Student's Performance Analysis and Prediction with the Help of Machine Learning Models

Author

Muhammad Faisal Masood

00000170850

MS-16(CSE)

Supervisor

Dr. Aimal Khan

Department of Computer Engineering College of

Electrical and Mechanical Engineering National

University of Sciences and Technology

Islamabad

August 2020

Student's Performance Analysis and Prediction with the Help of Machine Learning Models

Author
MUHAMMAD FAISAL MASOOD
00000170850

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Computer Software Engineering

Thesis Supervisor:
DR. AIMAL KHAN

Thesis Supervisor's

Signature:_____

Department of Computer Engineering College of
Electrical and Mechanical Engineering National
University of Sciences and Technology
Islamabad

August 2020

In the name of Allah most beneficent most merciful

وَلَا يُحِيطُونَ بِشَيْءٍ مِّنْ عِلْمِهِۦٓ إِلَّا بِمَا شَآءَ

And they can't
encompass anything
from His
knowledge, but to
extend He wills
[2:255]

Declaration

I certify that this research work titled "*Student's Performance Analysis and Prediction with the Help of Machine Learning Models"* is my work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged/referred to.

Signature of Student

Muhammad Faisal Masood

2020-NUST-MS-Soft-16

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. The thesis is also according to the format given by the university.

Signature of Student
Muhammad Faisal Masood
Registration Number
00000170850

Signature of Supervisor
Dr. Aimal Khan

Copyright Statement

Acknowledgments

I am thankful to Allah Almighty for giving me countless blessings and guidance in this research work. It is all possible because of His help. I want to thank my teachers **Dr. Aimal Khan**, **Dr. Farhan Hussain**, **Dr. Arslan** and **Dr. Farhan Riaz** for supporting me throughout the whole work and giving me ideas to improve my research. I couldn't have done this without them. They are the source of great knowledge and their expertise helped a lot in achieving my goals. I also want to thank and appreciate the efforts of my parents, siblings and friends especially "Salal Aslam", who encouraged me throughout the whole degree and made the things easy for me with their motivational assistance. I would like to express my gratitude to all the mentioned people and my department for assisting me in my publications.

Dedicated
to

To my Parents, family, friends and Advisors

Abstract

Educational Data Mining (EDM) has become one of the most important fields now a day because, with the development of technology, student's problems are also increasing. These problems can be related to student's behavior, parents' participation or academic etc. To tackle these problems and help students, educational data mining has come into existence. The prior prediction of student's performance is necessary so that useful steps can be taken out to help him and guide him in the correct direction. In this research, a systematic literature review (SLR) has been performed to get 20 studies (2012-2019) in the area of EDM. The reason behind SLR is to get advanced machine learning models that have been used by researchers in their field so that we can compare them with each other to get the most useful model among them. Feature extraction and data augmentation techniques will be used to enhance their performance and predictions. After detailed SLR, 11 highly advanced machine learning models have been obtained. These models are further applied to 2 public databases to check their performance and prediction rate. It is observed that "Random forest" and "Decision tree" are the best machine learning models having an accuracy score of 95% and 96% respectively. To validate results, database had been splited into 70/30 ratio. 70% database was used to train models and remaining 30% database was used to test and validate results. With the help of these experiments, weak students can be easily identified and proper precautions can be taken to help them. In future work, these models can be implemented on real-time university or school databases to further enhance their accuracy and performance scores. With the help of student performance, student's areas of interest and the future job can also be predicted.

Key Words: *Data Mining, Machine Learning Models, Student's Performance, Public Databases*

Table of Contents:

List of Figures

List of Tables

Chapter 1

INTRODUCTION

According to IBM, “The 90% of World’s data has been produced in the last two years. Every day, World creates 2.5 quintillion bytes of data” [1] and the situation has become so worse that it is becoming difficult to even store and save that data which World is generating every day in the form of comments, posts, articles, pictures, blogs, research papers etc. Today, 98% of the World’s data is in digital form and it is nearly impossible for a human to analyze and use such large amounts of data. This type of situation is both blessing and problematic for human beings. We can call it problematic because the human brain is not powerful enough to handle all of this data and use that data in the creation of something useful. It is a blessing because today’s computers have become so powerful that complex equations can be solved in the blink of an eye so they can not only process such large amounts of data but also create meaningful applications from it [1].

According to the International Data Corporation (IDC) and Seagate, the overall World’s data will grow up to 163 zettabytes by 2025 due to an increase in intelligent systems and their customer’s data collection capabilities. Each day, IoT based systems collect user’s data based on their interaction with systems and send it back to their company’s server so that complex algorithms developed by these companies can understand different choices of customers to help them improve their system’s performance. In the future, understanding big data generated by intelligent systems will become a key basis of competition between companies and it does not matter either that data is in a raw, semi-structured or structured form as long as it is being used for the creation of something meaningful for human beings. The one who can understand the meaning behind data and use it to create something

meaningful will win this race and that is the reason Facebook, Amazon and Microsoft are investing billions of dollars on Artificial intelligence [1].

1.1 Machine Learning:

Computers can understand digital data easily and they can solve complex patterns easily as compared to a human. This type of concept is called machine learning. Machine learning is basically the ability to learn different patterns and ways from big data (according to rules which are based on mathematics) and predict some results which can be useful in solving different real-life problems. These rules are not hardcoded as they can be defined again and again because they depend on the structure and quality of data fed to machine learning algorithms [2]. Machine learning consists of many algorithms that can be used to train every kind of dataset. It does not matter either database has labels or not as machine learning algorithms can be used for both labeled and unlabeled databases. Labeled databases are those databases that have labels with every column to define what is the data exactly about and unlabeled databases are those databases that do not have labels with columns.

1.2 Machine Learning Applications:

There are so many applications of machine learning currently being used by companies like Amazon, Facebook, and Google etc. Let's talk about search engines, machine learning is helping search engines to better understand and build a relationship between web pages and search keywords used by users. Machine learning models analyze website's content to extract those keywords which can better describe that website and according to search engine optimization techniques, search keywords are the most important thing for any website because user reaches website through these keywords [3].

Machine learning techniques can also be used for image recognition and to identify objects in a video or pictures such as faces, car's number plates, or students' thumbprints etc. At first machine learning models extract the background of any image to detect an object's edges then it compares that object with the required object to see either it is that object or not.

Machine learning is not recommending products to a user based on the searching history or the user. It learns which types of product user is mostly buying from the online stores and then it starts recommending similar products to that user in order to help him get different

choices. If we observe the above-described examples, we will know that machine learning models always follow a simple procedure, at first, it gets database, it trains its self with different scenarios and then it predicts similar results.

1.3 Types of Machine Learning Algorithms:

There are three types of machine learning algorithms supervised machine learning algorithms, unsupervised machine learning algorithms and reinforcement machine learning algorithms. Let's discuss them in detail one by one.

1.3.1 Supervised Machine Learning:

Supervised machine learning algorithms are used for those databases which have labels with them. Common supervised machine learning algorithms include decision tree, support vector machine, random forest and neural network etc.

1.3.2 Unsupervised Machine Learning:

Unsupervised machine learning algorithms are used for those databases which have no labels with them. Common unsupervised machine learning algorithms include k-mean clustering, apriori, mean shift and dimension reduction etc.

1.3.3 Reinforcement Machine Learning:

Reinforcement learning is based on punishment and reward systems. It is basically used in games where on any achievement, the character gets rewards in the form of points, XP or lives whereas, on a failed mission, the character gets punishment in the form of loss of health or life.

1.4 Machine Learning Process:

There are many steps involved in the gathering of data to making useful applications from it. At first, raw data will be collected and that data may consist of numeric values, images, audios etc. After that with the help of different tools and techniques, that data will be refined (discarding unnecessary data, filling out missing values, adding more values where necessary) making it more suitable for further processing to increase its quality. Data quality

is very important for machine learning-based applications because the application's accuracy depends on the quality and quantity of data. After refining data, it is time to split this data into training and testing part. Now the question arises, why we need to split the data into training and testing parts. In order to understand this, we first need to know about the concept of overfitting. In order to make a perfect application, it is necessary to make it able to accept all kind of data so that I can train itself and predict the useful outcome but when we restrict it around the same database then there is a possibility that it might now work perfectly with other databases so, in order to avoid this situation, we split data into training and testing parts. We use the training part to train the application with the help of machine learning algorithms (we will discuss all machine learning algorithms in detail in upcoming chapters) and the testing part is used to test that trained data to predict useful outcomes and check their accuracy against already available output. If accuracy is more than 90% then we can move forward and test the unseen database with the help of the developed application. This same process is used in pretty much every machine learning-based application to train and test algorithms. The most important thing in any machine learning-based application is its database which is used to train and test algorithms. The quality and quantity of database matters because the whole accuracy and prediction part depends on it.

1.5 Problem Statement:

The focus of this thesis is on the educational field especially on student's problems and their solution with the help of machine learning models. There are so many methods being used by teachers to detect underperforming students and to help them but none of them are producing reliable results. Furthermore, there are so many machine learning models and we do not know which models are best to tackle student's performance prediction problems.

The biggest problem is to get the student's database which is the key point in this research. We want to train all machine learning models on student's database and try to predict their performance either they will get success or not and also we want to find the best machine learning model which will give higher accuracy than all other models so that in future we can use that specific model and implement different changes in the resultant product.

1.6 Motivation:

With advancement of technology and low job rates, the pressure to get a job and have a successful life is making students depress. They are losing focus on study and trying to find different wrong ways to overcome their depression. We have a number of the classroom in each school, college and university and there are thousands of student getting an education but we cannot detect the emotional state of any student's mind so what if we can detect any student's future performance or predict their result, we can get to know about weak students and by using different methods we can get to know their problems and try to solve them. We can not only solve their problems but also recommend teachers to improve their teaching methods in case students are facing problems in understanding their lectures. With the help of the machine learning technique, we can not only detect weak students from the majority but also predict their performance in the final exam. With some suitable databases in the future, we can do wonders with the combination of machine learning with an educational field like predicting job opportunities for students, recommending jobs or different fields to students or helping students to self-evaluate themselves through our research.

1.7 Objective and Contribution:

Due to fewer jobs and high competition in the market, student's depression is increasing day by day as we are still following old school and teaching methods which are not helping the new generation of students in getting out of depression and finding paths to adopt for their future. This problem can be tackled with the early detection of student's future performance as it is very tough for teachers to detect underperforming students among hundreds of students. If the teacher becomes successful in the detection of underperforming students then possible measures can be taken to help them out and the teacher can guide them towards the right path. It can also try to solve some of the personal problems which can be making students depress and they are not able to focus on their studies. Our main objective is to solve this problem with modern technology which is called machine learning. With the help of machine learning, we can train models on student's database and that trained model can help us to predict student's future performance. We will first detect student's problems and current methods of the solution being used by teachers to tackle those problems. Then

we will find out which famous machine learning models are being used by researchers to tackle these kinds of problems so that we can train all of those machine learning models on the student's database and pick the best one among them based on accuracy score. We can also observe that how can this research opens new ways to help students in getting a successful future.

1.8 Outline:

The rest of the paper is organized as follows: In Section 2 presents related work, section 3 contains the model/method and performance of student prediction, while section 4 illustrates the results and analysis, and lastly, the paper is concluded in section 5.

1.9 Summary:

In this chapter, we have studied different problems being faced by the World due to the increase in digital data day by day. We have discussed how we are wasting this golden opportunity of not using billions of databases to get useful results and predictions. We have seen how by using machine learning we can use these databases to train models and predict useful outcomes. We have studied machine learning applications, machine learning model's types and complete process of implementation of the machine learning model with databases and predicting useful results from them. We have discussed our problem statement and why we are going to use machine learning models to analyze student's performance and predict their future results. In the end, we came to know that how this research can help students and teachers to tackle different problems related to study and teaching methods.

Chapter 2

LITERATURE REVIEW

In this research work, a detailed systematic research review has been carried out. There are two parts of this literature review. Part 1 is to find out problems faced by students during their semester in different subjects and part 2 is about finding different machine learning algorithms which can help out predict student's future performance successfully so that in case of weak students, useful measurements can be taken out to help them to get success in future exams. I have also tried to find out the best machine learning model among all identified machine learning algorithms by applying each one of them on 2 public datasets. The literature reviews, its summarizations and facts and figures are given below.

2.1 Literature review Steps:

The scope of this review is restricted to journal publications found between 2011 and 2019. The main search criterion was “Problems faced by students”, “Student performance analysis”, “Student performance with machine learning”, “Machine learning models”, “Machine learning techniques”, “Machine learning predictions” and ”Machine learning methods”.

2.2 Review Protocol

Development of a review protocol comprises 4 discrete steps namely: inclusion and exclusion criteria, search process, quality assessment, and data extraction and synthesis as recommended by “Hanny Tufail” [30] for SLR.

2.2.1.1 Inclusion and Exclusion Criteria

Only journals published from 2011 to 2019 are considered for inclusion. Parameters defined for inclusion criteria are a) Student performance Analysis. The dataset (real and virtual), Model used, the accuracy and performance and future work for further research b) Selected research work must belong to these databases: IEEE, SPRINGER, ACM, Taylor and Francis and ELSEVIER. c) Included research work must be results-oriented. In comparison to

inclusion criteria, exclusion criteria eliminate researches published before 2011 along with those researches with weak validation methods.

2.2.1.2 Search Process

This process is based on search terms that are defined. The resulting number of researches were filtered from 2011-2019. "Figure 1" represents the main steps of our search process.

2.2.1.3 Quality Assessment Checklist

This checklist has been developed to validate the selected research in accordance with the guidelines by [30] It comprises of following steps: 1) Include the researches from authenticated databases 2) Include the studies where outcomes are validated in proper way 3) Include the latest studies as much as possible. The aforementioned quality points ensure the reliable outcomes of this SLR.

2.2.1.4 Data Extraction and Photosynthesis

This subsection refers to the selection of relevant information from 24 research papers.

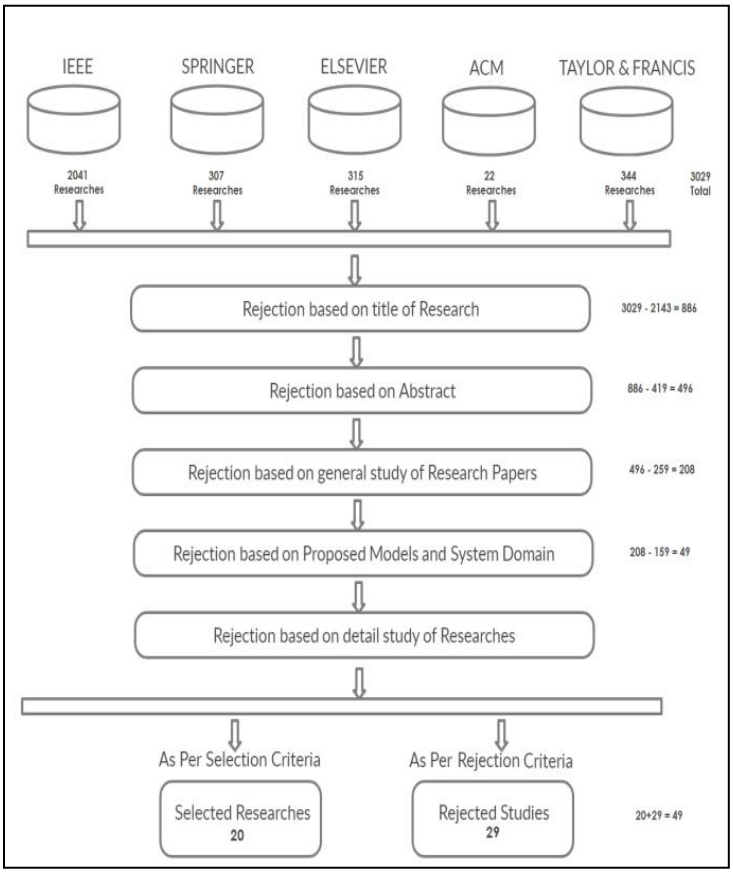

Figure 1: The search process of literature review

Table 1: Search results of keywords

Sr. #	Search Terms	No. of Search Results				
		IEEE	SPRINGER	ELSEVIER	ACM	Taylor and Francis
1	"Machine Learning Models"	448	60	122	4	61
2	"Machine Learning Techniques"	606	166	29	5	39
3	"Machine learning Predictions"	550	6	59	6	74
4	"Machine Leaning Methods"	398	72	102	4	72
5	"Student Performance Analysis"	33	1	1	0	97
6	"Student Performance with Machine Learning"	6	2	2	3	1

In the above Table 1, summarizes the number of results for each keyword. I have used both AND and OR operators to search for the related research papers for this systematic literature review. The total research papers which I have gathered were 3029 and after applying data extraction techniques like target years, topics related to my research abstracts related to my research etc. I got 25 final research studies that were related to my thesis.

Table 2: Data extraction elements

Sr. #	Description	Details
1	Information of Bibliographic	Paper's title, Author's name and year of publication, etc.
Data Extraction		
2	Overview	The first object of this study is to find out which machine learning models have been used from 2012 to 2019 for training data and results prediction. The second objective is to implement those models on 2 public databases to find out which one of them is more accurate.

3	Results	Models, Predictions and Solutions of problem
4	Data Collection	Quantitative methods used
5	Validation	Proper research has been done and verified by experiment on databases
Data Synthesis		
6	Models	Models Used for Machine Learning Process. (Table 3)
7	Predictions	Results obtained by applying machine learning models on databases. (Table 4)
8	Solutions	Actions recommended to solve students problems (Table 5)

2.3 Literature Review:

Students are the main asset of any institution and without them; there is no meaning of having an institution in the first place. Student's performance plays a key role in any institution's success because the success of any intuition depends on the numbers of researches and graduates by that particular institute. There are numbers of factors which affects student's performance and many institutes try to deal with them so that they can help their students in getting good marks and bright future. According to the author of [1], there are many factors which affect student's performance like parent –based factors, school-based factors, student-based factors and teacher-based factors. It also includes the student's motivation level, classroom management, learning environment, family background, parent's income, their behavior with their kids, attitude towards their jobs, time given to their kids, parent's education etc.

In [2], author did a detailed research on the factors affecting student's performance. According to the author, there are three factors which normally disturb student i.e. school-based factors, socio-cultural factors and personal factors. These factors include under staffing, poor teaching skills, inadequate learning material, poor attitude and less motivation by the teachers. These factors disturb a student's life and make it less confident, expressive and friendly towards people. The author "WycliffeAmukowa" [3] also claimed that school and its facilities, student's previous academic background, student's house environment and qualification of teachers effect student's performance a lot. According to the research by author of [4], the factors which can create positive thinking in students are good family environment, healthy hostel facilities, adequate

educational material and resources, good teaching methods, good and motivational attitude towards study and involvement of student in group discussion related to subjects.

Since the introduction of the modern technology, data is increasing day after day and according to the latest figures [5], World creates around 2.5 quintillion bytes of data every day and it is still increasing due to the introduction of smart devices in everyday life. Now it is impossible for humans to read and locate meaningful combinations of pattern from such large amount of data because even capturing and storing that data is a problem due to its size. Technology giants are using different tools, techniques and graphs to manage the data and find different patterns among them. They are using artificial intelligence and machine learning techniques to organize that data, process it, train it by using machine learning algorithms and then predicting meaningful outcomes through it [5].

Wireless sensor networks with the combination of machine learning models can be used to detect and predict pest/disease in agricultural areas in the country. In [6], author got the database of pest/diseases in agriculture lands in India and with the help of Naïve Bayes Kernel algorithm; he successfully detected the disease's patterns and predicted when diseases are going to appear in agriculture land in the future. With the help of wireless sensor, Raspberry Pi and machine learning model named "Naïve Bayes Kernel"; it was possible to predict future attack of disease on the agricultural land.

Finding the shortest route has always been a problem for experts and they have used different methods to find out the shortest distance path between two points but in [7], author decided to use a machine learning model named "K-nearest Neighbors" to solve this problem. He used a database consist of map of stores and a list of available products. After dividing database into 30/70 ratio, he used 70% data to trained machine learning model and then used remaining 30% to check and predict results. The accuracy of this process was reasonable and perfectly predicting the shortest route between two points. The complexity of the machine learning model was O(VlogV).

Online education has become so much popular due to the introduction of modern technologies and distant learning techniques. A person can easily get enrolled in any university located at the other side of the World and go through the learning process including monthly tests, midterms and final exams and get certification from that university by sitting at home comfortably. It is also making it difficult to identify those students who are struggling in some particular subjects

and unable to perform well in exams. In [8], author used some machine learning algorithms to identify those students with the help of their performance throughout the semester. He used their database which includes their test marks, midterm marks and quiz's marks with machine learning algorithms i.e. Neural Networks, Support Vector Machine, decision trees and cluster analysis. He divided database into two parts and used one part for training and other part for testing and with the help of machine learning model, he successfully identified weak students which were not performing well and who could fail that particular subject in finals. The accuracy of these machine learning models was 83% which could be increased with the increase in database and by using more accurate and quality data of students [8].

One of the many treatments of cancer is "Radiation" but it has a lot of side effects as it can also destroy normal cells along with effected ones. In order to protect normal cells p53 inhibitors are being used during the treatment process because they have low toxicity against human body but the design of the p53 is very complex and it requires a lot of cost and time. In [9], author proposed a new method by using machine learning models named "Random Forest" and "Support Vector Machine" to train models on database including two and three dimensional structures of the compounds. They have used Pareto Ranking method for ranking purpose. Their proposed method shows useful ranking for drug discovery.

There are different types of data which can be used to train and predict useful results i.e. Stock exchange data, different games like cricket, hockey, football's score data, educational field data which can be used to predict student's performance in future etc. Educational field data is very important as it can help students to organize their life and save their failing subjects if use correctly and efficiently. In [10], author has decided to use student's final year data which is called GPA to train three models on it so that he can predict final year result for students to help them identify failing subjects. He used Deep Learning, Decision Tree and Generalized linear model with "Rapid Miner" software to perform this experiment. According to Author, from this experiment, useful factors can be known which are affecting student's performance so that they can better prepare themselves for the final exams.

According to [11], Machine will play a big role in educational field in upcoming years. It will not only provide some very effective tools to predict the future progress of students but will also help them to pin point those factors which can affect student's progress. Author has used Gartner Analytics Ascendancy Model which needs 4 types of data analytics including descriptive,

diagnostic, predictive and prescriptive. Different types of machine learning algorithms i.e. Naive Bayes classification, decision tree classification, random forest classification, linear regression, k-near neighbor classification, artificial neural network regression and classification, support vector machine classification and logistic regression have been analyzed and tested. Based on accuracy of these machine learning models, author described some recommendation to select, setup and utilization of machine learning models in educational field. In his research author further did a survey to know the opinion of undergraduate and graduate students of computer science field about what they think about the role of machine learning in educational field in future.

With the evolution of modern technology, everything is converting into digital and producing so much data that is even difficult to handle now. Educational fields are also evolving and with the introduction of e-learning, it has also become a big data problem. In [12], Author is researching on different data mining techniques so that he can get insights of students' learning patterns which effects academic performance. In order to achieve his goal, author has used three machine learning algorithms i.e. decision tree, neural networks and naïve bayes along with predictive analysis technique. Author believes that integration of big data with educational field will not only show promising results but also help universities to improve their education system and help their students in their fields.

For any educational institute, students are the most important asset and their performance in their specific fields matters a lot. With the evolution of modern technologies, educational institutes are also trying to discover new ways to help their students and identify their problems in education. According to the author of [13], there are so many factors which can improve performance of students in their fields. In present day, educational institutes are using data mining techniques along with classification methods to improve their quality of education. Author has proposed a new technique to improve the accuracy of classification methods by combining different classifiers together. Author has used "AdaBoost" ensemble with an algorithm named "Ada-Ga". With the help of new technique, Author believes that they will be able to identify those student who are at risk of failing their semester at early stage based on their previous progress in that particular semester.

In [14], Author focused on a single machine learning classifier named "Quadratic Discriminant Analysis" and explained its basic background so that readers from any level of understanding can

understand how to use and implement it on different applications. This paper is all about quadratic discriminant analysis, its mathematical forms, numerical implementation and comparison between linear and quadratic classifier and how to solve singularity problem by using high dimensional datasets.

Software development industry is a huge and risky industry in which a single error in software can cause a lot of damage in business. Many techniques are being used by software developers and researchers to early detect errors so that patch can be implemented on time to save the customer from huge loss. According to the author of [15], with the help of machine learning models named "Random Forest" and "Neural Network", software defects can be identified on larger scales now. Author has used a dataset which has been collected from more than 500 software applications in which more than 150 software static analytic analysis measurements have been used. Author is trying to determine that either early prediction of defects can reduce defect rates or not.

Machine learning in medical field can show wonders by saving so many lives by predicting diseases at early stage. Researchers have been trying to find ways to combine machine learning algorithms with medical data sets which can show high accuracy in predicting diseases and damaged cells especially in breast cancer treatment. Author of [16] has presented its research in which the main focus was the prediction of survival time in breast cancer based on clinical data. Three machine learning models named decision tree, linear regression and support vector machine have been used which are also predicting promising results with high accuracy in detecting survival time in breast cancer. In order to confirm results of the models, author has used cross-validation techniques which are based on four parameters for error evaluation.

With the evolution in modern technologies, online education is also getting popularity because any person can get education from popular university of its choice while sitting at home according to its own timetable and secure a certification from that university however it is also difficult to find out those students who are struggling and getting problems in online learning. In order to resolve this problem author of [17] first got open database of students of e-learning platform named "edX" and then applied different machine learning algorithms on it to predict students' performance in future in order to find out weak students who can fail that particular subject or semester. Author used many machine learning models including logistic regression, support vector machine, naïve bayes, k-near neighbor and bayes network. At first author refined

students' dataset by removing those features which are unnecessary or unimportant. Then author used different feature selection approaches in order to find out those features which matters a lot and they can help in predicting student's result accurately [17]. The final result proved to be effective in predicting student performance and helped teachers to find out weak students so that particular precautions can be taken out in order to help them.

Since the introduction of the modern technology, data is increasing day after day and according to the latest figures [5], World creates around 2.5 quintillion bytes of data every day and it is still increasing due to the introduction of smart devices in everyday life. Due to increase in data, it has become difficult to get important results and extract meaning information from educational databases. In [18], Author has done a systematic literature review to find out the answers of some questions related to the problems in student's performance in education. According to author, Malaysia lacks systems to analyze educational databases and it cannot monitor student's progress due to the same reason. The first reason is that researchers do not have sufficient methods to predict student's future performance in particular semester or subject and second reason is that they do not have enough knowledge about the factors which are affecting student's performance. The main focus of the research is on finding suitable data mining techniques which can help in prediction of student's performance and also to find out those factors which are affecting student's results [18]. Author has find out that with the help of student's previous GPA, assignment marks, quizzes, lab work, class test and attendance it is possible to predict student's future marks in some particular subject or semester. Author found out some machine learning algorithms like decision tree, artificial neural network, naïve bayes, k-near neighbor and support vector machine can help us in prediction process [18]. Author believes that with the help of machine learning algorithms student's achievement can be improved and they can get success in final exams more effectively. It will not only help students and teachers but also academic institutions because their overall performance will improve.

In today's World it is very important to find out those students who are going through stress and mental illness because due to increase in competition in this modern age and low job rates, every other student is stressed out about its CGPA and final results. Some can handle this stress and some cannot handle it properly and as a result unpleasant incidents might occur repeatedly. In previous researches, researchers mostly used static factors like educational background and results from many questionnaires in order to analyze student's performance but due to the

introduction of data mining techniques and machine learning algorithm, more important factors like student's previous semester CGPA, GPA, assignments, quizzes, midterms and attendance etc. have being considered to find out about students future performances. In [19], Author has used Naïve bayes, rule learner and decision tree machine learning algorithms to train them by using student's educational dataset and predict their future performance based on it. According to author, with the help of machine learning algorithms, they have successfully identified low and high performing students with reliable accuracy.

This will help teachers to target those students who are under achievers and help them in their education and give them more importance than other students. In the same way in [20], author has used naïve bayes classifier to predict student's performance who are at the risk of being drop out in order to help them in their study. The main goal of author is to help teachers to find out weak students who are under a lot of stress due to their poor grades and in order to train naïve bayes machine learning algorithm, author has used database of three undergraduate engineering courses of one of the largest Brazilian public university.

Although the World's data is increasing at an alarming rate but machine learning is also proving itself an effective tool to use that huge database and with the help of some machine learning algorithm, predict useful results. Educational data mining is a new field and a lot of researchers are trying to come up useful techniques which can help student's performance and predict their future results so that suitable measurements can be taken out to help them. In [21], author has done some research on internet usage by the undergraduate students. According to author, internet usage has a civilizing influence on student's performance and their living. Author used three machine learning algorithms i.e. decision tree, neural network and support vector machine on a database of 4000 students which includes features of student's online duration, internet traffic volume and their connection frequency [21] to predict the effect of internet on student's performance. According to author, behavior discipline plays a key role in student's academic success and with the help of machine learning algorithm, the effects (positive and negative) can be easily identified however the accuracy depends on the numbers of features used to train machine learning algorithms. Accuracy is directly proportional to the numbers of features used during machine learning training [21].

According to the author of [22], higher educational institutes mostly struggle with problems related to student's performance because they mostly depend on on teacher's experience of

identifying those students which are going through problems in semester. Author has tried to use a different approach to deal with this problem and according to him, machine learning algorithms are the best solution to identify those students as they can predict student's performance in future exams by learning previous examples of similar students. Author has used different machine learning algorithms i.e. random tree, random forest, multi-layer perceptron, naïve bayes and decision tree and trained them on student's dataset and successfully identified weak students with reliable accuracy. Author believes that with the help of machine learning algorithm, we can help students who are going through difficult time and with more teacher's attention, we can solve their problems and help them to get success in future.

In [23], Author used machine learning algorithms to predict in-hospital mortality after transcatheter aortic valve replacement in United States. According to author, hospital's current risk prediction tool has been designed by using statistical modeling approaches and it has some limitation which can be fixed by using machine learning algorithms. Author took patient's database from National inpatient sample database from 2012 to 2015 and divided that database into 70/30 ratio. Author used 70% data for model training and 30% data for testing purpose. Author used logistic regression, neural network, random forest and naïve bayes machine learning algorithms for training purpose. Total 10,833 TAVRS were analyzed in this process and accuracy was more than 80% which is considered reliable in this case. The best machine learning model which gave more accuracy was logistic regression. Author believes that machine learning algorithms can be used to predict in-hospital mortality for TAVR better than previous tools which are being used in hospital.

Although machine learning is a new technique to find out weak students by analyzing their results and predicting their future marks and on the base of results, special attention can be given out to those students to help them but aside from machine learning, different methods are being used to solve this student's performance issue. In [24], author research on those factors [1], [2], [3] and applied problem solving strategies to involve and motivate students for study. According to the author, problem based learning can help students to think, analyze and solve a problem in different ways. Author created some groups of students and gave them some problems to solve. In that group, there were some senior students who were gone through same method before. Author played a role of teacher and he tried to expand student's student by giving them directions and new ways of thinking [28]. Students used different brain storming methods, self-

learning and problem analyzing techniques and they did research about the ways to solve similar type of problems on internet and successfully solved that problem. Author believes that in this way all the students (including smarts and weak students) can share their ways of solving problems and can learn a lot from each other. It will also motivate students to solve problems before other groups to get a reward if they are also competing with each other [26].

"Lisa Flook" and "Channa Cook-Harvey" did detailed research on how to solve student's problems and improve their performance. According to them [25], a student's life revolves around academic, cognitive, ethical physical, psychological and socio-emotional problems.

Figure 2: Principals of practices for the teachers [25]

In order to solve these problems, we need to improve and change supportive environment in which we need to promotes strong attachment, relationship, a sense of safety and belonging so that student could feel safe and confident around teachers and enjoy a safe environment around him. Teachers need to improve teaching strategies, connect students with each other through study groups and help them to develop cognitive abilities. Teachers should create a system of support around them so that in case of any problem, they do not hesitate to contact their teachers and share their problems with them with confidence. Teachers need to help them in social and

emotional development and help them in learning and exploring their skills and special gifts so that they can use them in their future to improve their lives [25].
George Polya, a famous mathematician wrote in this book "How to solve it" [27] that "teacher should put himself in the student's place, he should see the student's case, he should try to understand what is going on in the student's mind and ask a question or indicate a step that could have occurred to the student himself." According to the author of [30], students can improve their performance if teachers motivate them in group discussion and promote self-learning. They should at first propose an open ended problem to engage students and get their attention then they should ask students to self-study about it and discuss it among themselves to share ideas. At the end teacher can compare their ideas and tell them the solution or give them right direction and boost their motivation and confidence.
Although students face many problems during their educational period but teachers are also trying different methods to support them and motivate them. According to the above literature review, students are facing problems including poor family background, socio-economic status, poor motivation towards education, slow to learn, low level of parents' income, low parents' literacy, poor teacher' attitude towards their job, unqualified teacher, poor class management, poor class environment, Inadequate Teaching and learning material, low Student's academic background, lower type of institute and its facilities, adequate hostel facilities, involvement in too much sport activities, low Attendance and less time allocation for studies

2.4 Problems faced by Students:

Table 3: Problems faced by students

No.	Problems	References
1	Poor family background	[1], [2]
2	Socio-economic status	[1], [2]
3	Slow to learn	[1], [3]
4	Poor student's motivation towards study	[1], [2], [4]
5	Level of income of Parents	[1], [3]
6	Parents' literacy	[1], [3]
7	Poor Teacher's attitude	[1], [2]
8	Teacher's teaching methods	[1], [2]
9	Teacher's attitude towards their job	[1], [2]
10	Under Qualified Teacher	[1], [2], [3], [4]
11	Poor Classroom management	[1], [3]
12	Poor Learning Environment	[1], [3]
13	Distance from learning space	[32]
14	Inadequate Teaching and learning material	[2], [3]
15	Student's academic background	[2], [4]
16	Type of institute and its facilities	[4], [1], [3]
17	Adequate hostel facilities	[4]
18	Involvement in too much sport activities	[4], [31]
19	Low Attendance	[31], [3], [1]
20	Time allocation for studies	[31], [3]

2.5 Method used by Teachers to solve problems:

Table 4: Method used by teachers to solve student's problems

No.	Methods	Papers
1	Problem Solving Based Strategies	[24], [26], [28]
2	Create Supportive Environment	[25], [27]
3	Create Study Groups Among Students	[24], [28]
4	Give Extra Time and Attention To Weak Students	[24], [28], [8], [11], [13]
5	Help To Develop Communication Among Students	[25], [27]
6	Encourage Students To Be Independent	[25], [29]
7	Teacher Puts Himself In His Student's Place To Understand Problem Through Student's Perspective	[27]
8	Machine Learning to learn about student's problems	[8], [10], [11], [12], [13], [17]
9	Motivate students to self-learn and group discussion	[30]

2.6 Machine Learning Models used to solve problems:

Table 5: Machine learning models

No.	Models	Papers
1	Support Vector Machine	[6],[8],[10,[14],[15],[18],[19],[22]
2	Neural Network	[5],[9],[10],[13],[16],[19],[21]
3	Naive Bayes	[3],[8],[9],[10],[15],[16],[17],[18],[20],[21],[22]
4	K-Nearest Neighbors	[4],[10],[15],[22]
5	Random Forest	[6],[8],[10],[13],[16],[20],[21],[22]
6	Linear Regression	[7],[10],[12],[14]
7	Logistic Regression	[10],[15],[21],[22]
8	Decision Tree	[7],[8],[9],[10],[14],[16],[17],[18],[19],[20],[22]
9	AdaBoost	[11]
10	Quadratic Discriminant Analysis	[12]
11	Multi-Layer Perceptron	[18],[20]

2.7 Machine Learning Models:

2.7.1 Support Vector Machine:

Support vector machine is a division of supervised learning algorithm. It is a powerful method to perform regression, classification and outlier detection of data. The original inventor of SVM is Vapnik and Chervonenkis. They proposed SVM hyperplane with linear classifier. Vapnik, Boser and Guyon proposed SVM hyper plane with non-linear classifier using the kernel concept. It can perform Support Vector Clustering for the unsupervised learning algorithm. The main objective is to find a hyper plane that best divides the two classes. Based on this hyper plane, the new data is best classified to which class it belongs to [32]. Such a hyper plane is illustrated in Figure 2.

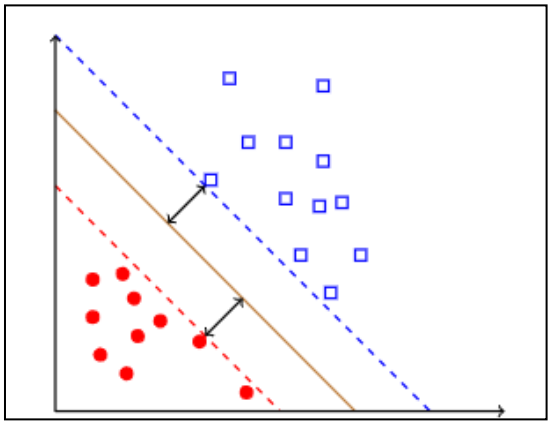

Figure 3: Maximizing the margin in support vector machine

2.7.1.1 Advantages

- Support vector machine is good for high dimentional data
- If number of samples are lower than number of dimention, it will be a good choice
- By using kernel trick, it can support classification of non-linear data
- In case of prediction problems, it is proved to be a robust classifier

2.7.1.2 Limitations

- In case of wrong kernel selection, error percentage will increase
- Very slow during test phase
- Due to quadratic programming, it need high memory space
- Due to its high training time, not recommended for high and noisy data
- In order to calculate probability measures, expensive cross validation is needed

2.7.1.3 Real Time Applications

- Image classification and detection of faces from imges or videos
- Recognition of hand written words and categorization of text and hyper text
- Use of SVM in Bioinformatics field
- Protein fold and Remote homology detection
- Environmental and Geo sciences based applications

2.7.2 Neural Network:

Artificial neural network works like a human brain and it also process information just like human brain in the form of neurons. In order to learn and predict something, artificial neural network gets information from data set and tries to learn its patterns and relationship between different entities in it. According to [33], "An ANN is formed from hundreds of single units, artificial neurons or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are organized in layers." In artificial neural network, there are three types of layers i.e. input layer, middle layer (consists of many layers because they are used

to balance the output layer and responsible for main calculations) and output layer. Each layer consists of many processing elements which have also weighted inputs with them [33][34].

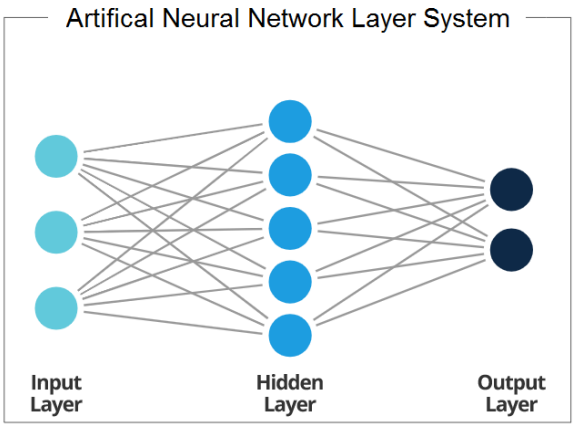

Figure 4: Artificial Neural Network Layer System [34]

There are two main types of neural networks

- Feed-forward NN
- Feed-back NN

2.7.2.1 Feed-forward networks

Feed-forward Artificial Neural Networks (ANN) moves in a forward direction from input towards output, it is also called a bottom-up approach. The data is given at the input which process to the hidden layer and to the output layer as Shown in Figure 7. Feed-forward ANNs (Figure 7) allow signals to travel one way only; from input to output. They are not considered good for time series problem n[34].

2.7.2.2 Feedback networks

In feedback neural networks in the input is given at the output. It is also called out-star/fan-out. The data is given at the output with is then processed in the hidden layer and move towards the inputs. It has a very useful utility in recurrent neural networks, when the network is moving in both directions to get the best solution. Learning laws are Hebbian law and Delta law [34].

2.7.2.3 Advantages:

- Requires less statistical training
- It has ability to detect non-linear relationship between independent and dependent variables
- It has ability to detect all possible interaction between predictor variables

2.7.2.4 Limitation:

- It has black box type of nature

- Greater computational burden
- Proneness to over fitting
- It has empirical nature of model development

2.7.2.5 Real Time Application:

- Medical Diagnosis
- Credit Rating
- Targeted Marketing
- Voice Recognition
- Financial Forecasting
- Fraud Detection

2.7.3 Naive Bayes:

Normally for classification tasks in machine learning, we use Naive Bayes algorithm. It does not matter how much records a database has or either it has binary and multi class classification problem or not, Naive bayes can easily handle it. Natural language processing and text analysis are the main application of Naive bayes. In order to understand naive bayes algorithm efficiently, we need to understand bayes theorem because it combines multiple classification algorithms in order to form a naive bayes classifier which is based on conditional probaability which means that events will occur one by one based on those events which are already occured.

Formula:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Where,

P(A) = Prior probability of an event A.
P(A|B) is a conditional probability of an event A with conditioned an event B.
P(B) = Prior probability of an event B.

Types:

1) Bernoulli
2) Multinomial
3) Gaussian

Bernoulli works well when values are binary-values in database whereas Multinomial naive bayes works well when valus are dispersed multi-nominal. Gussain works well when all the values in database are continuous.

2.7.3.1 Advantages

- Simple, Quicker and Accessible.
- It can easily handle continuous, binary and multinomial distributed attribute values
- It is best choice for classification realted problems.

- Very simple to build for small and big data set.
- It is not sensitive for irrelevant attributes.

2.7.3.2 Limitations

- The biggest drawback of this algorithm is that it can not find relationship between attributes.
- If the attribute class has zero frequency data items then there is a possibility to occur "Zero Conditional Probability Problem".
- Not useful for regression problems
- Naïve bayes assume high independence of attribute vairbles which are not possible in rea life.

2.7.3.3 Real Time Applications:

- Natural language processing based applications.
- Symbols, names, emails and text classification
- Recommendation Systems.
- Sentiment Analysis.
- Text analysis based applications like Udictionary etc.

2.7.4 K-Nearest Neighbors:

K-Nearest Neighbors is another famous model used for classification problem. In order to find target label of data, it calculates the ditance between nearest class labels with new data point with the help of k value.It then count the numbers of closest points and based on those points, it predicts the label of target data. The value of K always varry between 0 and 1 normally. The most popular distance function used by KNN are Manhattan distance, Euclidean distance, Hamming distance and Minkowski distance [32].

2.7.4.1 Advantages

- In case of large training dataset, it will be a good and strong option
- For distance functions and attributes, it is very simple and flexible
- In case of multi class data set, it will support it

2.7.4.2 Limitations

- Finding appropriate K value is a problematic task.
- Computation cost is very high.
- For any specific dataset, it is difficult to choose suitable distance function.
- Change in K value is directly proportional to change in target class label.
- It needs large storage space as well as large sample for reliable accuracy.

2.7.4.3 Real Time Applications

- Credit ratings and Finance
- Agriculture and medicine
- Text and video recognition
- Image and handwriting recognition.

2.7.5 Random Forest:

Random forest is an ensemble classifier, which constructs a group of independent and non-identical decision trees based on the idea of randomization. Random forest can be defined as

$$\{h(x,\theta_k), k=1,\cdots,L\}$$

In which theta (k) is a kind of mutual independent random vector parameter, and x is the input data. Each decision tree uses a random vector as a parameter, randomly selects the feature of samples, and randomly selects the subset of the sample data set as the training set. The construction algorithm of random forest is as follows [35].

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Where,

- N is the number of data points
- Fi is the value returned by the model
- Yi is the actual value for data point i

2.7.5.1 Advantages

- Random forest can easily overcome the problem of over fitting
- Less sensitive to outlier data during data training
- Very easy to set the parameters in it
- Accuracy and variable importance are generated automatically
- Accurate predictions results for a variety of applications
- Through model training, the importance of each feature can be measured
- Trained model can measure the pair-wise proximity between the samples

2.7.5.2 Limitations:

- For regression problem, it does not gives precise continuous nature prediction
- In case of regression, it doesn't predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy.
- Random forest can feel like a black box approach as we have very little control on what the model does.

2.7.5.3 Real Time Applications:

- Urban Planning in a Visual Semantic Decision Support System by using real world city data.
- Video classification
- Internet traffic interception
- Image classification
- Voice classification

2.7.6 Linear Regression:

In order to predict the future, only one independent variable is being used by simple regression and two or more independent varibales are being used by multiple regression.Dependent and independent varibales have also different values i.e. continuous values for dependent variables and discrete values for independent variables. Regression models have two kinds, linear and non linear models. Normally linear regression models utalize stright line relationship between dependent and independent variables while curved line relationship will be used by non linear regression model.

2.7.6.1 Advantages

- Linear regression is good to find relationship with independent and dependent variables with reliable results
- Linear regression is very easy to use
- It is very simple to train data and predict results

2.7.6.2 Limitations

- Only good for numeric output.
- Not suitable for datasets containing nonlinear data .
- It is sensitive with outliers.
- It need independent data to work effectivily.

2.7.6.3 Real Time Applications:

- Studying engine performance from test data in automobiles.
- OLS regression can be used in weather data analysis.
- Linear regression can be used in market research studies and customer survey results analysis.
- Linear regression is used in observational astronomy commonly enough.
- Predictive Analytics and Trend Lines.
- Epidemiology and Finance.

2.7.7 Logistic Regression:

Logistic regression is normally used for those classification problems and based on the concept of probability, it is also called predictive analysis algorithm. Basically logistic regression is similar to linear regression but it has more complex cost function which is also called "sigmoid function" or "logistic function" [36]. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

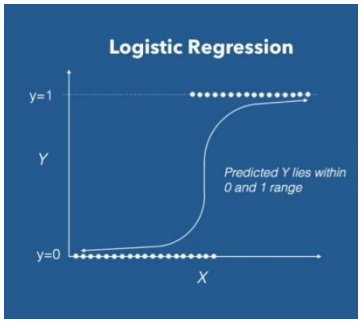

Figure 5: Logistic Regression [36]

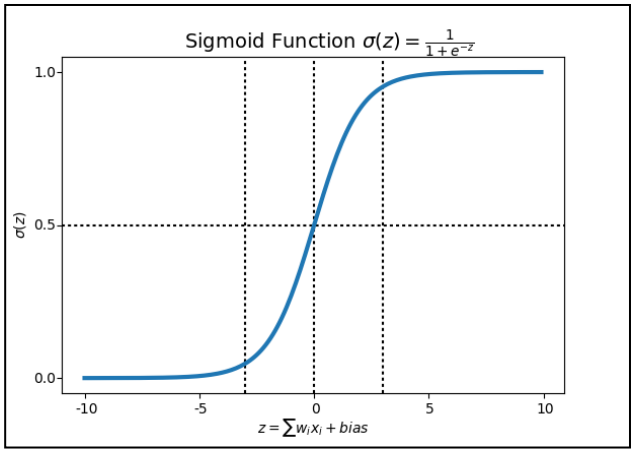

Figure 6: Sigmoid function mapping real values with predictive values

After getting predicted values, we need to map them with probabilities so that we can know about the difference between them and for that purpose we use sigmoid function. Sigmoid function converts real values in to form of 0 to 1 and then map them into predicted values. It will look like following;

Sigmoid function will be presented mathematically as;

$$f(x) = \frac{1}{1+e^{-(x)}}$$

When using *linear regression* we used a formula of the hypothesis i.e.

$$h\Theta(x) = \beta o + \beta 1X$$

For logistic regression we are going to modify it a little bit i.e.

$$\sigma(Z) = \sigma(\beta o + \beta 1X)$$

We have expected that our hypothesis will give values between 0 and 1.

$$Z = \beta_0 + \beta_1 X$$

$$h\Theta(x) = \text{sigmoid}(Z)$$

$$\text{i.e. } h\Theta(x) = 1/(1 + e\text{^-}(\beta_0 + \beta_1 X)$$

The hypothesis of logistic regression will be calculated as;

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Decision Boundary: The logistic regression classifier should give us the probability score between 0 and 1 when we pass input values through its prediction function. With the help of cost function, we can improve the accuracy of our result and reduce the error rate to minimum.

If we will try to implement cost function of linear regression, it will be very difficult to minimize the cost value because it would end up being non-convex function.
So, we will use a very special cost function for logistic regression which is defined as;

$$-log(h\theta(x)) \text{ if } y = 1$$

$$-log(1-h\theta(x)) \text{ if } y = 0$$

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Figure 7: Logistic Regression Cost Function

With the help of gradient descent, we can successfully reduce the cost value in logistic regression function. For this purpose we need to run gradient descent on each parameter. Gradient decent is as follow;

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2.$$

And we will repeat following gradient decent function on each parameter to minimize cost value.

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

(simultaneously update all θ_j)

$\}$

Gradient descent has an analogy in which we have to imagine ourselves at the top of a mountain valley and left stranded and blindfolded, our objective is to reach the bottom of the hill. Feeling the slope of the terrain around you is what everyone would do. Well, this action is analogous to calculating the gradient descent, and taking a step is analogous to one iteration of the update to the parameters.

2.7.7.1 Advantages:

- When dataset is linearly separable, logistic regression performs really well.
- Less prone to over-fitting
- It not only gives measure of how relevant a predictor but also its association's direction (negative or positive)
- Easier to implement
- Very efficient to train

2.7.7.2 Limitations:

- Can be over-fit in high dimensional datasets
- Features should always be greater than number of observations otherwise over-fit problem will appear
- Assumption of linearity between independent variables and dependent variables
- Can only be used to predict discrete functions

2.7.7.3 Real Time Applications:

- Handwriting Recognition
- Image Segmentation
- Geographic Image Processing
- Predication of a person's emotion
- Healthcare

2.7.8 Decision Tree:

Decision tree algorithm mainly used to construct a training/classification/regression model in the form of a tree structure (root, branch and leaf), which is based on (inferred from) previous data to classify/predict class or target variables of future/new data with the help of decision rules or decision trees. As a supervised learning algorithm, decision trees can be used for both numerical and categorical data. ID3 (Iterative Dichotomiser3) act as a base algorithm for construction of decision trees based on Entropy and Information Gain and CART (Classification and Regression Trees) algorithm is based on Gini Index as a metric. It uses greedy search methodology from top to bottom without backtracking. In a complete decision tree, the root node in each level is a starting point or the best splitting attribute in that position which helps to test on an attribute. The output of the test will produce branches. Leaf node will be acting as a final class label or target variable to classify/predict the new data [32].

2.7.8.1 Advantages

- It is easy to implement and interpret.
- It can classify/predict categorical/numerical data
- It takes less data preprocessing.
- Statistical test helps to validate the decision tree model.
- It resembles the human decision making methodology
- The complex tree structure will easily understand by visualization

2.7.8.2 Limitations

- High Probability of over fitting in the decision tree
- Prediction accuracy is low compared to other machine learning algorithms.
- If the class labels are huge, then the calculation may leads complexity.
- Prediction result will not be good, if evaluation data and sampled training data are different.
- Outliers may produce sampling errors.
- Small change in the data set will lead different tree structure, so it is not a stable one.
- Redrawing is needed for every addition of information to the data set.

2.7.8.3 Real Time Applications

- Agriculture, Astronomy
- Biomedical Engineering, Control Systems
- Financial analysis, Manufacturing
- Medicine, Molecular biology
- Object recognition, Pharmacology
- Software debugging, Text processing

2.7.9 AdaBoost:

The full form of adaboost is "Adaptive Boosting" and it was proposed by "Freund" and "Schapire" in 1996. The main purpose of adaboost machine learning classifier is to solve classification problem and it can easily convert weak classifiers into strong ones. It can be represented as,

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)),$$

Here, f_m represent the weakest classifier and theta(m) represent corresponding weight [37]. The full adaboost algorithm process is explain under,
Let's suppose there is a dataset with n points,

$$x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}.$$

In above equation, 1 represent positivity and -1 represent negativity. Now let's initialize weight for each data point,

$$w(x_i, y_i) = \frac{1}{n}, i = 1, ..., n.$$

Step 1: At first, we need to fit weak classifiers into the dataset and after that we will select that one, which has lowest weighted classification error,

$$\epsilon_m = E_{w_m}[1_{y \neq f(x)}]$$

Step 2: Now calculate the weight for weak classifier

$$\theta_m = \frac{1}{2} ln(\frac{1 - \epsilon_m}{\epsilon_m}).$$

If weight is positive then it means that classifier has higher accuracy than 50%. It shows that higher the accuracy of a classifier, higher the weights will be and if the weight is negative it means that classifier has less accuracy than 50% which shows that we can combine its prediction by flipping the signs [37]. In short, we can easily convert 40% accuracy of a classifier into 60% by only flipping the sign of prediction. Due to random guessing, even a good classifier will guess wrong but still it will contribute in result prediction at the end.
Step 3: Now in order to improve accuracy, we will update weights of each data points,

$$w_{m+1}(x_i, y_i) = \frac{w_m(x_i, y_i) exp[-\theta_m y_i f_m(x_i)]}{Z_m},$$

In above equation, Z_m is normalization factor and it make sure that sum of weights is equal to 1. If a misclassified case is from a positive weighted classifier, the "exp" term in the numerator would be always larger than 1 (y*f is always -1, theta_m is positive). Thus misclassified cases would be updated with larger weights after iteration. The same logic applies to the negative weighted classifiers. The only difference is that the original correct classifications would become misclassifications after flipping the sign.
After M iteration we can get the final prediction by summing up the weighted prediction of each classifier [37].

2.7.9.1 Advantages:

- Easier to use with less need for tweaking parameters
- Can be used with Support vector machine algorithm to improve accuracy
- Not prone to over-fitting because parameters are not jointly optimized
- Can be used to improve accuracy of other classifiers making it flexible
- Less error based on ensemble method

2.7.9.2 Limitation:

- It needs quality data in order to learn progressively
- Extremely sensitive to noisy data and outliers
- Time and computation expensive
- Hard to implement in real time platform
- Complexity of the classification increases

2.7.9.3 Real Time Applications:

- Prediction of customer churn
- Use to solve classification problems
- Person recognition
- Medicine recognition

2.7.10 Quadratic Discriminant Analysis:

Quadratic discriminant analysis is very similar to linear discriminant analysis when normally distributed measurements are required. However there is a small difference between them which is that in QDA, there is no assumption that the covariance of each of the classes is identical. As compare to linear discriminant analysis, more computational and data is required in order to estimate those parameters which are required in quadratic discriminant analysis. In case of no difference in group covariance matrices, both quadratic discriminant analysis and linear discriminant analysis can perform well [38]. We can also call quadratic discriminant analysis a general form of Bayesian discrimination.

Normally quadratic discriminant analysis is used to find out which variables discriminate between two naturally occurring groups. Let's suppose, a student wants to find out which variables discriminate between high school graduates who decide (1) to go for more higher education, (2) Not go for higher education. In order to find out about it, student needs to collect data from a lot of high school graduates because he knows that students will choose one of the above options after their graduation. Student can then use quadratic discriminant analysis to find out about those variables which can help to predict the best results. In short QDA is very similar to ANOVA (Analysis of Variance)[38].
Discriminant Analysis may be used for two objectives: either we want to assess the adequacy of classification, given the group memberships of the objects under study; or we wish to assign objects to one of a number of (known) groups of objects. Discriminant Analysis may thus have a descriptive or a predictive objective. In both cases, some group assignments must be known before carrying out the Discriminant Analysis. Such group assignments, or labeling, may be arrived at in any way [38]. Hence Discriminant Analysis can be employed as a useful complement to Cluster Analysis (in order to judge the results of the latter) or Principal Components Analysis.

2.7.10.1 Advantages:

- Fast Classification
- More accuracy
- Outperform K-near neighbor and Linear discriminant analysis
- Quadratic decision boundary

2.7.10.2 Limitations:

- Complex matrix ops
- High training time
- Time and computation expensive
- Only based on Gaussian distribution

2.7.10.3 Real Time Applications:

- Classification Problems like cat and dog
- Student performance analysis
- Teacher performance analysis
- Recommendation systems

2.8 Research Questions:

This section described the research questions made for this systematic literature review. The following are the research questions.

Research Question 1: What are the main issues and problems being faced by the weak students in colleges/universities?

Research Question 2: Which methods are being used by institutes to solve student's problems related to their education?

Research Question 3: Which important machine learning models have been used by researchers from 2012 to 2019 in their studies?

Research Question 4: Which machine learning model is best to predict student's performance on the base of accuracy?

Research Question 5: How machine learning algorithms can solve student's performance problem and what are its advantages?

2.9 Answers to research questions:

This section answers all the research questions. Based on these research questions answers we conducted a second literature review only on blood glucose level prediction. The following are answers to the research questions.

Answer of Research Question 1: In Table 3, problems being faced by students have been described and we have taught about them in detail in above literature review.

Answer of Research Question 2: Methods which are mentioned in Table 4 have been used to help students to solve their problems by the teachers.

Answer of Research Question 3: With the help of detailed literature review, we have identified latest machine learning algorithms used by many researchers in their research and we have described them in Table 5.

Answer of Research Question 4: According to the comparison between all machine learning model's results in table 33 and 34, Decision Tree and Random forest are the best machine learning models to deal with student's performance issue. Neural Network is also good but it accuracy was less than decision tree and random forest.

Answer of Research Question 5: From results, applications and future work, it can be easily observed that with the help of machine learning, we can revolutionize education by predicting student's performance earlier and taking some reasonable steps to help them. Furthermore application and future work section can better answer this question.

2.10 Summary:

In this chapter, we have done a detailed literature review by selecting papers according to specified criteria. We have identified those problems which mostly students faced during their education phase. In order to solve those problems and help students, some teachers use special methods which also have been identified in literature review. Further we have identified which popular machine learning algorithms have been used by the researcher from the years 2011 to 2019 so that we can apply them on a public dataset of students to identify weak students so that significant steps can be taken by teachers to help those students.

This page is intentionally left blank

Chapter 3

METHODOLOGY

3.1 Introduction:

Research methodology is the systematic, theoretical analysis of the procedures applied to a field of study. Methodology involves procedures of describing, explaining and predicting phenomena so as to solve a problem; it is the 'how'; the process, or techniques of conducting research.

(Kothari, 2004)

In this chapter, we will discuss how we are going to analyze student's progress and how we are going to predict their future performance. We will discuss about properties of student's database on which we will apply machine learning algorithms which we have gotten from our research. We will discuss complete process from data cleaning to prediction of performance step by step.

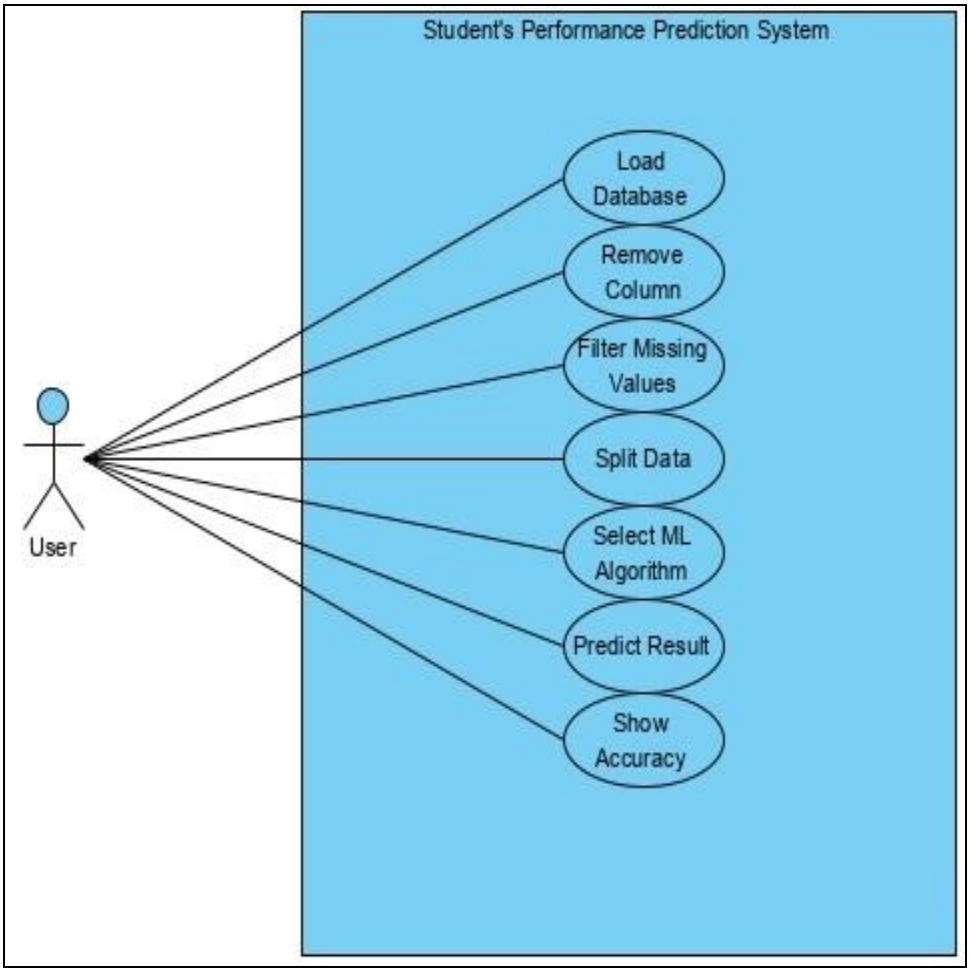

Figure 8: Use Case Diagram of Requirements

3.2 Methodology Requirements:

The methodology of this research consists of many requirements which should be understood in order to move forward. We will explain and discuss each requirement here so that it will be easy to understand step by step procedure later in the chapter.

3.2.1 Databases:

In order to implement our methodology, we need student's database because machine learning algorithms will train themselves by using databases and then we can use them

to predict student's performance. Database is the most important part of machine learning modeling [39] because the accuracy of machine learning models depends on the size and quality of database which has been used during training process. If there are missing values in database, its size is small and there are a lot of garbage values in it, the accuracy will be close to nothing because machine learning algorithm will use that database to train them and they will predict results according to their learning [39].

3.2.2 Remove Columns:

Every database consists of a lot of garbage (unimportant) data (features) which are not useful during machine learning training. Only those features are recommended for machine learning which has some effect on output or result which we desire. So we will remove all those features from database which are not necessary for machine learning training so that proper machine learning algorithm can be trained on it [39].

3.2.3 Filter Missing Values:

Database is the most important thing for any machine learning algorithm because without database, it is impossible to train machine learning algorithm. Every database consists of rows and columns. These rows and columns consist of records which sometimes can have null or empty values which can disturb accuracy of machine learning algorithm after training so in order to remove this fault; we add some values and fill those null and empty records.

3.2.4 Split Database:

Mostly machine learning models fail because the testing data which was used to test that machine learning model was totally different from that data which was used to train it. Although we will recommend that model which has been trained with database which has sample of every type of data so that in case of any unusual database, it should handle it pretty well and present good accuracy [40]. There are so many techniques which can be used to counter this difficulty somehow. One of them is splitting your database into some ratio i.e. 70-to-30 ratio in which 70% database can used to train machine learning model and 30% database can be used to test the accuracy of model. Or we can also use 100% database for training purpose and we can take some value or record from user to test that model which has been trained earlier [40].

3.2.5 Selecting Machine Learning Model:

Selecting a most suitable machine learning model is the second most important part in machine learning model training and prediction. There are three types of machine learning models and they are used for different purposes.

3.2.5.1 Types of Machine Learning Algorithms:

There are three types of machine learning algorithms supervised machine learning algorithms, unsupervised machine learning algorithms and reinforcement machine learning algorithms. Let's discuss them in detail one by one.

3.2.5.2 Supervised Machine Learning:

Supervised machine learning algorithms are used for those databases which have labels with them. Common supervised machine learning algorithms include decision tree, support vector machine, random forest and neural network etc.

3.2.5.3 Unsupervised Machine Learning:

Unsupervised machine learning algorithms are used for those databases which have no labels with them. Common unsupervised machine learning algorithms include k-mean clustering, apriori, mean shift and dimension reduction etc.

3.2.5.4 Reinforcement Machine Learning:

Reinforcement learning is based on punish and reward systems. It is basically used in games where on any achievement, character gets rewards in the form of points, XP or lives whereas on failed mission, character gets punishment in the form of loss of health or life.

3.2.6 Predicting Result:

After selecting a suitable machine learning model, it is time to train it by using database which was selected earlier. Machine learning model will use this database and train itself. This process will take some time as it depends on the quantity of database (numbers of rows and columns).

3.2.7 Calculating Accuracy:

Accuracy is the final and third most important thing in any machine learning model based project because we are doing all the hard work just to predict some result whose accuracy should be more than 80% at least. Here one thing is important to know that if database's size is large, its accuracy will also be good [40] because model's result depends on the quantity and quality of database. The meaning of quality is that database should have less number of faulty records, null values and it has more meaningful data which can help in model training and predicting useful result. The meaning of quantity is that database should have at least 10,000 records so that machine learning model can train itself properly [40]. Remember that more data means more proper training and more accuracy.

3.2.8 Data Visualization:

Data visualization means comparing predicted results with actual results and shows the difference between them on graph. This is the final step in any machine learning model based project which can help to understand the accuracy of predicted results and difference between actual and predicted results more effectively.

3.3 Flow Chart of Whole Process:

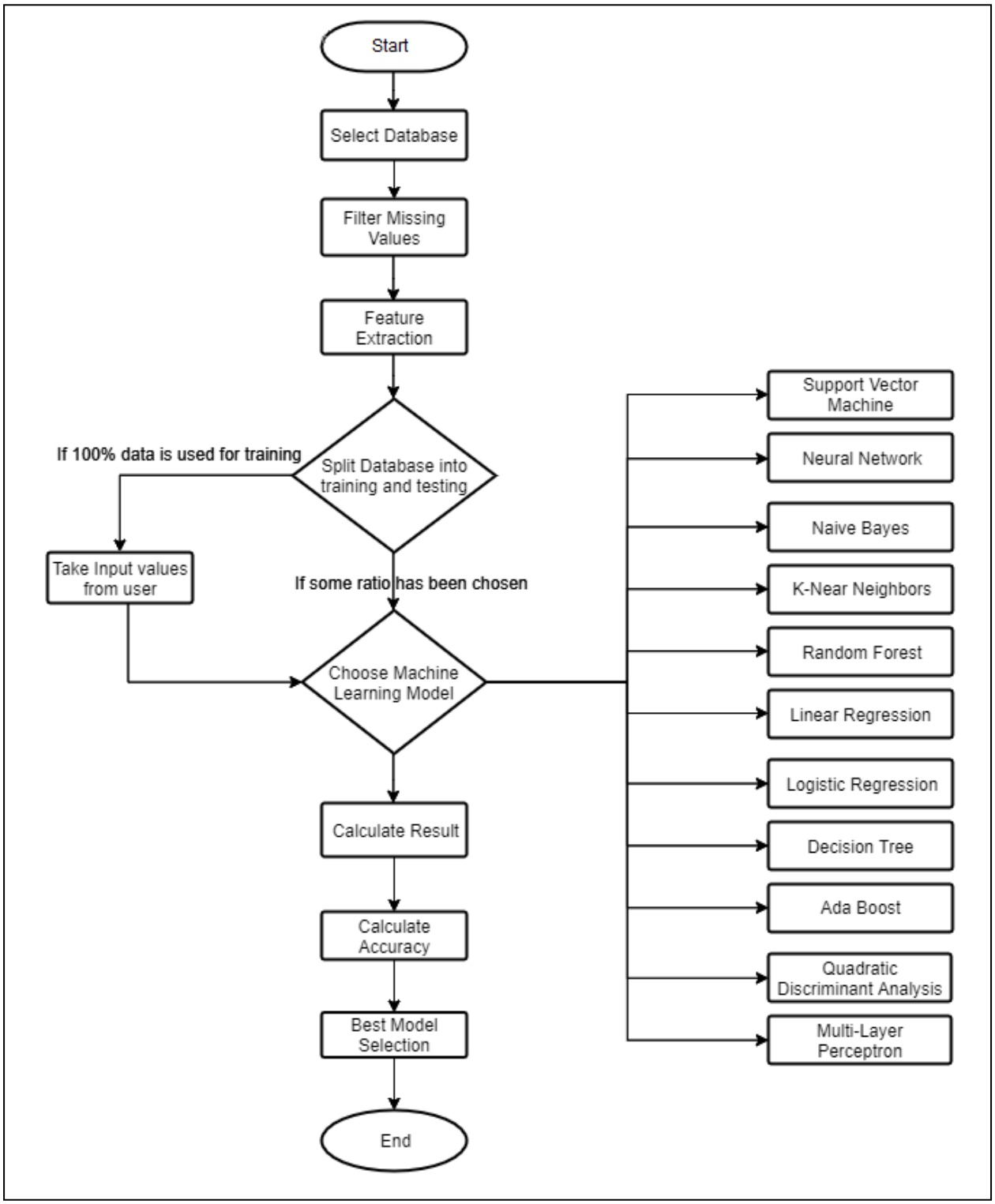

Figure 9: Flow Chart of Complete Project

3.3.1 Explanation of the process of user, using the Software:

Step 1: User will load all the databases which are available but user has to select that database which is related to its problem. For example, for "Student Progress Analysis and Prediction" problem, user needs to load databases which have data about student's performance in it. So user will load all related database and select one of them which is more relevant.

Step 2: After loading database, user need to filter all the missing and null values to improve database's quality because null values will affect accuracy of model after training. User can ask to put 1 or 0 on all the null record's location.

Step 3: User will ask to show all database's columns and their records so that relevant columns can be selected and remaining columns could be removed in order to make database more relevant for machine learning model. User will choose one column as an output column and remove remaining irrelevant columns from the database. All the remaining columns will act as input for machine learning model.

Step 4: Now user has to split database into training and testing part. There will be some options like 70/30, 80/20, 90/10 or user can even select 100% database for training purpose. If user selects any ratio other than 100%, then software will split database according to that ratio. Training database will be used for training purpose and testing database will be used to test that model's accuracy. In case user select 100% ratio to train model, software will ask to give input value to test model's accuracy after training and user has to enter similar data which is in training database.

Step 5: After selecting ratio, software will ask user to select one machine learning model (There are 11 machine learning model which researchers have used to solve their machine learning related problems in the past and which we have detected from research papers during our literature review process). User will select one machine learning model and software will start training that machine learning model by using training database. It will take some time to train model as it depends on the size of database.

Step 6: After some time, machine learning model's training will complete and user will select show result option, software will show the result in the form of 0 and 1. If the result is "zero" then it means student might get fail in future and if the result is "one" then it means student will get success in future.

Step 7: User will ask to show accuracy of the predicted result. If user has selected splitting ratio in the beginning of the training then software will compare predicted results with actual results and based on comparison, it will show accuracy. In case user has selected 100% data for training purpose then there will be no accuracy prediction because there is nothing to compare our result.

Step 8: Software will show the comparison's result on graph to help user to understand the difference more clearly and easily.

3.4 Sequence diagram about how user will use the system:

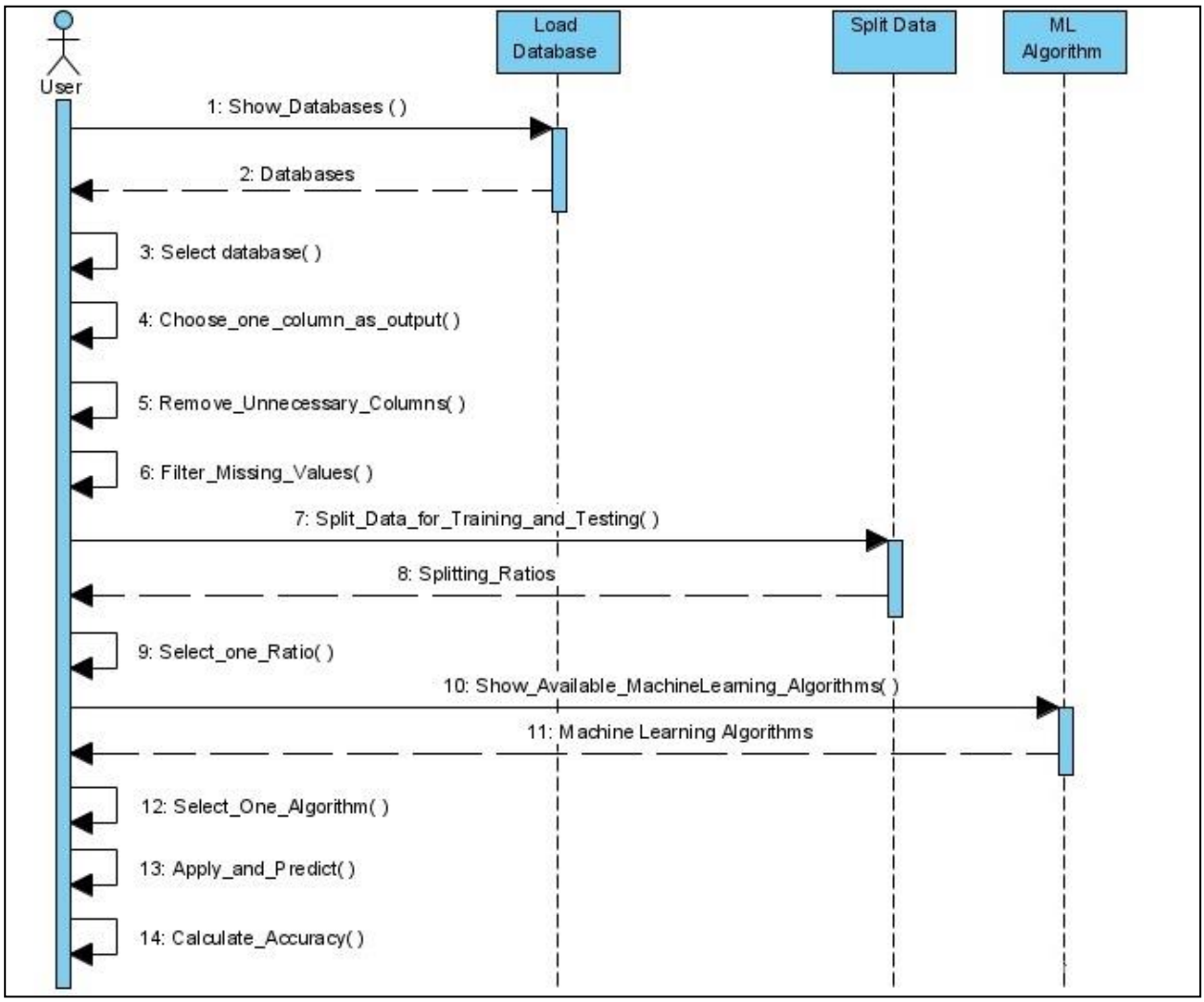

Figure 10: Sequence Diagram of the Complete Process of User Using the Software

3.5 State Machine Diagram:

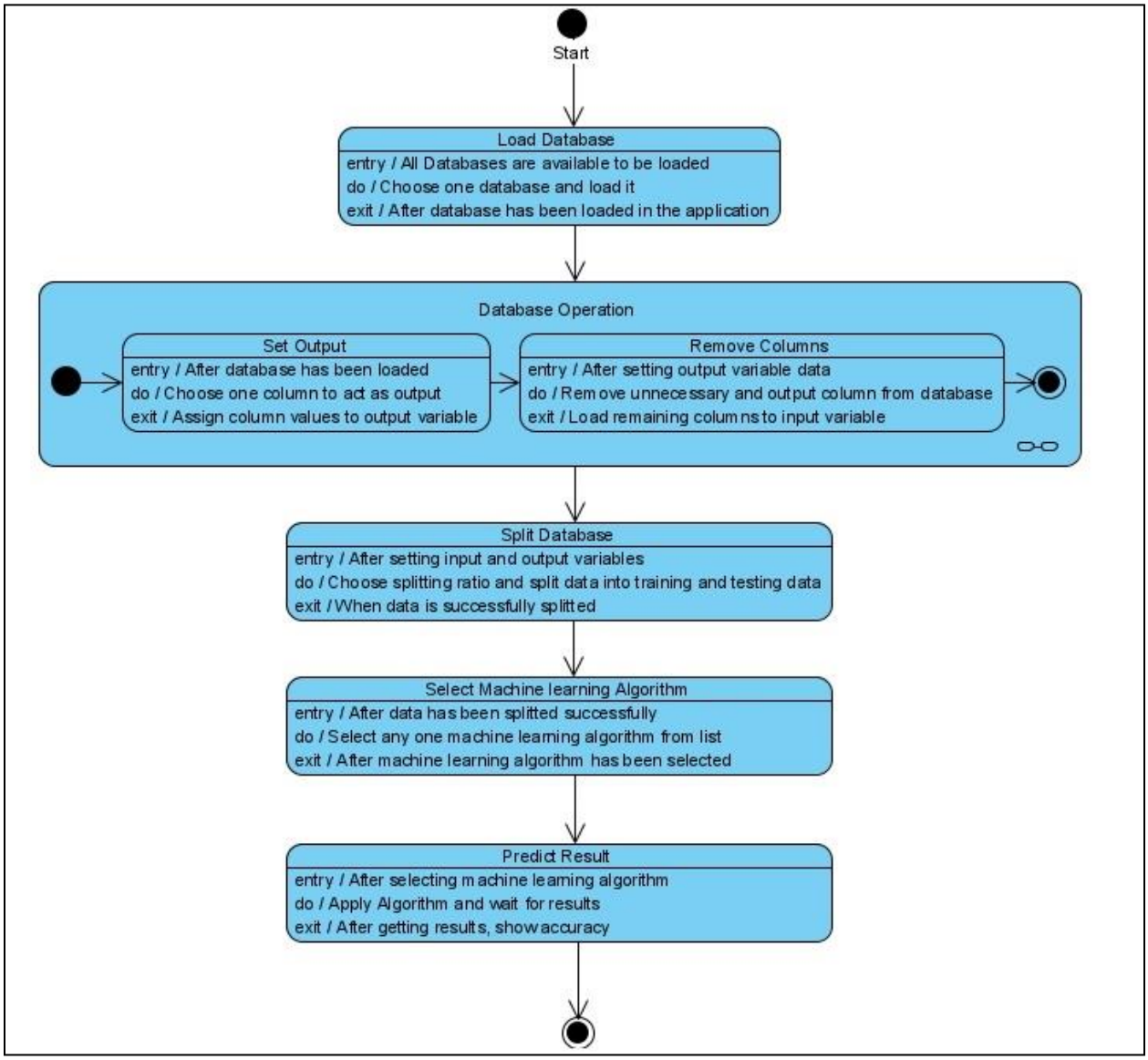

Figure 11: State Machine diagram of the Whole Machine Learning Process

3.6 Dataset Used

3.6.1 Dataset Number 1:

In order to evaluate student's performance, we need student's database which consists of student's marks, its attendance and other related features. The first public database was taken from "KAGGLE WEBSITE" [41] which contains data of 480 students. This database has 17 attributes including "Gender", "Nationality", "Stage ID", "Grade ID", "Place of birth", "Section ID", "Topic", "Semester", "Relation with Parents", "Raised hands", "Visited Resources", "Announcement View", "Discussion", "Parent Answering Survey", "Parent School Satisfaction", "Class" and "Student Absence Days". In this database, student's success level is divided into three categories which are "M" (Middle), "H" (High) and "L" (Low). Categories "M" and "H" are considered successful in the exam whereas category "L" is considered fail in exam [41].

Table 6: 1st Dataset's Columns' Description [41]

Column	Description	Type
Gender	Gender of student	Nominal
Nationality	Nationality of student	Nominal
PlaceofBirth	Country of birth for student	Nominal
StageID	Educational stage, for example Middle school, high school	Nominal
GradeID	Grade level of the student	Nominal
SectionID	Classroom of the student	Nominal
Topic	Course topic	Nominal
Semester	Semester of the year	Nominal
Relation	Parent responsible for the student	Nominal
Raisedhands	Number of times the student raised hands during the class	Quantitative
VisitedResources	Number of times the student visited the course content	Quantitative
AnnouncementsView	Number of times the student checked new announcements	Quantitative
Discussion	Number of times the student joined the discussion groups	Quantitative
ParentAnsweringSurvey	Did the parent answer the school surveys	Nominal
ParentschoolSatisfaction	Parents level of satisfaction for the school	Nominal
StudentAbsenceDays	Number of days the student has been absent	Quantitative
Class	Grade of student for the course	Quantitative

Table 7: A Sample of Data from 1st Dataset [41]

Gender	Nationalit	Place of B	StageID	GradeID	SectionID	Topic	Semester	Relation	Raised ha	Visited Re	Announce	Discussior	ParentAn	Parentsch	StudentAl	Class
M	KW	KuwaIT	lowerleve	G-04	A	IT	F	Father	15	16	2	20	1	Good	0	1
M	KW	KuwaIT	lowerleve	G-04	A	IT	F	Father	20	20	3	25	1	Good	0	1
M	KW	KuwaIT	lowerleve	G-04	A	IT	F	Father	10	7	0	30	0	Bad	1	0
M	KW	KuwaIT	lowerleve	G-04	A	IT	F	Father	30	25	5	35	0	Bad	1	0
M	KW	KuwaIT	lowerleve	G-04	A	IT	F	Father	40	50	12	50	0	Bad	1	1
F	KW	KuwaIT	lowerleve	G-04	A	IT	F	Father	42	30	13	70	1	Bad	1	1
M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	35	12	0	17	0	Bad	1	0
M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	50	10	15	22	1	Good	0	1
F	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	12	21	16	50	1	Good	0	1
F	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	70	80	25	70	1	Good	0	1
M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	50	88	30	80	1	Good	0	1

3.6.2 Dataset Number 2:

The second public database was also taken from "KAGGLE WEBSITE" [42] which contains 395 students. This database has 33 attributes including "School", "Sex", "Age", "Address", "Family Size", "Parent Status", "Mother Education", "Father Education", "Mother Job", "Father Job", "Reason to choose school", "Guardian", "Travel Time", "Study Time", "Failures", "School Extra Education Support", "Family Support", "Extra Paid Classes", "Extra-Curricular Activities", "Nursery", "Higher Education", "Internet Access", "Romantic Relationship", "Family Relationship", "Free Time", "Going Out", "Health", "Absences", "Daytime Alcohol Consumption", "Weekend Alcohol Consumption", "First Period Grade", "Second Period Grade" and "Final Grades".

Table 8: 2nd Dataset's Columns' Description [42]

Column	Description	Type
School	Name of student's school	Nominal
Sex	Gender of student	Nominal
Age	Age of student	Quantitative
Address	Whether the student lives in urban or rural area	Nominal
Famsize	Student's family size	Nominal
Pstatus	Whether the parents are living together or apart	Nominal
Medu	Mother's education	Quantitative
Fedu	Father's education	Quantitative
Mjob	Mother's job	Nominal
Fjob	Father's job	Nominal
Reason	Reason to choose the school	Nominal
Guardian	Student's guardian	Nominal
Traveltime	Travel time between home and school	Quantitative
Studytime	Study time in a week	Quantitative
Failures	Number of times student failed in past	Quantitative
Schoolsup	Educational support from school	Nominal
Famsup	Educational support from family	Nominal
Paid	Extra paid classes	Nominal

Activites	Extra activities	Nominal
Nursery	Attended nursery school	Nominal
Higher	If the student wants to pursue higher education	Nominal
Internet	If the student has internet at home	Nominal
Romantic	Does the student have a relationship	Nominal
Famrel	Family relations quality	Quantitative
Freetime	Student's amount of free time	Quantitative
Goout	Going out with friends	Quantitative
Dalc	Alcohol take during weekdays	Quantitative
Walc	Alcohol take during weekends	Quantitative
Health	Student's health	Quantitative
Absences	Number of times student was absent	Quantitative
G3	Final grade	Quantitative

Table 9: A Sample of Data of 2nd Dataset [42]

School	Sex	Age	Address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	1	0
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	0	1
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	1	0
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	0	1
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	0	1
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	0	1
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	0	0
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	1	1
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	0	1
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	0	1
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	0	1

Table 10: A Sample of Data of 2nd Dataset [42]

paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	1
no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	1
no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	1
no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	1
no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	1
no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	1
no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	1
no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	1
no	no	yes	yes	yes	no	4	2	2	1	1	1	0	15	16	1
no	yes	yes	yes	yes	no	5	5	1	1	1	5	0	12	12	1
no	no	yes	yes	yes	no	3	3	3	1	2	2	2	14	14	1

3.7 Justification of using Data sets:

In order to analyze student's performance it was necessary to train machine learning model on actual dataset of students but a) It is extremely difficult to obtain real student' data because of issues that are private and ethical for school, colleges and universities as I have tried to get dataset from NUST university but NUST has distributed database system. It means that each department has some portion of database and it was impossible to collect all of them and combine them and perform feature engineering on it, b) It is very expensive and time-consuming to conduct experiments on human subjects, c) "Kaggle" an online platform to do machine learning projects by using actual dataset was offering two student's dataset to solve student's performance issue by permission of colleges , and d) with the help of a personally made data generator software, I was able to generate more data which was based on original dataset got from kaggle website. [41] [42].

3.8 Summary

In this chapter, we have discussed about the requirements which user need in order to successfully train and get prediction from machine learning model. We have analyzed all the steps which user will take or go through while using the software. We have discussed and analyze both dataset which we are going to use with 11 machine learning models. We have justified why we can only use described dataset with machine learning models and why it was impossible to get other datasets.

Chapter 4

RESULTS AND ANALYSIS

4.1 Introduction:

In this chapter, we are discussing the results which we have got from training eleven machine learning models (retrieved from literature review) with two dataset. We will discuss result of each machine learning model with both datasets one by one and compare their accuracy with each other to get the best machine learning model suitable for student performance analysis process. We will discuss following model's results one by one.

Machine Learning Models:

- Support Vector Machine
- Neural Network
- Naive Bayes
- K-Nearest Neighbors
- Random Forest
- Linear Regression
- Logistic Regression
- Decision Tree
- AdaBoost
- Quadratic Discriminant Analysis
- Gussain Process

4.2 Results of First Dataset:

4.2.1 Support Vector Machine:

Table 11: Support Vector Machine Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.64	0.10	0.18	88
1	0.82	0.99	0.89	355
Accuracy	81%	**Total Data**		443

4.2.2 Neural Network:

Table 12: Neural Network Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.74	0.32	0.44	88
1	0.85	0.97	0.91	355
Accuracy	84%	**Total Data**		443

4.2.3 Naïve Bayes:

Table 13: Naïve Bayes Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.51	0.51	0.51	88
1	0.88	0.88	0.88	355
Accuracy	80%	**Total Data**		443

4.2.4 K-Nearest Neighbor:

Table 14: K-Nearest Neighbor Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.87	0.83	0.85	88
1	0.96	0.97	0.96	355
Accuracy	94%	**Total Data**		443

4.2.5 Random Forest:

Table 15: Random Forest Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.97	0.82	0.89	88
1	0.96	0.99	0.98	355
Accuracy	96%	**Total Data**		443

4.2.6 Linear Regression:

Table 16: Linear Regression Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.65	0.17	0.27	88
1	0.83	0.98	0.90	355
Accuracy	82%	**Total Data**		443

4.2.7 Logistic Regression:

Table 17: Logistic Regression Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.59	0.23	0.33	88
1	0.83	0.96	0.89	355
Accuracy	82%	**Total Data**		443

4.2.8 Decision Tree:

Table 18: Decision Tree Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.94	0.82	0.87	88
1	0.96	0.96	0.97	355
Accuracy	95%	**Total Data**		443

4.2.9 AdaBoost:

Table 19: AdaBoost Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.64	0.41	0.50	88
1	0.87	0.94	0.90	355
Accuracy	84%	**Total Data**		443

4.2.10 Quadratic Discriminant Analysis:

Table 20: Quadratic Discriminant Analysis Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.60	0.41	0.49	88
1	0.86	0.93	0.89	355
Accuracy	82%	**Total Data**		443

4.2.11 Gaussian Process:

Table 21: Gaussian Process Model Results of First Dataset

Output	Precision	Recall	F1-Score	Support
0	0.98	0.59	0.73	88
1	0.90	1.00	0.95	355
Accuracy	91%	**Total Data**		443

4.2.12 Result Graph for First Dataset:

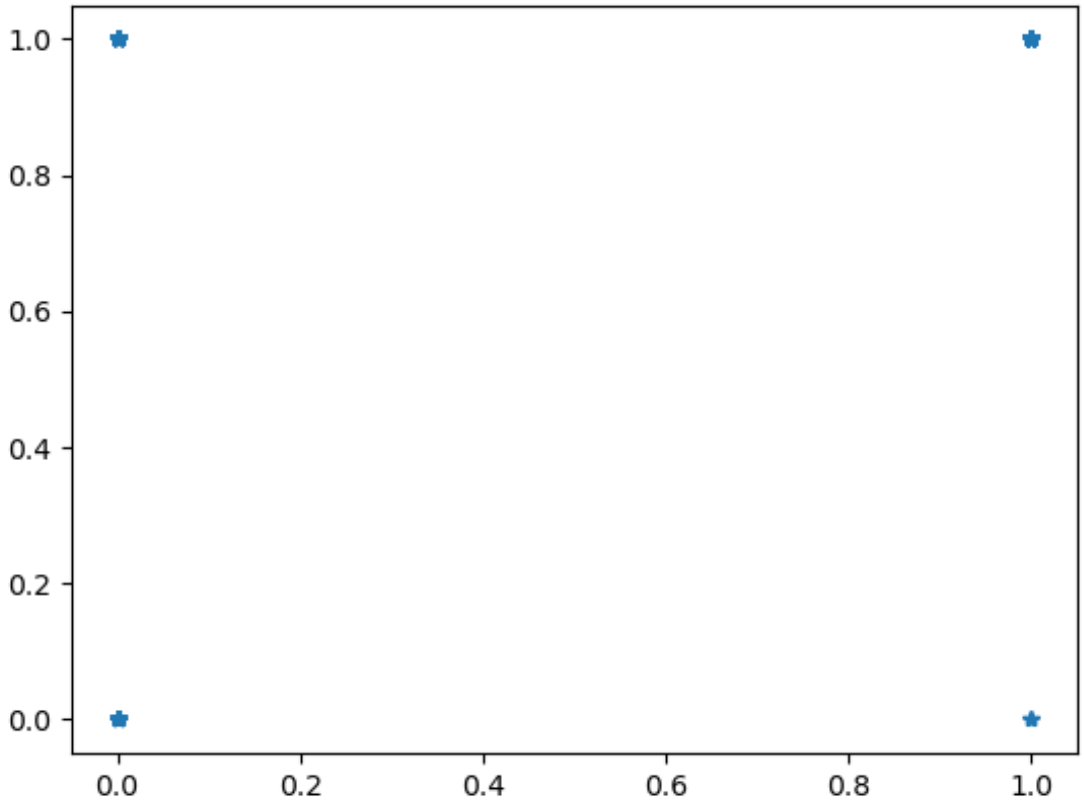

Figure 12: Result Graph for First Dataset

4.2.13 Accuracy Graph for First Dataset:

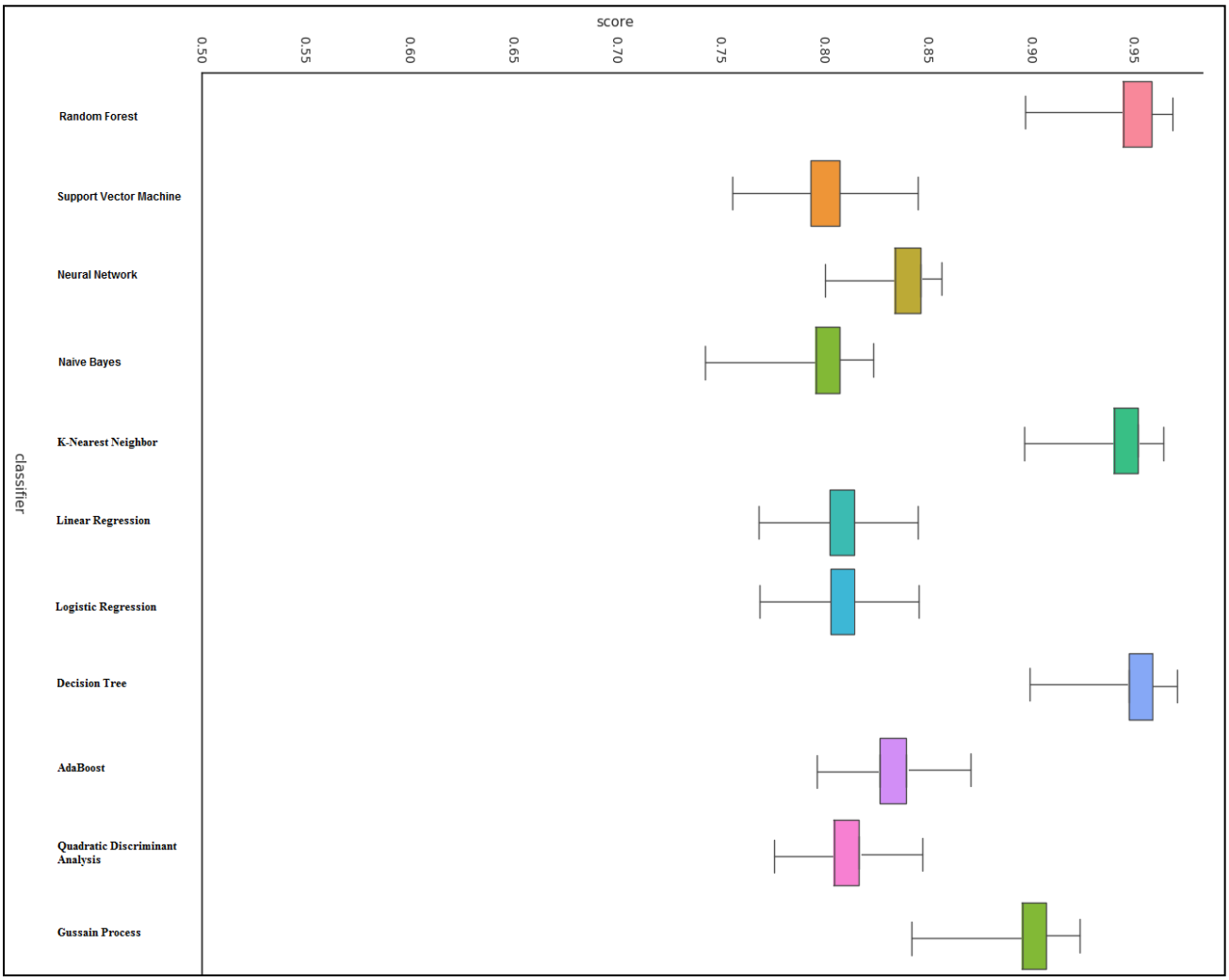

Figure 13: Accuracy Graph for First Dataset

4.3 Results of Second Dataset:

4.3.1 Support Vector Machine:

Table 22: Support Vector Machine Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.80	0.83	0.81	116
1	0.94	0.92	0.93	316
Accuracy	90%	**Total Data**		432

4.3.2 Neural Network:

Table 23: Neural Network Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.96	0.85	0.90	116
1	0.95	0.99	0.97	316
Accuracy	95%	**Total Data**		432

4.3.3 Naïve Bayes:

Table 24: Naïve Bayes Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.85	0.91	0.88	116
1	0.97	0.94	0.85	316
Accuracy	93%	**Total Data**		432

4.3.4 K-Nearest Neighbor:

Table 25: K-Nearest Neighbor Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.87	0.83	0.85	116
1	0.96	0.97	0.96	316
Accuracy	94%	**Total Data**		432

4.3.5 Random Forest:

Table 26: Random Forest Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.98	0.82	0.90	116
1	0.97	0.99	0.99	316
Accuracy	97%	**Total Data**		432

4.3.6 Linear Regression:

Table 27: Linear Regression Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.86	0.88	0.87	116
1	0.96	0.95	0.95	316
Accuracy	92%	**Total Data**		432

4.3.7 Logistic Regression:

Table 28: Logistic Regression Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.86	0.87	0.86	116
1	0.95	0.95	0.95	316
Accuracy	93%	**Total Data**		432

4.3.8 Decision Tree:

Table 29: Decision Tree Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.94	0.93	0.95	116
1	0.96	0.96	0.98	316
Accuracy	97%	**Total Data**		432

4.3.9 AdaBoost:

Table 30: AdaBoost Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	116
1	0.97	0.96	0.96	316
Accuracy	85%	**Total Data**		432

4.3.10 Quadratic Discriminant Analysis:

Table 31: Quadratic Discriminant Analysis Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.86	0.87	0.86	116
1	0.95	0.95	0.95	316
Accuracy	92%	**Total Data**		432

4.3.11 Gaussian Process:

Table 32: Gaussian Process Model Results of Second Dataset

Output	Precision	Recall	F1-Score	Support
0	0.92	0.79	0.87	116
1	0.89	0.95	0.95	316
Accuracy	93%	**Total Data**		432

4.3.12 Result Graph for Second Dataset:

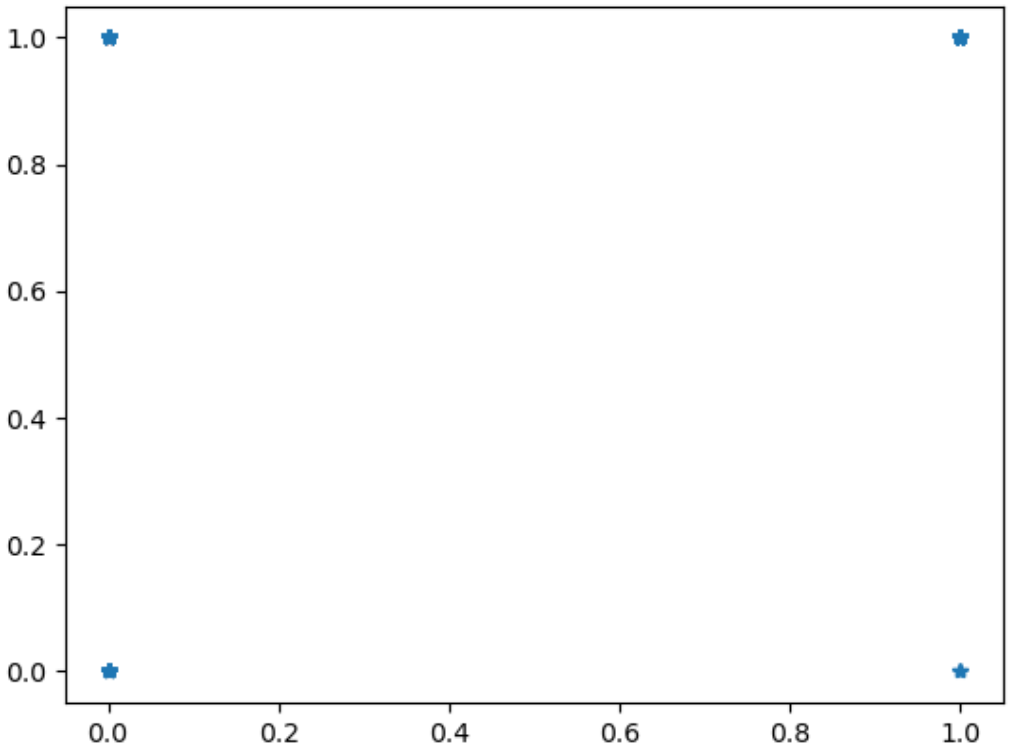

Figure 14: Result Graph for Second Dataset

4.3.13 Accuracy Graph for Second Dataset:

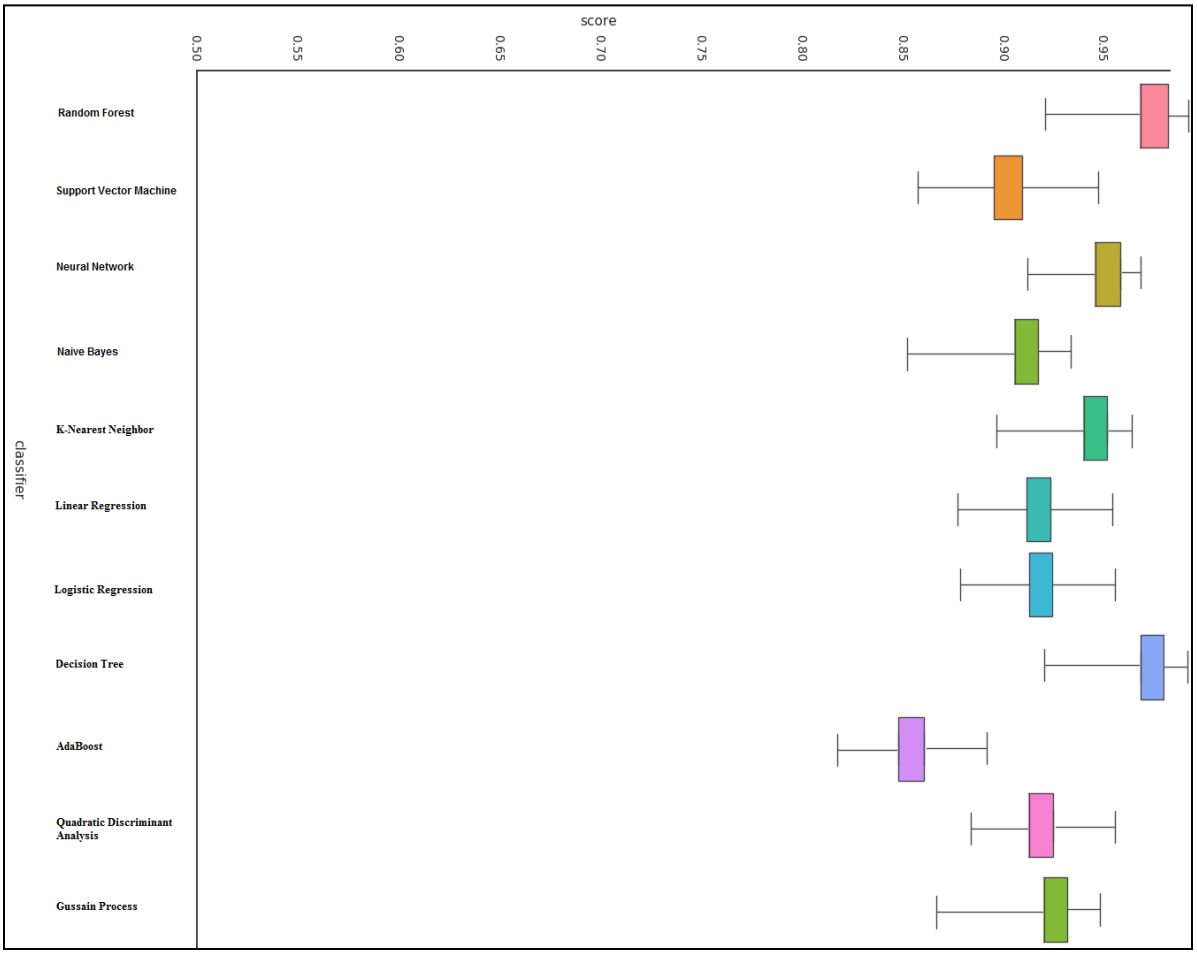

Figure 15: Accuracy Graph for Second Dataset

4.4 Comparison of Accuracies with First Dataset:

Table 33: Comparison of accuracies with first dataset

No.	Machine Learning Models	Accuracy
1	Support Vector Machine	81%
2	Neural Network	84%
3	Naive Bayes	80%
4	K-Nearest Neighbors	94%
5	Random Forest	96%
6	Linear Regression	80%
7	Logistic Regression	80%
8	Decision Tree	95%
9	AdaBoost	84%
10	Quadratic Discriminant Analysis	82%
11	Multi-Layer Perceptron	91%

4.5 Comparison of Accuracies with Second Dataset:

Table 34: Comparison of accuracies with second dataset

No.	Machine Learning Models	Accuracy
1	Support Vector Machine	90%
2	Neural Network	95%
3	Naive Bayes	93%
4	K-Nearest Neighbors	94%
5	Random Forest	97%
6	Linear Regression	94%
7	Logistic Regression	92%
8	Decision Tree	95%
9	AdaBoost	84%
10	Quadratic Discriminant Analysis	82%
11	Multi-Layer Perceptron	93%

According to the comparison of the accuracies of all machine learning models, we can clearly observe that Random Forest and Decision tree predicted result more accurately than all other machine learning model. Neural network also performed well but it was not as accurate then Random Forest and Decision Tree. May be if we had more student's record, it might give more accuracy than winning machine learning models.

4.6 Summary:

In this Chapter, we have trained 11 Machine learning models with two student's datasets and analyzed each machine learning model's result one by one with the help of their f-measure factors, recalls and precisions. We have compared their accuracies with each other and observed that **Random Forest** and **Decision Tree** are the best machine learning models as their accuracies for this specific "Student's Performance analysis and prediction" project were higher than other machine learning models. Neural Network's accuracy was also good and it might give more accuracy if we had more records in database as for any machine learning model, database's quality and quantity are the most important things to improve its accuracy. We have also observed all machine learning models' accuracy with the help of Graph to make it easier to understand.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Introduction

After discussion of related work, proposed methodology, and results of the proposed methodology in detail, we are able to conclude our research work. In this chapter, overall research done in this thesis is concluded and future work is discussed.

At first we have done some detailed literature review in which we have found problems being face by students during their education, teacher's strategies to solve student's problems and enhance their performance. We have observed that with the help of machine learning models we can also solve this problem but we do not know which machine learning model was best to use to tackle this problem and we do not even know how many machine learning models are there in the field. So we have done literature review and found that there are 11 machine learning models put the next problem was to find out student's database which was necessary to apply machine learning model and luckily with our research we found two student's databases and we have trained all machine learning models on both database and got our results.

5.2 Applications of this research work

Due to advancement of technologies and less number of jobs, Students are facing severe depression now a days and it not only affecting their performance but also depressing them from their future. With the help of our research,

- Student's future performance can be predicted
- Possible measurements can be taken to help those students

- With right database, we can also suggest future jobs and fields for the students
- With early detection of student's performance, Teacher can also focus and improve their teaching methods to help those students
- It will help to evaluate teacher's performance as well

5.3 CONCLUSION:

In this research work, we had studied about all the problems being faced by students during their education. We have also studied all the solutions being used by teachers to tackle those problems which are being faced by students. We have observed that mostly old teaching techniques are being followed to teach students and teachers are using old methods to help students which are not giving suitable results. With the advancement of technology and low number of jobs, students are facing severe depression and it is also affecting their performance.

With the help of machine learning, we can tackle student's performance problem. We have done detailed literature review to get to know about all famous machine learning models being used by researchers to solve different problems. We got 11 machine learning models through our research. In order to solve this problem, we needed to apply machine learning models on student's database which was not easy to get at first place. Luckily we got two public student's database on Kaggle website which contain 1000 records of students combined.

After that with some feature engineering process to polish database, we have applied machine learning models on two student's dataset to predict their future outcomes so that possible measurements can be taken to improve their performance and save their future. After machine learning models implementation and comparing their accuracies with each other, we discovered that "Random Forest" and "Decision Tree" are the best machine learning models to tackle student's performance issue because their accuracies were higher than other machine learning models.

5.4 FUTURE WORK:

In this research work, we have used two public databases which have helped us to achieved good accuracy but we need more detailed university database in order to make it a fully functional product which can help a lot of universities to predict their student's future performance.

Our main aim is to get database which has student's basic data including attendance, student's participation in class, marks of midterms, quizzes, final term etc. and also the job details after they graduate from university so that machine learning model can also predict which job fields are good for students based on their performance. It might give them a lot of new ideas and it will decrease their depression about their future.

Right now we are only predicting either student will get success or not but in future, we want to predict full marks, student's health, and student's carrier and suggest best strategies which can help to improve their performance.

REFERENCES

1. R.Devakunchari, "Analysis on big data over the years", International Journal of Scientific and Research Publications, vol. 4, no. 1, 2014.
2. Y. Altujjar, W. Altamimi, I. Al-Turaiki, and M. Al-Razgan, "Predicting critical courses affecting students' performance: a case Study", *Procedia Computer Science*, vol. 82, pp. 65-71, 2016. Available: 10.1016/j.procs.2016.04.010
3. Hemant Kumar Wani and Nilima Ashtankar, "An appropriate model predicting pest/diseases of crops using machine learning algorithms", *International Conference on Advanced Computing and Communication Systems*, 2017.
4. Olga Narushynska, Vasyl Teslyuk, and Bohdan-Dmytro Vovchuk, "Search model of customer's optimal route in the store based on the algorithm of machine learning A", *CSIT*, 2017.

5. Monica Ciolacu, Ali Fallah Tehrani, Rick Beer and Heribert Popp "Education 4.0 – fostering student's performance with machine learning methods", *23rd International Symposium for Design and Technology in Electronic Packaging*, 2017.
6. Haruka Motohashi, Tatsuro Teraoka, Shin Aoki and Hayato Ohwada, "Regression models and ranking method for p53 inhibitor candidates using machine learning", *International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018.
7. Evawaty Tanuar, Yaya Heryadi, Lukas, Bahtiar Saleh Abbas and Ford Lumban Gaol "using machine learning techniques to earlier predict student's performance", *INAPR International Conference*, 2018.
8. Ijaz Khan, Abir Al Sadiri, Abdul Rahim Ahmad and Nafaa Jabeur, "Tracking student performance in introductory programming by means of machine learning", IEEE, 2019.
9. M. Alloghani, D. Al-Jumeily, A. Hussain, A. Aljaaf, J. Mustafina and E. Petrov, "Application of machine learning on student data for the appraisal of academic performance", *11th International Conference on Developments in eSystems Engineering (DeSE)*, 2018.
10. Vladimir L. Uskov, Jeffrey P. Bakken, Ashok Shah and Adam Byerly "Machine learning-based predictive analytics of student academic performance in STEM education", *IEEE Global Engineering Education Conference*, 2019.
11. A. Elden, M. A. Moustafa, H. M. Harb, and A. H.Emara "Adaboost ensemble with simple genetic algorithm for student prediction model", *International Journal of Computer Science and Information Technology*, vol. 5, no. 2, pp. 73-85, 2013.
12. A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial", *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, p. 145, 2016.
13. R. MacDonald, "Software defect prediction from code quality measurements via machine learning", *Advances in Artificial Intelligence*, pp. 331-334, 2018.
14. I. Mihaylov, M. Nisheva, and D. Vassilev, "Machine learning techniques for survival time prediction in breast cancer", *Artificial Intelligence: Methodology, Systems, and Applications*, pp. 186-194, 2018.

15. C. Ma, B. Yao, F. Ge, Y. Pan and Y. Guo, "Improving prediction of student performance based on multiple feature selection approaches", *Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology - ICEBT 2017*, 2017.
16. A. Shahiri, W. Husain and N. Rashid, "A review on predicting student's performance using data mining techniques", *Procedia Computer Science*, vol. 72, pp. 414-422, 2015.
17. A. Shahiri, W. Husain and N. Rashid, "A review on predicting student's performance using data mining techniques", *Procedia Computer Science*, vol. 72, pp. 414-422, 2015.
18. L. Manhães, S. da Cruz and G. Zimbrão, "WAVE", *Proceedings of the 29th Annual ACM Symposium on Applied Computing - SAC '14*, 2014.
19. X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms", *Computers in Human Behavior*, vol. 98, pp. 166-173, 2019.
20. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes", *Computers & Education*, vol. 131, pp. 22-32, 2019.
21. D. Hernandez-Suarez et al., "Machine learning prediction models for in-hospital mortality after transcatheter aortic valve replacement", *JACC: Cardiovascular Interventions*, vol. 12, no. 14, pp. 1328-1338, 2019. Available: 10.1016/j.jcin.2019.06.013
22. R. Suguna, M. Shyamala Devi, R. Bagate and A. Joshi, "Assessment of feature selection for student academic performance through machine learning classification", *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 729-739, 2019.
23. "Students' Academic Performance Dataset", Kaggle.com. [Online]. Available: https://www.kaggle.com/aljarah/xAPI-Edu-Data.
24. Anderson, W., Mitchell, S. and Osgood, M., 2008. Gauging the Gaps in Student Problem-Solving Skills: Assessment of Individual and Group Use of Problem-Solving Strategies Using Online Discussions. CBE—Life Sciences Education, 7(2), pp.254-262.
25. Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B. and Osher, D., 2019. Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2), pp.97-140.
26. Hmelo-Silver, C., 2004. Problem-Based Learning: What and How Do Students Learn?. *Educational Psychology Review*, 16(3), pp.235-266.

27. Summary taken from G. Polya, "How to Solve It", 2nd ed., Princeton University Press, 1957, ISBN 0-691-08097-6
28. Abrandt Dahlgren, M., and Dahlgren, L. O. (2002). Portraits of PBL: Students' experiences of the characteristics of problem-based learning in physiotherapy, computer engineering, and psychology. Instr. Sci. 30: 111-127
29. Ames, C. (1992). Classrooms: Goals, structures, and student motivation. J. Educ. Psychol. 84: 261-271
30. Intaros, P., Inprasitha, M. and Srisawadi, N., 2014. Students' Problem Solving Strategies in Problem Solving-mathematics Classroom. *Procedia - Social and Behavioral Sciences*, 116, pp.4119-4123.
31. Stamp, M., 2018. A Survey of Machine Learning Algorithms and Their Application in Information Security. *Computer Communications and Networks*, pp.33-55.
32. Obulesu, O., Mahendra, M. and ThrilokReddy, M., 2018. Machine Learning Techniques and Tools: A Survey. *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*,
33. Agatonovic-Kustrin, S. and Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22(5), pp.717-727.
34. Tu, J., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11), pp.1225-1231.
35. Ren, Q., Cheng, H. and Han, H., 2017. Research on machine learning framework based on random forest algorithm.
36. Meskele Ashine, K. and Tessema Zewude, B., 2016. Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students' Academic Achievement: The Case of College of Natural and Computational Science, Wolaita Sodo University, Ethiopia. *Journal of Education and Practice*, 7, p.5.
37. Tharwat, Alaa. (2018). AdaBoost classifier: an overview. 10.13140/RG.2.2.19929.01122
38. Tharwat, A., 2016. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2), p.145.

39. Wang, W., Zhang, M., Chen, G., Jagadish, H., Ooi, B. and Tan, K., 2016. Database Meets Deep Learning. *ACM SIGMOD Record*, 45(2), pp.17-22.
40. Namatēvs, I., 2017. Deep Convolutional Neural Networks: Structure, Feature Extraction and Training. *Information Technology and Management Science*, 20(1).
41. "Students' Academic Performance Dataset", Kaggle.com. [Online]. Available: https://www.kaggle.com/aljarah/xAPI-Edu-Data.
42. "Student Grade Prediction", *Kaggle.com* [Online]Available: https://www.kaggle.com/dipam7/student-grade-prediction.

APPENDIX A: Simulation/ Software

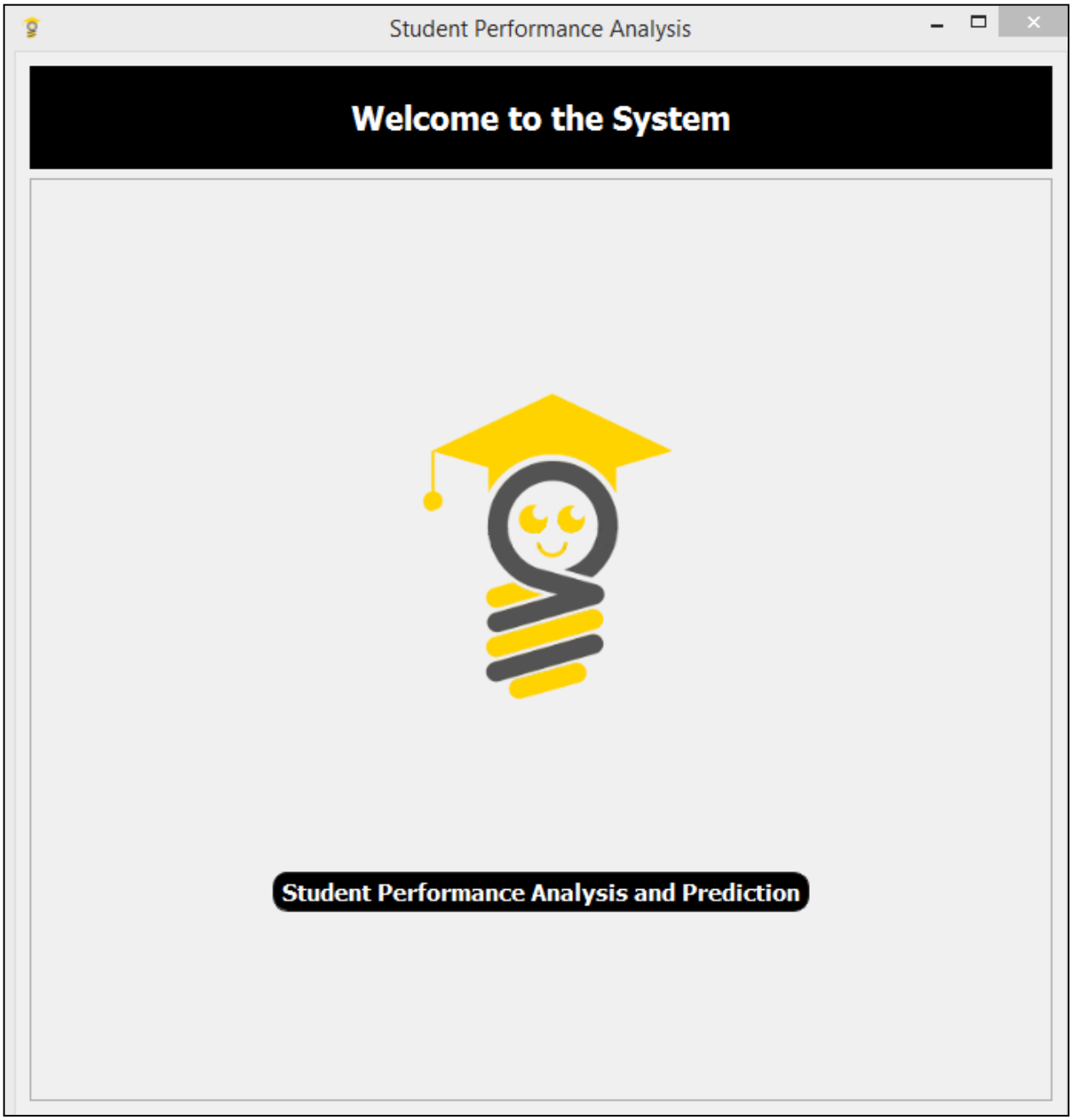

Figure 16: Welcome Screen

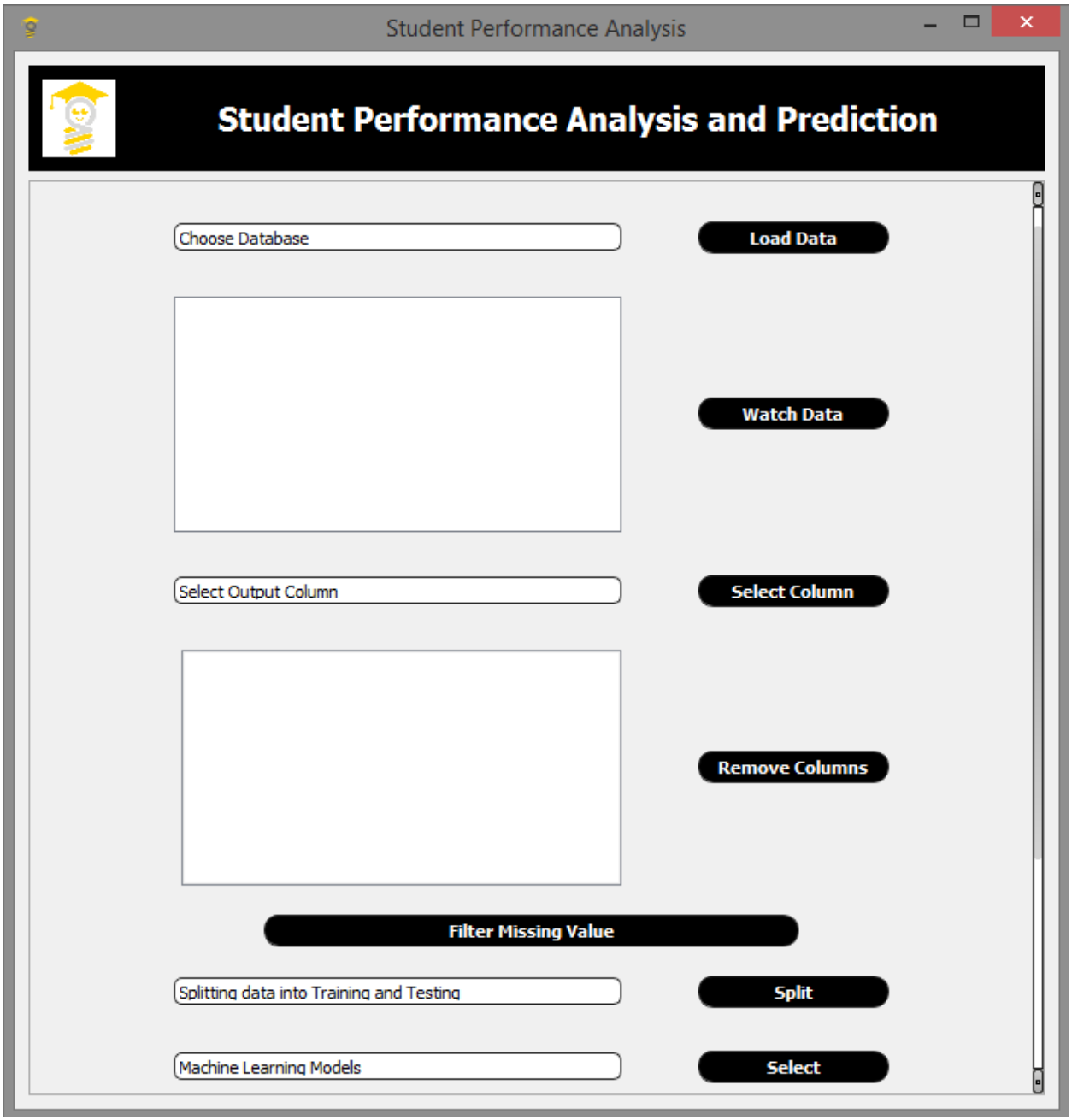

Figure 17: Screen 2, Fill in the inputs

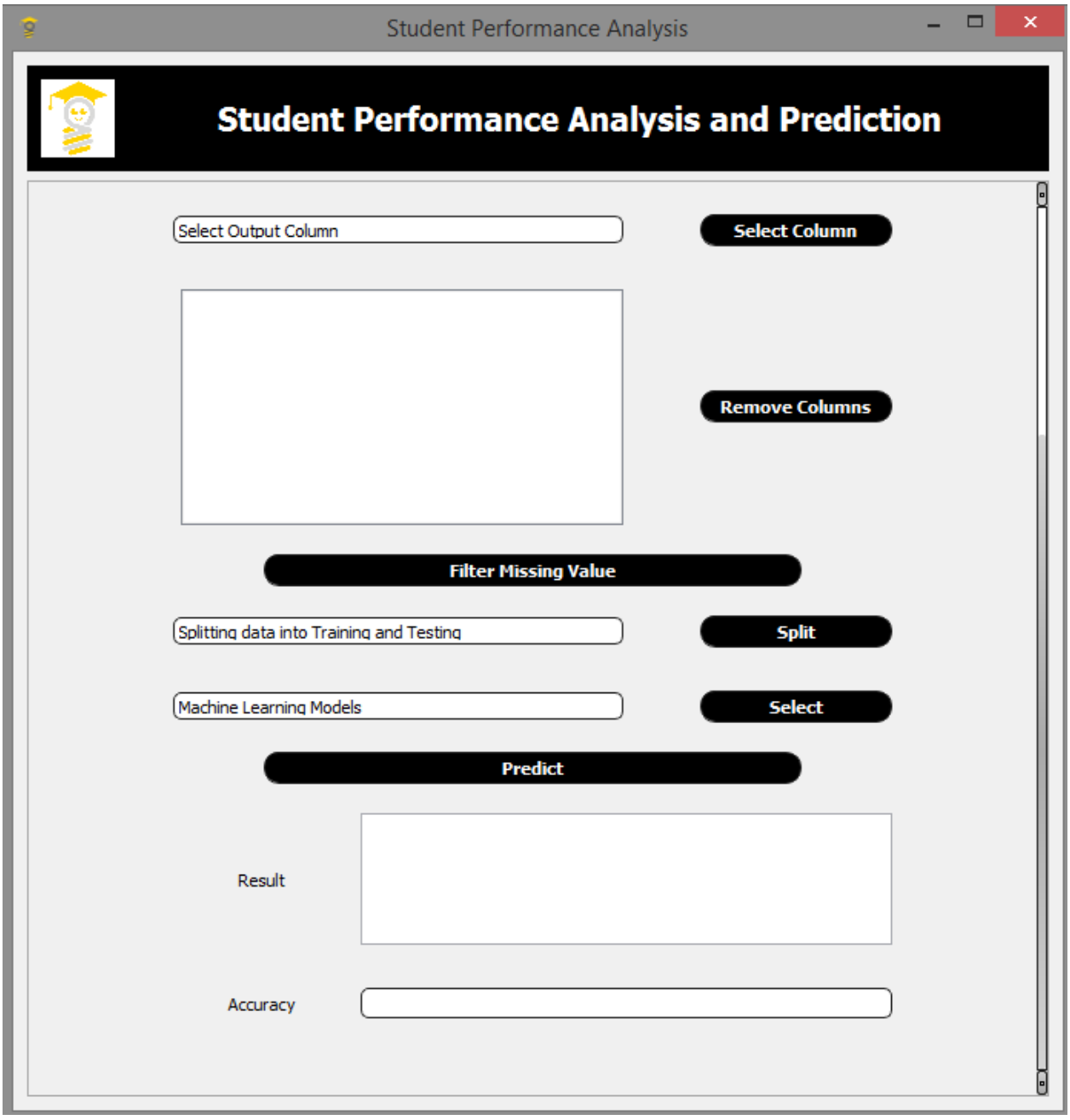

Figure 18: Screen 3, Choose Machine Learning Model

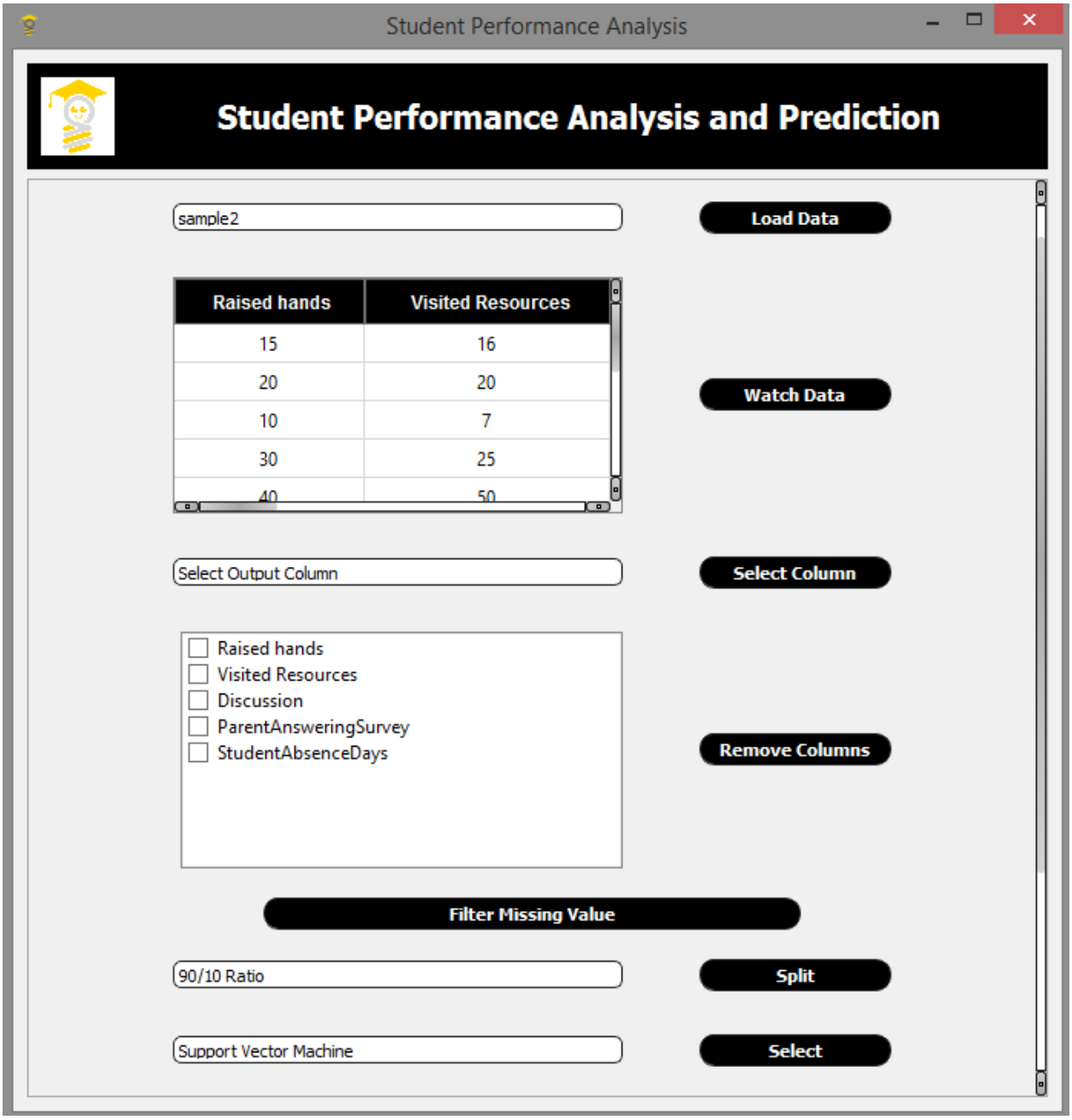

Figure 19: Data Filled in, Software in Running Condition

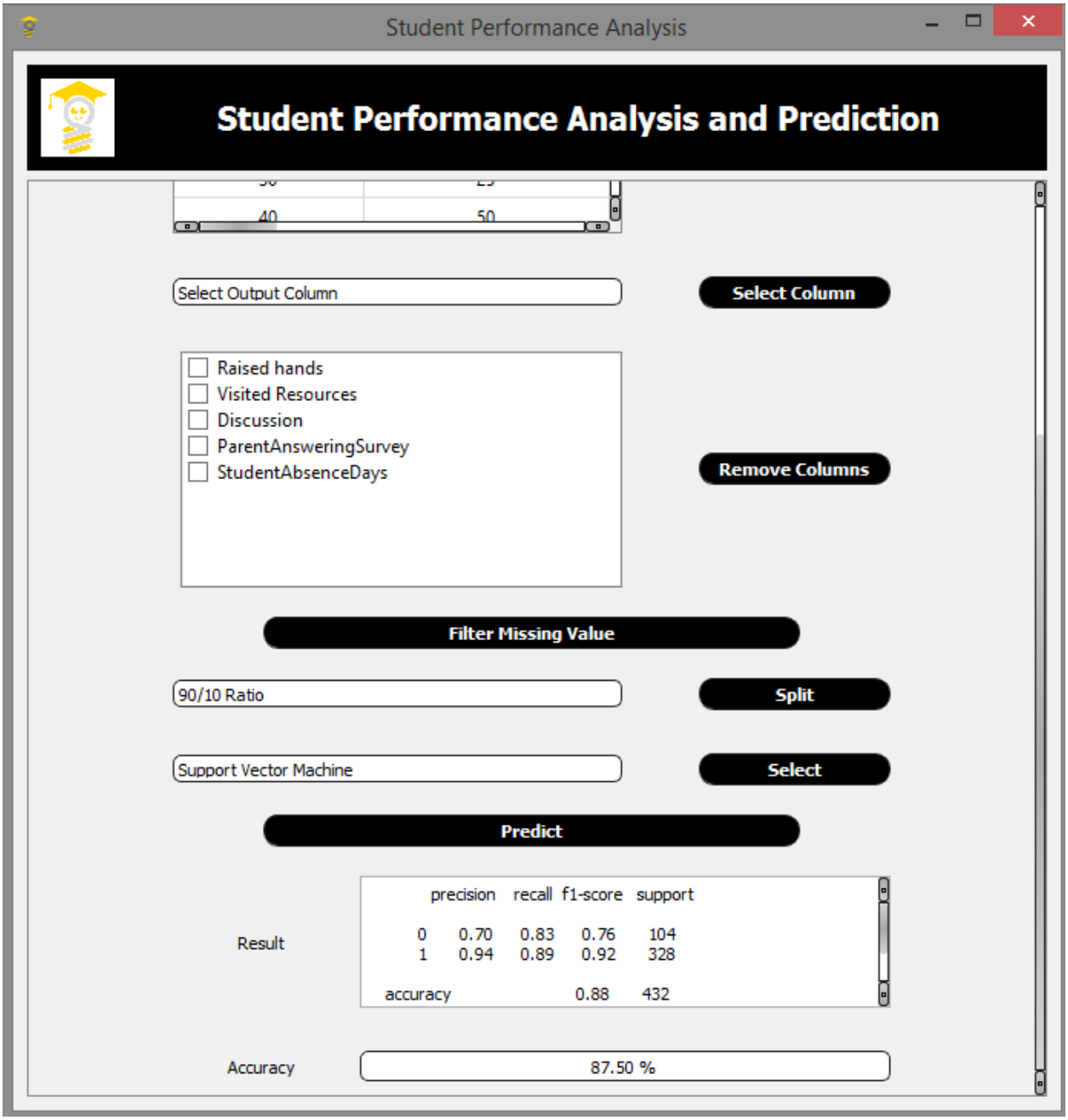

Figure 20: Machine Learning Model Result

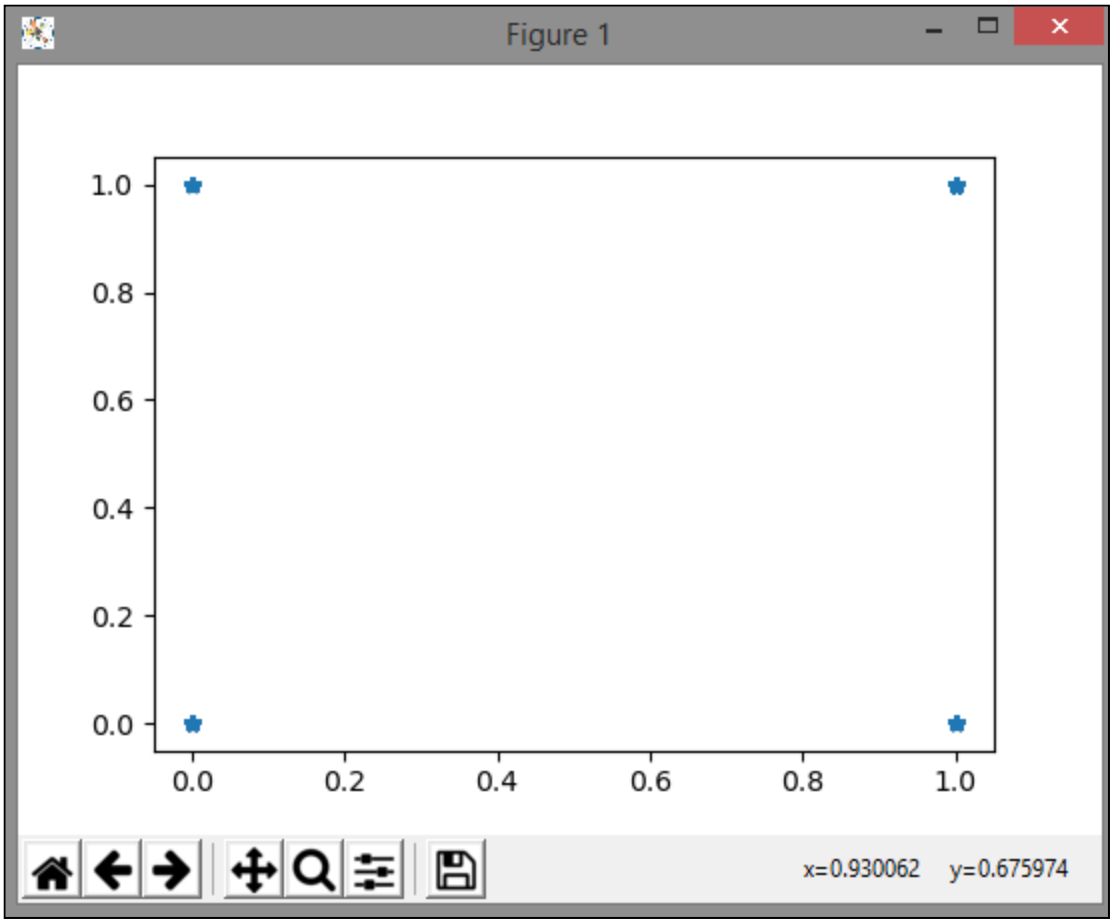

Figure 21: Result Graph