ratio is 82 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 56.25%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students are middle in their family.

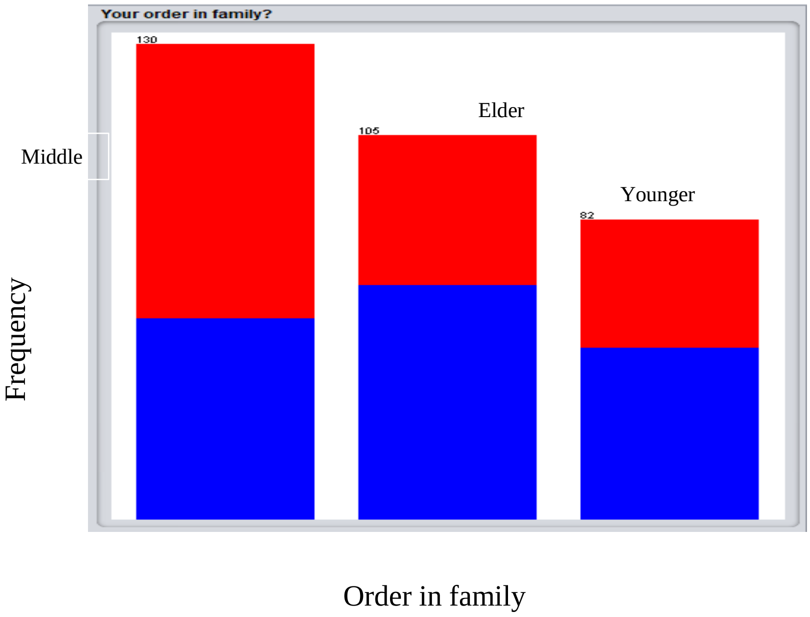


Figure 4.26: Bar chart of Order in Family Time (This bar chart shows the feature "order in family" which belongs to three categories elder, middle and younger. This chart show the elder ratio that is 105, middle ratio is 130 and the younger ratio is 82.)

The next feature of our dataset is "Parents Relationship Status" which belongs to two categories living together or Apart/Separate. The living together ratio came according to our chart is 303 and the data received the Apart/Separate ratio is 14 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students' parents happily live together.

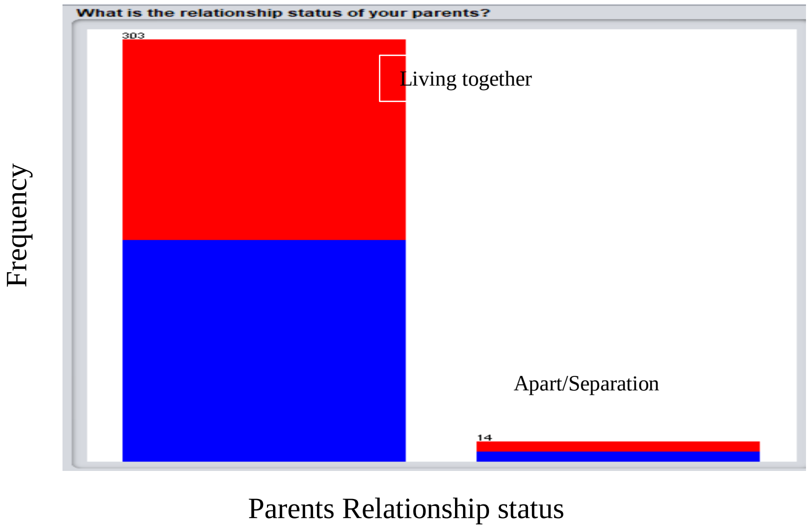


Figure 4.27: Bar chart of Parents relationship status (This bar chart shows the feature "parent's relationship status" which belongs to two categories living together and apart. This chart show the living together ratio that is 303 and the apart ratio is 14.)

The next feature of our dataset is "Occupation of Father" which belongs to three categories government servant, businessman and doing job in the private sector. The Government servant ratio came according to our chart is 116, for the businessman ratio is 67 and for the private sector, the ratio is 134 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 40.63%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students' parents work for the government, a few open a business, as well as a few works in the private sector.

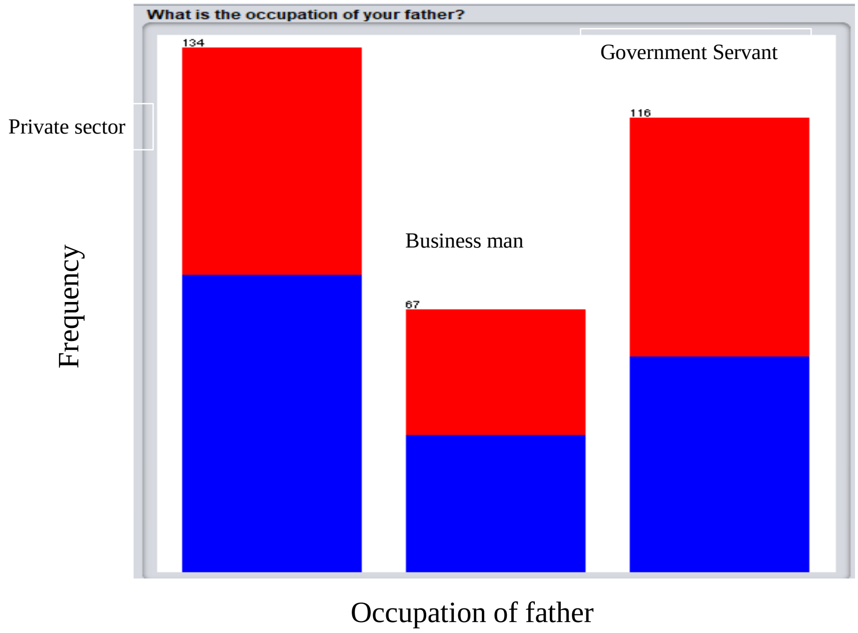


Figure 4.28: Bar chart of Occupation of Father Time (This bar chart shows the feature "occupation of father" which belongs to three categories government servant, businessman and private sector. This chart shows the government servant ratio that is 116, business man ratio is 67 and the work in a private sector ratio is 134.)

The next feature of our dataset is "Occupation of Mother" which belongs to two categories holding a job or not holding the job. For not a job holder, the ratio came according to our chart is 288 and for holding a job, the ratio is 29 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 65.62%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of student mothers are housewives rather than working women.

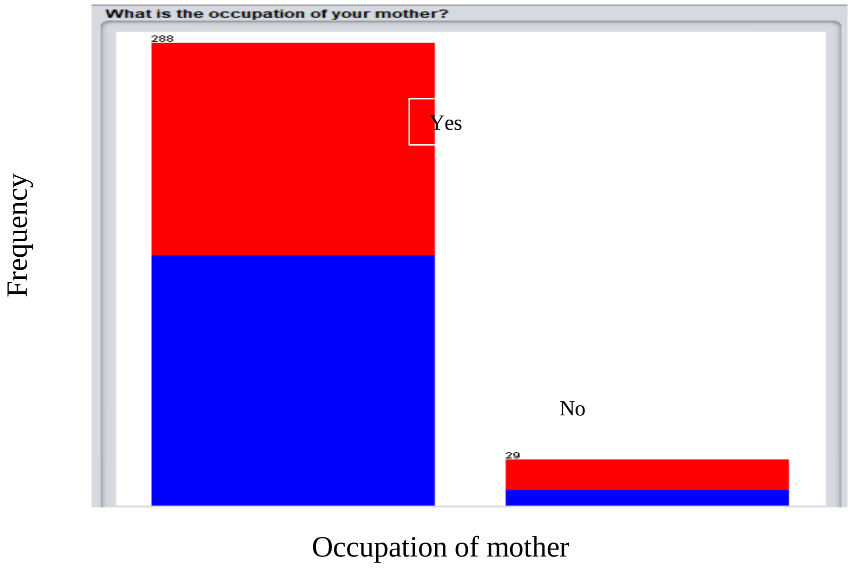


Figure 4.29: Bar chart of Mother Occupation (This bar chart shows the feature "encourage by your teacher" which belongs to two categories Yes and No. This chart show the Yes ratio that is 177 and the No ratio is 140.)

The next feature of our dataset is "Guardian" which belongs to three categories Father, Mother and Other. The Father ratio came according to our chart is 263, for the Mother ratio is 34 and for the other, the ratio is 20 according to our chart. The Accuracy of this feature against the target class "CGPA" is 68.75%. This accuracy is reported with the best classifier Support vector machine. The majority of students have their father as a guardian, as shown in this graph and very few students have their mother and another person as a guardian.

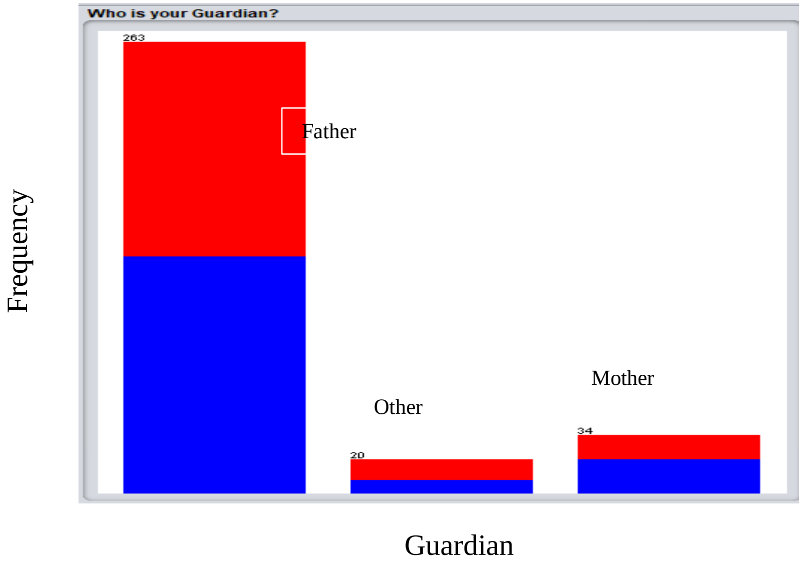


Figure 4.30: Bar chart of Guardian Time (This bar chart shows the feature "guardian" which belongs to three categories father, mother and other. This chart shows the father ratio that is 263, mother ratio is 34 and the other ratio is 20.)

The next feature of our dataset is "Educational Expense" which belongs to two categories Parents and Guardian. The parent's ratio came according to our chart is 242 and the data received for the Guardian ratio is 75 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students' parents bear or support their educational expenses, whereas only a small percentage of students' guardians finance their educational expenses.

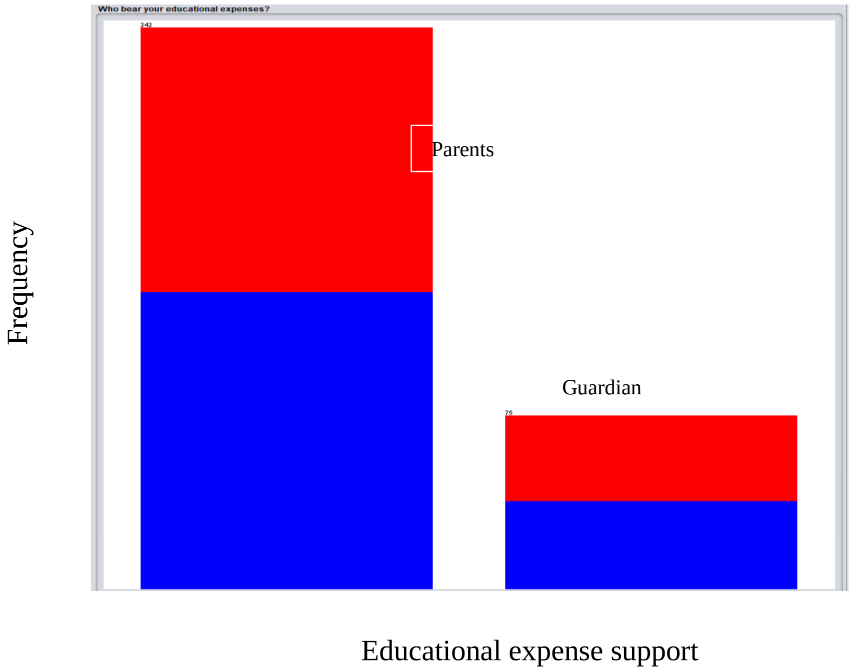


Figure 4.31: Bar chart of Educational Expense Time (This bar chart shows the feature "educational expense support" which belongs to two categories parents and guardian. This chart show the parents ratio that is 242 and the guardian ratio is 75.)

The next feature of our dataset is "Internet Facility at home" which belongs to two categories Yes and No. Yes ratio came according to our chart is 264 and the data received for the No ratio is 53 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 40.63%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students have access to the internet.

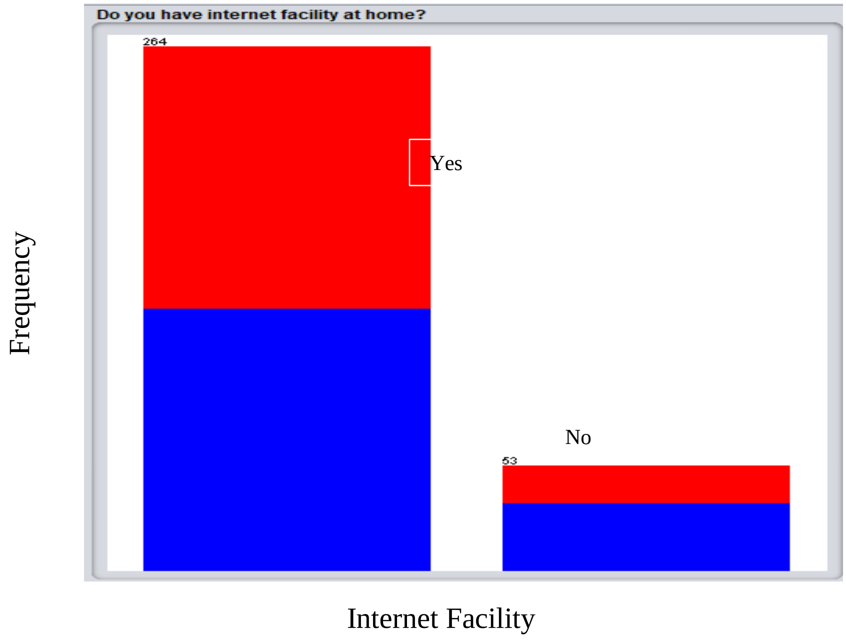


Figure 4.32: Bar chart of Internet Facility (This bar chart shows the feature "internet facility" which belongs to two categories Yes and No. This chart show the Yes ratio that is 264 and the No ratio is 53.)

- **Social Feature**

Social data refers to information of students based on their social characteristics such as

- Extra-curricular activities
- Go out with friends

The next feature of our dataset is “Extracurricular Activities” which belongs to two categories Yes and No. Yes ratio came according to our chart is 169 and the data received for the No ratio is 148 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph indicates that all students participate equally in extracurricular activities.

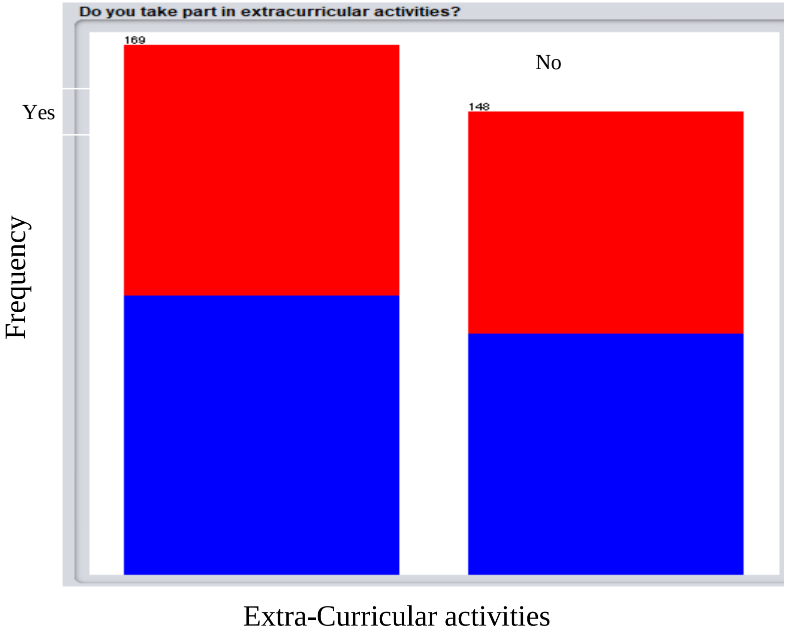


Figure 4.33: Bar chart of Extra-Curricular Activities (This bar chart shows the feature “extra-curricular activities” which belongs to two categories Yes and No. This chart show the Yes ratio that is 169 and the No ratio is 148.)

The next feature of our dataset is “Go out with friends” which belongs to three categories Weekly, Monthly and Twice a year. The weekly ratio came according to our chart is 44, the data received for the Monthly ratio is 98 and the data received for the twice a year ratio is 176 according to our chart. The Accuracy of this feature against the Target class “CGPA” is 40.63%. This accuracy is reported with the best classifier that is the Support vector machine. This graph indicates that the majority of students go out with their friends twice a year, with only a minority going out monthly and even fewer going out weekly.

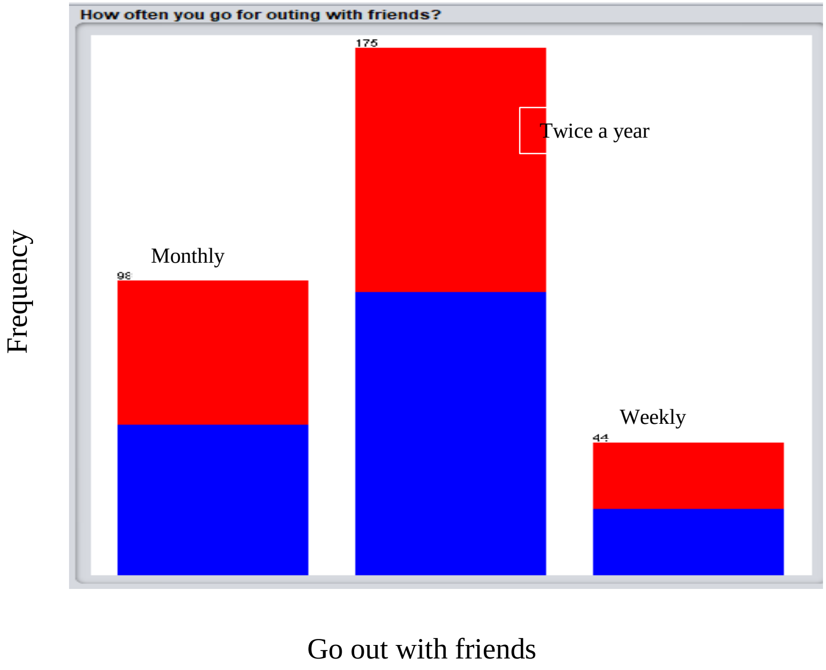


Figure 4.34: Bar chart of Go outing with Friends (This bar chart shows the feature "Go out with friends" which belongs to three categories weekly, monthly and twice a year. This chart shows the weekly ratio that is 44, monthly ratio is 98 and the twice a year ratio is 176.)

- **Course Related Feature**

Course related data refers to information about the course of the students which are currently studies in the university such as

- Course work provides sufficient information
- Improve analytical skill
- Course work market-oriented

The next feature of our dataset is "Course work Provide Sufficient Information" which belongs to two categories Yes and No. Yes ratio came according to our chart is 154 and the data received for the No ratio is 163 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 43.75%. This accuracy is reported with the best classifier that is the Support vector machine. This graph indicates that some students say no, university coursework does not provide enough information, while others say yes, and university coursework does provide enough knowledge.

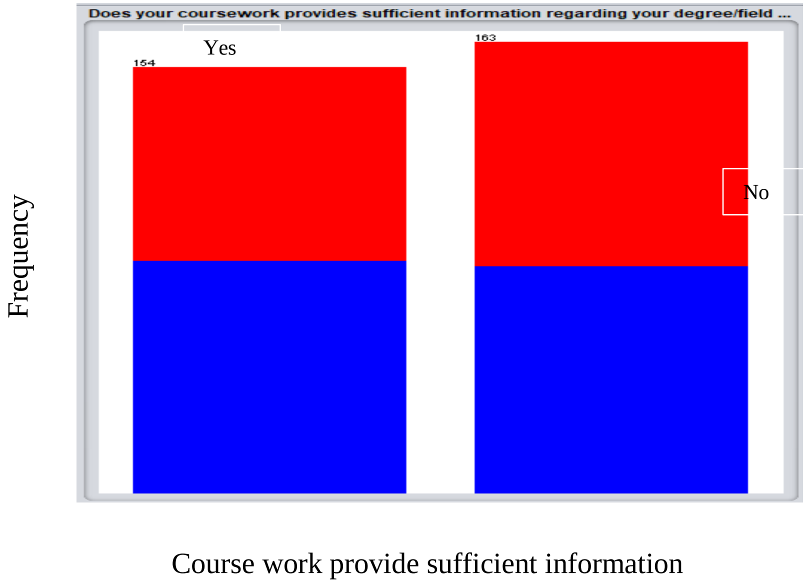


Figure 4.35: Bar chart of Course work Provide Sufficient Information (This bar chart shows the feature "coursework provide sufficient information" which belongs to two categories Yes and No. This chart show the Yes ratio that is 154 and the No ratio is 163.)

The next feature of our dataset is "Improve Analytical skills" which belongs to two categories Yes and No. Yes ratio came according to our chart is 162 and the data received for the No ratio is 155 according to our chart. The Accuracy of this feature against the Target class "CGPA" is 53.13%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students believe the coursework enhanced their analytical skills, while some students believe the coursework did not improve their analytical skills.

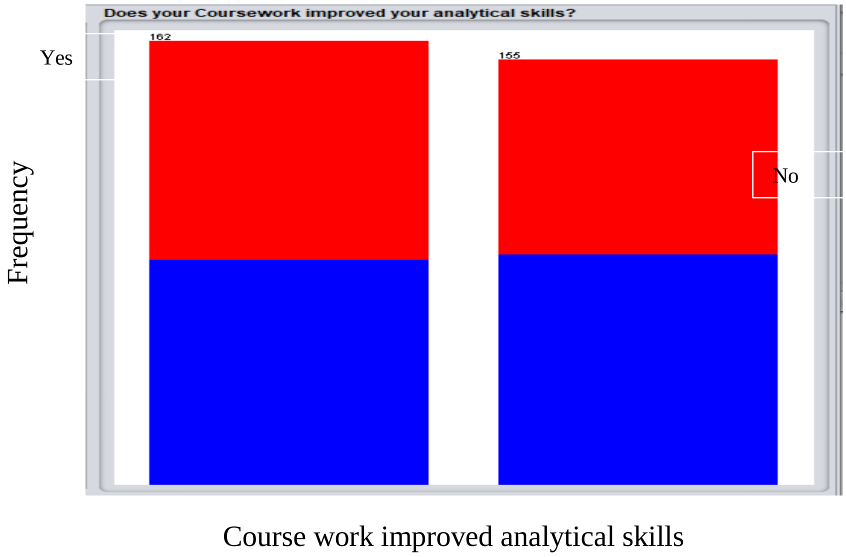


Figure 4.36: Bar chart of Improved Analytical Skills (This bar chart shows the feature "course work improves your analytical skills" which belongs to two categories Yes and No. This chart show the Yes ratio that is 162 and the No ratio is 155.)

The next feature of our dataset is "Coursework market Oriented" which belongs to two categories Yes and No. Yes ratio came according to our chart is 105 and the data received for the No ratio is 212 according to our chart. The Accuracy of this feature against the target class "CGPA" is 62.50%. This accuracy is reported with the best classifier that is the Support vector machine. This graph demonstrates that the majority of students answer no to the coursework being market-driven, while some say yes to the coursework being market-oriented.

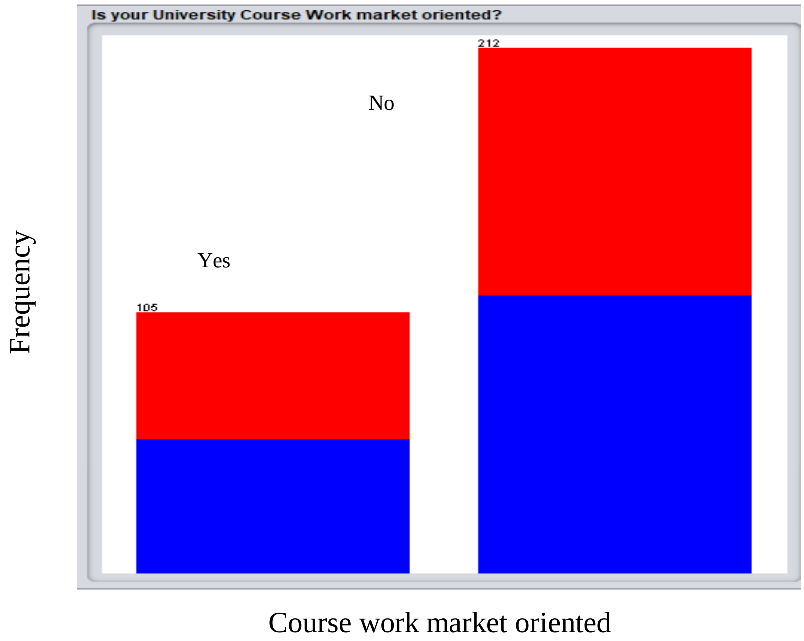


Figure 4.37: Bar chart of Course work Market Oriented (This bar chart shows the feature "course work market oriented" which belongs to two categories Yes and No. This chart show the Yes ratio that is 105 and the No ratio is 212.)

4.5 One to One Comparison

This research make the one to one comparison between the accuracy of existing study and the proposed study. This table show the accuracy of both the studies.

Table 4.1: Attributes with their accuracy (this table shows the accuracy of the individual features/ attributes of the research dataset. To find the individual feature accuracy, support vector machine classifier are used because support vector machine shows the highest accuracy.)

Attribute	**Accuracy**
Gender	62.5%%
Previous Degree	68.75%
Level of study	68.75%
Current Semester	68.75%
Address	46.875%
Family size	62.5%

Order in Family	56.25%
Parents Relationship Status	62.5%
Occupation of Father	40.625%
Occupation of Mother	65.62%
Guardian	68.75%
Travel time	40.625%
Weekly study hours	50%
Extracurricular activities	62.5%
Internet Facility	40.625%
Go outing with friends	40.625%
Educational Expense	62.5%
Health Status	50%
Encourage your teacher	62.5%
Marks increase or decrease during Covid-19	56.25%
Suitable System	62.5%
Coursework provides sufficient Information	43.75%
Analytical skills	53.125%
Coursework market Oriented	62.5%

4.5.1 Discussion

Table 11 shows the accuracy of individual features/attributes using the classifier. In this research, used four classifiers to find the accuracy but the results are reported with the best classifier. We obtain the accuracy against the target class which is “CGPA”. To find the individual features accuracy against the target class. Some of the features don’t show the significant results as expected for certain reasons. In the dataset, some of the features show the lowest accuracy/results. Let’s talk about the lowest accuracy features one by one. The feature “Father of Occupation” obtained lowest accuracy as compared to expected accuracy there are certain reasons associated with their lower accuracy one of them is that the accuracy has been distributed among three different trials that’s why we can’t get the expected results. “Weekly Study Time” obtained low accuracy as compared to expected accuracy and the reason is that nobody has known the definite study time in weeks. “Education Expense Support” obtained low accuracy as compared to expected accuracy and the reason is that fewer trials because in Pakistan the most parents are supported/ bear the educational expense of our children. “Go Out with Friends” obtained low accuracy as compared to expected accuracy and the reason is that most of the students meet with our friends in university regularly and go out with the friends in a year. “Travel time from home to university” obtained low accuracy as compared to expected accuracy and the reason is that most of the students’ lives in urban and the students who belong to the rural life and get admission in the university which is situated in urban then these students live in university hostels. The Overall accuracy of our dataset is good and is quite nearby the accuracy of the existing dataset. The accuracy of this dataset is 58.75% which is nearby the accuracy of the existing dataset.

4.6 Individual features Analysis

This section analyses the impact of each feature on student performance. For the experiments, twenty-four (24) features (picked by the feature extraction procedure) are chosen. Five classifiers (NB, SVM, DT, and RF) are employed in trials to examine the impact of each feature on performance of the student. Out of these classifiers, Support Vector Machine (SVM) shows the highest accuracy. To find the accuracy of an individual feature we used the SVM classifier. Using the SVM classification algorithm, we discover that the "Guardian, Previous degree, level of study in the university and current semester" are the highest accuracy features and the accuracy is 68.75% of the desired student's performance. Other characteristics are also significant. The feature "Travel time from home to university" has the lowest accuracy features as expected because students don’t know the exact travel time and the accuracy is 40.625%. Father’s occupation and Go out with friends are those features that have the lowest accuracy as expected because of the fewer trials.

The given proposition based on the Final CGPA feature set improve classification accuracy by analyzing the performance of the best and worst features. The second-best feature, "mothers working" is likewise part of the planned feature set of student personal information. When we compare the effects of other suggested features to the impacts of previous features, we find that our proposed feature provides higher accuracy than the old feature because we collect the data from the students of Pakistan.

4.7 Overall Student performance results

There are 319 tuples in the complete data set. The Weka tool, which is detailed in start of this chapter, is used to select an acceptable algorithm for the data set under evaluation. SVM is used to classify students into two categories based on the analysis: "pass" and "fail." In order to achieve the best results in terms of accuracy. The accuracy is measured using ten-fold cross validation, and the result is 62.5 % accuracy. The accuracy of all of the classifiers evaluated in this analysis is tabulated in Table 8. This table compares the accuracy of proposed and existing studies. These two studies utilized the same classifiers, but the outcomes were different, as seen in the table 8. This table shows that both the studies have the different target class. The existing study used "G3" as the target class and the proposed study used "CGPA" as the target class. These two studies used the four classifiers which are Naïve bayes, J48, Random forest and support vector machine (SVM). These four classifiers shows the different accuracies. The highest accuracy classifiers in both studies were the support vector machine classifier but the accuracies were different. The existing study highest accuracy were 64.14% and the proposed study highest accuracy were 62.5%. The proposed study accuracy was nearby by the existing study accuracy due to some geographical or cultural difference. Support vector machine algorithms outperforms the other algorithms because on out-of-sample data, SVM performs well and generalizing well. Another significant benefit of the SVM Algorithm is its ability to handle high-dimensional data. Discovering a hyperplane can be important in appropriately classifying data amongst distinct groups. Support Vector Machine is effective in finding the separating Hyperplane.

4.8 Visual Representation of individual features with accuracies

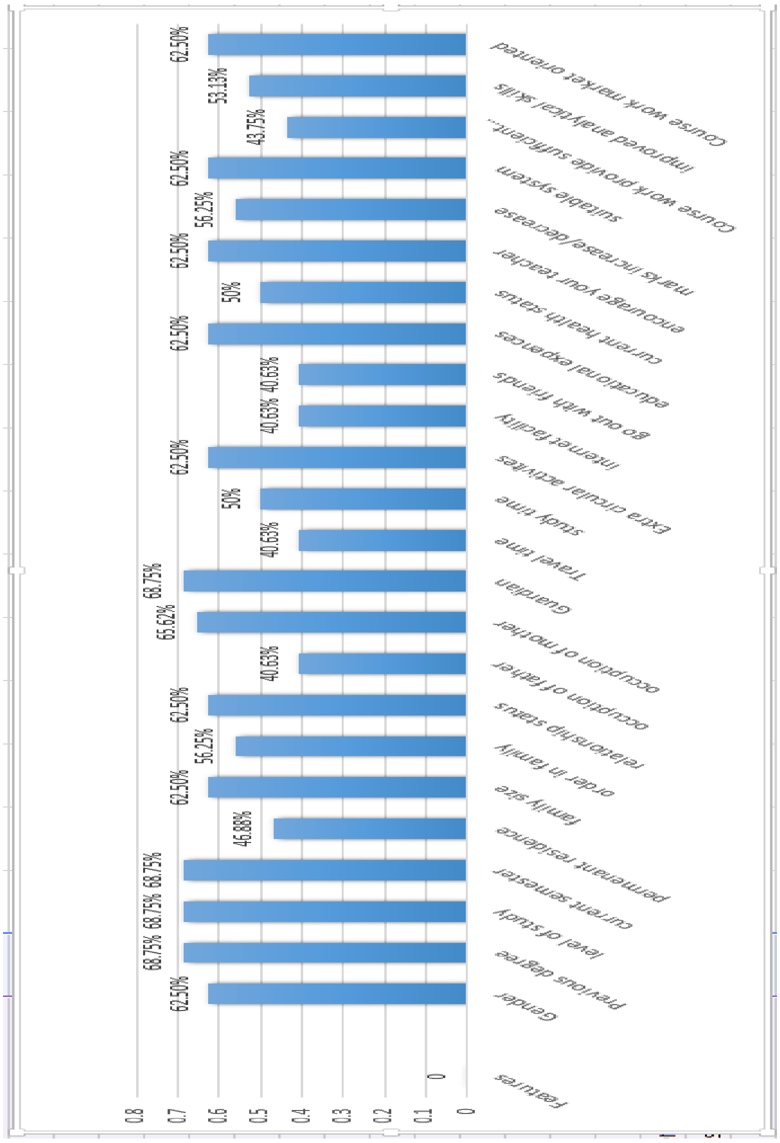


Figure 4.38: Visual representation of dataset with their accuracies

4.9 Weighted Average of Support Vector Machine (SVM) Classifier

Confusion matrix, specificity, sensitivity, false positive and the accuracy rate are the weighted average of the every classifiers. Here are the weighted average of the support vector machine classifiers.

- **Confusion Matrix**

a	b
14	6
6	6

- **Sensitivity**

Table 4.2: Confusion Matrix Values (this table shows the true positive, true negative, false positive and false negative values)

N=32	Predicted No	Predictive Yes
Actual No	TN=14	FP=6
Actual Yes	FN=6	TP=6

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{6}{6+6} = \frac{6}{12} = \boxed{0.5}$$

- **Specificity**

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{14}{14+6} = \frac{14}{20} = \boxed{0.7}$$

- **False Positive**

$$FP = \frac{FP}{FP+TN} = \frac{6}{6+14} = \frac{6}{20} = \boxed{0.3}$$

- **Accuracy Rate**

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{6+14}{6+14+6+6} = \frac{20}{32} = \boxed{0.62}$$

$$\text{Accuracy of support vector machine Classifier} = \boxed{62.5\%}$$

Specificity and sensitivity are inversely proportional: as sensitivity rises, specificity falls, and vice versa. Specifically, sensitivity and specificity refer to a test's consistency with a particular comparable, and PPV and NPV, respectively, indicate to the likelihood that a test can correctly determine whether persons have a target condition based on their test findings.

4.10 Research Validation

A validation dataset is a sample of data from your model's training that is used to estimate model competence while tuning the model's hyper parameters. The validation dataset is distinct from the test dataset, which is likewise kept from the model's training and is instead utilized to provide an independent evaluation of the final adjusted model's skill for comparing or selecting between models.

4.10.1 Existing Studies to validate the data

We have to validate the data base on the base paper accuracy. Existing studies are used to validate proposed study findings. One of the methods for validating the model is to comparing the existing studies and this method we use to validate the proposed study data. We checked the data against the data from previous research that they were utilizing. In our research, we employ the same data attributes that have already been used in previous studies.

4.11 10-fold Cross Validation

Cross validation is a model validation approach that is used to examine the statically analysis results in a separate dataset. This study used a standard cross validation method known as 10-fold cross validation.

The 10-fold cross validation approach is used in this study, which divides the dataset into 10 subsets and uses one of the subsets as the test set and the other nine subsets as the training set each time, and then repeats the process ten times. 10-fold cross validation is a technique for evaluating predictive models that divides the original sample into a training set and a test set for training and evaluating the model. In cross-validation, a set number of folds (or partitions) of the data are created, the analysis is conducted on each fold, and the total error estimate is averaged. This strategy ensures that this model's score is independent of how to choose the train and test set. The data set is partitioned into k subsets, and the holdout approach is applied to each of them k times.

4.12 Weighted Average of Classifiers

Table 4.3: Weighted Average of classifiers (this table shows the weighted average of the classifiers which are used in this research. In this research, four classifiers are used and these four classifiers weighted average are shown in this table.)

Classifiers	Sensitivity	Specificity	Accuracy
Naïve Bayes	0.5	0.55	53.125%
J48	0.66	0.5	56.25%
Random Forest	0.58	0.6	59.37%
Support Vector Machine(SVM)	0.5	0.7	62.5%

4.13 Confusion Matrix of highest accuracy Classifier

a	b
14	6
6	6

In this research, four different classifiers are used to find the accuracy of the dataset. Random forest, naïve bayes, support vector machine and j48 are the classifiers used in this research and these classifiers shows the different accuracies. This the confusion matrix of highest accuracy classifiers which is support vector machine. In this confusion matrix shows "a" works as the no class and b works as the yes class and "yes" and "no" are the target class variables.

4.14 Comparison of proposed study with the existing study

Table 4.4: Accuracy of combine features of existing dataset (this table shows the accuracies of these features which are common in this research dataset and the existing dataset.)

Proposed study Target class	Classifiers	Accuracy of proposed study	Accuracy of exiting study

Student’s CGPA	Naive Bayes	53.125%	55.3%
Students Final grade G3 Existing study Target class	J48	56.25%	58.97%
	Random Forest	59.37%	61.5%
	Support Vector Machine	62.5%	64.10%

The above table shows the different accuracies of proposed study dataset and existing study dataset using the four different classifiers. From many classifiers, these classifiers are used in this research because, in many existing papers, the authors was mostly used these classifiers to find the accuracies. ZeroR, Multilayer perceptron are the classifiers which are also used to find the accuracy but most of the existing studies were used these classifiers that’s why this proposed study considered these four classifiers to find the accuracy. This table show that both the studies used the same classifiers but they shows the different accuracy. The existing study highest accuracy was 64.10% and the SVM classifiers shows the highest accuracy and the proposed study highest accuracy was 62.5% and in proposed study SVM classifier shows the highest accuracy this thing is common this table. The proposed study accuracy was nearby the existing study accuracy. The comparison of both studies accuracy was shows graphically in the next sub section.

4.15 Visual Representation of proposed study with the existing study comparison

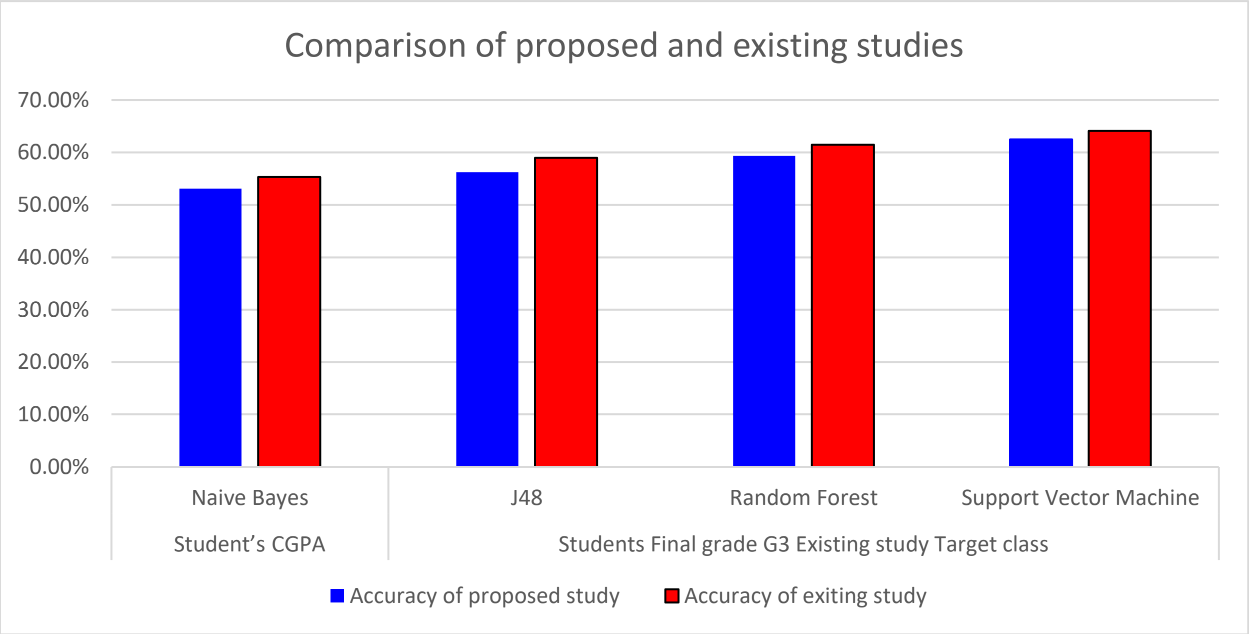


Figure 4.39: Visual Representation of Existing dataset Combine Features (this bar chart shows the accuracies of these features which are common in this research dataset and the existing dataset.)

Figure 42 shows the accuracies of the proposed study dataset and the existing study and these two studies used different classifiers. There are many classifiers used to find the accuracy. In this research, to find the accuracies used these four classifiers Naive Bayes, J48, SVM (Support Vector Machine) and Random Forest and these four classifiers shows the different accuracies. In this bar chart, blue color show the propose study accuracy and the red color shows the existing study accuracy. The visual representation of this bar chart show that the Support vector machine (SVM) classifier has the highest accuracy in both of the studies. The existing study have the highest accuracy which is 64.10% and proposed study have 62.5% and the classifier which have the lowest frequency that is Naïve Bayes in both the studies. The existing study have the lowest accuracy which is 55.3% and the proposed study have the lowest accuracy which is 53.125%.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview

Predicting student performance is a work-in-progress that includes contributions from a variety of researchers. In underdeveloped countries, this issue is still in its infancy. The student's high school achievement, ongoing evaluation, and demography are all used to predict academic development. Universities in developed countries have begun to make great progress in terms of evaluating student data. Apart from high school and university grades, the active involvement, behaviour, social activities, and financial position of students must also be examined. To predict the student’s performance, machine learning depends on data usage and algorithms. One should select the right machine learning solution to address the right problem to achieve maximum accuracy. We have compared methods selection in terms of their ability to improve the prediction results.

5.2 Conclusion

Our research dataset was analyzed with four different machine learning classifiers. Classifiers that were used were Naive Bayes, Random Forest, Support Vector Machine and J48. We have compared the results of our dataset with the existing dataset using the same classifiers. We have collected the data from the university level students and then pre-process the data using pre-processing techniques like data cleaning, missing values etc. On our dataset, the Support Vector Machine classifier shows the highest accuracy and outperforms the other classifiers. We have selected some features from the existing dataset because these features were used in 4 to 5 research papers, some of the features were not selected due to the demographic change. While trying to analyze the student's performance, the modification of input data is an important factor besides selecting the right method for data. When applied to different methods on student data, the result showed that- support vector machines outperformed other methods in predicting student performance.

5.3 Limitation

Due to the COVID-19 pandemic, we have implemented the questionnaire methodology to gather data from students of different universities. Google forms have been used for the creation questionnaire and after that mailed to target audiences for collecting their responses.

We did not use Deep learning algorithms like Convolutional Neural Network (CNN) due to the low sample size. The existing studies which they didn't use the CNN, because of sample size their sample size is the same as our sample size is.

We have a percentage of drop data because we cannot verify the data due to the COVID-19 pandemic. We don't have a data collection verification mechanism due to the COVID-19 pandemic.

5.4 Future Work

In future other machine learning algorithms like deep learning and Convolutional Neural Network, (CNN) can be implemented in predicting the student's performance. The research can be extended to college-level students. We can take the data from the students through interviews to get the verified data.

References

[1] M. Ramaswami, "A CHAID Based Performance Prediction Model in Educational Data Minning," 2017.

[2] S. S. K. S. A. B. M. N. &. H. P. G. S. Athani, "Student performance predictor using multiclass support vector classification algorithm," 2018.

[3] V. U. K. A. N. P. &. B. C. Z. Kumar, ". Advanced Prediction of Performance of a Student in a University using Machine Learning Techniques.," 2020.

[4] M. Mehil B Shah, "Student Performance Assessment and Prediction System using Machine Learning," 2019.

[5] S. Jayaprakash, "Predicting Students Academic Performance using an Improved Random Forest Classifier," 2020.

[6] B. &. B. M. M. Sravani, "Prediction of student performance using linear regression.," 2020.

[7] R. &. G. R. Ghorbani, " Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," 2020.

[8] J. Cheng, ""ata-Mining Research in Education," 2017.

[9] F. Ahmad, "The prediction of students' academic performance using classification data mining techniques," *2015.*

[10] M. Dayalan, ""Top Challenges in Data Mining Research," 2019.

[11] J. A. G.-P. *. a. A. D.-D. Juan L. Rastrollo-Guerrero, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," 2020.

[12] B. b. Hana Bydžovská, "A comparative analysis of techniques for predicting student performance," 2016.

[13] X. Wang, "Student Performance Prediction with Short-Term Sequential Campus Behaviors," 2020.

[14] A. S. A. I. S. A. B. a. A. C. ". I. Singh, ""STUDENT PERFORMANCE ANALYSIS USING,," 2016.

[15] A. Y. A. M. Ihsan A. Abu Amra, "Students performance prediction using KNN and Naïve Bayesian," 2017.

[16] R. J. A. M. R. &. S. R. (. Ahuja, " Analysis of educational data mining. Advances in Intelligent Systems and Computing," 2019.

[17] A. M. S. A. A. a. N. G. H. R. Asif, "R. Asif, A. Merceron, S. A. Analyzing undergraduate student's performance using educational data mining,"," 2017.

[18] F. (. Widyahastuti, "Wi Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron," 2017.

[19] M. V. T. P. a. V. H. M. T. Devasia, "M. T. Devasia, M. V. T. P, and V. Hegde, "Prediction of Students Performance using Educational Data Mining,," 2016.

[20] P. S. A. T. D. &. F. R. Cortez, "USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE," 2003.

[21] M. M. A. E.-h. A. M. Tair, "Mining Educational Data t o Improve Students ' Performance : A Case Study," *International Journal of Information and Communication technology Research,* 2012.

[22] M. S. Mythili and A. R. M. Shanavas, " "An Analysis of students ' performance using classification algorithms An Analysis of students ' performance using classification algorithms,," 2017.

[23] R. P. V. S. a. P. P.-,. .. P. Veeramuthu, ""Analysis of Student Result Using Clustering Techniques,"," 2014.

[24] ". A. I. Adekitan and O. Salau, "The impact of engineering student's performance in the first three years on their graduation result using educational data mining,," 2019.

[25] B. K. Bhardwaj, "Data Mining : A prediction for performance improvement using classification,"," 2011.

[26] A. A. T. A. S. a. S. A. T. H. Hashim, ""Data mining methodologies to study student's academic performance using the C4. 5 algorithms,"," 2015.

[27] A. Alhassan and B. Zafar, " "Predict Students ' Academic Performance based on their Assessment Grades and Online Activity Data,"," 2020.

[28] L. (. Gerritsen, " Predicting Student Performance with Neural Networks. May, 1–30.," 2017.

[29] C. &. S. S. R. Felix, "Predicting students' performance using survey data," 2020.

[30] E. &. D. D. Alyahyan, "Predicting academic success in higher education: literature review and best practices," 2020.

[31] M. R. K. S. S. S. P. I. &. E. M. S. F. Rimadana, "Predicting Student Academic Performance using Machine Learning and Time Management Skill Data.," 2019.

[32] A. M. D. A. J. M. F. R. &. C. E. B. Morais, "Monitoring Student Performance Using Data Clustering and Predictive Modelling.," 2014.

[33] H. B. U. &. F. A. Zeineddine, " Enhancing prediction of student success: Automated machine learning approach," 2021.

[34] J. L. &. K. S. A. P. Harvey, " A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning.," 2019.

[35] B. Z. R. X. G. S. C. &. Y. L. (. Guo, "Predicting Students Performance in Educational Data Mining.," 2016.

[36] J. V. N. S. S. &. C. A. (.] Dhilipan, "). Prediction of Students Performance using Machine learning.," 2021.

[37] E. H. Y. L. A. B. S. &. G. F. L. (. Tanuar, " Using Machine Learning Techniques to Earlier Predict Student’s Performance.," 2019.

[38] A. A. H. &. Y. F. (. Tarik, " Artificial intelligence and machine learning to predict student performance during the COVID-19.," 2021.

[39] D. Shanavas, "Application of Educational Data mining techniques in e-Learning- A Case Study," *International Journal of Computer Science and Information Technologies,* vol. 6, 2015.

[40] P. H. V. Devasia M, "Prediction of Students Performance using Educational Data Mining," *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE),* 2016.

[41] L. M. A. N. A. F. A. R. A. J. Daud A, "Predicting student performance using advanced learning analytics," *26th International World Wide Web Conference 2017,* 2017.

[42] W. K. P. Singh, "Comparative Analysis of Classification Techniques for Predicting Computer Engineering Students' Academic Performance," *International Journal of Advanced Research in Computer Science,* 2016.