A COMPARATIVE STUDY OF CLASSIFIERS IN THE CONTEXT OF CRITICAL STUDENTS MANAGEMENT

BY

Dewan Mamun Raza
ID: 163-25-548

This Report presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Co-supervised By

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
May 2019

APPROVAL

This thesis titled **"A comparative study of classifiers in the context of critical students management"**, submitted by Dewan Mamun Raza, ID No: 163-25-548 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on May 05, 2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

....

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Co-supervised by:

....

Md. Tarek Habib
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Submitted by:

....

Dewan Mamun Raza
ID: 163-25-548
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGMENT

First and foremost, I would like to pay my sincerest gratitude to the almighty ALLAH for keeping me in sound and health during the work and giving me the ability to work hard successfully.

I would like to show gratitude to my research supervisor Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head, Department of CSE, Daffodil International University, Dhaka. Deep knowledge & keen interest of our supervisor in the field of “Data mining” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would also like to show gratitude to my research Co-supervisor Md. Tarek Habib, Assistant Professor, Department of CSE, Daffodil International University, Dhaka. Without his assistance and dedicated involvement in every step throughout the process, this thesis would have never been accomplished. His guidance helped me in all the time of research and writing of this thesis.

We would like to express our heartiest gratitude to Dr. Syed Akhter Hossain, Professor and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents and other family members.

ABSTRACT

Education can determines the standard of society, it also make the nation empowered providing new thoughts and implementation. In the last decades, it is found that number of higher level educational institutes grows rapidly in Bangladesh. Besides ensuring quality this increasing number causes tight competition of attracting students to get admitted in the institutes. This institute have higher rating tendency to fill all the available seats emphasizing on counting the number of students not on their academic excellence. Therefore, a remarkable number of student drop the course due to inability of adjustment with the academics which causing an ultimate loss to the family, society and educational institute. None knows the proper reason of their leave and what percent or who of student is going to become critical student. This paper investigated the prediction of dropout student through data mining approaches. The study predicts critical students applying different classification algorithm who tend to need support and essential guidelines from the different perspective. The outcomes are compared with each and also the models with the best.

TABLE OF CONTENTS

Contents	Page
Board of Examiners	I
Declaration	II
Acknowledgement	III
Abstracts	IV

Chapters

Chapter 1: Introduction	**1-4**
1.1 Introduction	1
1.2 Motivation	2
1.3 Research Question	3
1.4 Expected Output	3
1.5 Acquainting of this Research	4
Chapter 2: Background	**5-9**
2.1 Introduction	5
2.2 Related Works	6
2.3 Research Summary	8
2.4 Scope of the Problem	8
2.5 Challenges	9
Chapter 3: Research Methodology	**10-17**
3.1 Introduction	10
3.1.1 Data Mining Methodology	10
3.2 Key Techniques	11
3.2.1 Association	11
3.2.2 Classification	12

3.2.3 Clustering 12

3.2.4 Prediction 13

3.3 Adopted Approach 14

3.4 Data Collection Procedure 16

3.5 Implementation Requirements 17

Chapter 4: Experimental Results and Discussion 18-26

4.1 Introduction 18

4.2 Experimental Results and Descriptive Analysis 19

4.3 Summary 26

Chapter 5: Conclusion and Implication for Future Research 27-27

5.1 Conclusion 27

5.2 Recommendations 27

5.3 Implication for Further Study 27

REFERENCES 28-29

LIST OF FIGURES

Figure 3.1: Types of Data mining 10
Figure 3.2: Association Example 11
Figure 3.3: Clustering Techniques 13
Figure 3.4: Adopted Approach for recognizing critical Student’s 14
Figure 4.1: Critical Student’s Data set 19
Figure 4.2: Comparison of Accuracy for each classifier 22
Figure 4.3: Comparison of True Positive (TP) rate for each classifier 23
Figure 4.4: Comparison of False Positive (FP) rate for each classifier 23
Figure 4.5: Comparison of Precision for each classifier 24
Figure 4.6: ROC area curve 25
Figure 4.7: Comparison of Time complexity for each classifier 26

LIST OF TABLES

Table 4.1: Confusion matrix for each classifier 20
Table 4.2: Comparison of experimentally evaluated classifiers 21

Chapter 1: Introduction

1.1 Introduction

We are live in modern civilization which known as information-era, where accumulating data is easy and storing in inexpensive. Today the amount data generating by peoples, social media, digital gadgets and other devices increases day by day in all areas. The ability to understand and make use of it does not keep pace with its growth. To extract the novel and most useful information form huge amount of data is known as data mining. Method and concepts of Data mining can be applied in different areas like Education, Banking, marketing, customer relationship, Fraud detection, telecommunication and web mining etc. the data is related to the field of education is known as educational data mining. Educational data mining is an emerging increasing research interests in data mining. This new emerging field's uses many techniques such as Decision tress, Classification, neural networks, Naïve Bayes, k-nearest neighbor and many more techniques as the required of what types of data mining.

Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can used for predicting the performance of particular student in a particular course, level of regular-irregularity and other performance in an educational institutions so on. In the earlier prediction of students performance was most challenging task in higher education, therefor in that case one way to scale down the rate of dropout students and identify the critical students is Data analysis. There may have several reasons i.e. attribute to dropout or becoming the level of students critical. Ratio of student's dropout is quite often in first year of their graduation. Becoming critical students from normal students is caused by academic, family, personal reasons, campus environment and infrastructure of university and various issues.

In an educational institute examination plays a vital role of a student's life. Overall performance of a student's determined by assessment of internal as well as external also. Assessments are like attendance, class test, and lab-work, results of pre-requisite course and involvement of extracurricular activities. The individual academic performance in the examination of a student decides his future. Therefore it becomes essential to predict whether the students are critical or not. If the prediction says that a student tends to become critical prior to the examination then extra efforts can be taken to improve his studies and improving the quality of their enrolled institute.

1.2 Motivation

Education is a key factor for any nation which considered as backbone for sustainable development and stimulates the society to participate in their own development. It plays a crucial role not only in expanding further educational opportunities, but also in fostering basic intellectual abilities such as literacy that are crucial to success in a world where power closely linked with knowledge. So education is very important for a country like Bangladesh to develop the powers of reasoning and judgment and to create a space for self-actualization. More importantly this should not be education only; it must be quality education and the effectiveness of quality education of a country mainly depends on the policy guideline. It is very unfortunate or us that we failed to formulate and implement a proper education policy for our education system until 38 years of independence.

In 2018 After successful assessed by United Nations Economic Social Council declared that Bangladesh is recognized as a developing country from the list of least developing countries. Based on the up going score of three factors- Economic risk index, Human development index and capital income.

Among the three factors Human development was the most crucial for Bangladesh to recognize as a developing country form the list least developing country in 2018 under the assessment of UN Economic and social council. Though the level of score was satisfactory but still ratio of dropout from the institution is quite high which is not expect for a nation. Very unfortunate that the reason behind that a normal student becoming a critical student in their institution is not addressed properly there might be the lack of existing system to analyze and monitor the student progress and performance. Reasons are likely- existing traditional prediction methods is still insufficient to identify or due to lack of investigations on the factors effecting student performance. Therefore, in this research we proposed a systematic analysis using data mining techniques to identify the level of a student's criticality. Also this research helps to focus the prediction algorithm that how can be used to identify the most important attributes in a student database. Through this research study we could be able to improve the success ration of students in efficient way using Educational data Mining that is more impactful to institutions, educators, guardian, and students overall to build a successful nation.

1.3 Research Question

Significant research question is indeed for a meaningful research. The proposed research focused on two or more distinct yet related problems of predicting critical students.
All types of empirical studies such as preliminary studies, questionnaires, experiments, and case studies.

Therefore, the research questions proposed in this study are:

Q1: What are the important attributes used in predicting students’ performance?
Q2: What is the prediction methods used for students’ performance?

Results of educational data mining can be used by different members of education system. Understudies can utilize them to distinguish the activities, resources and learning and learning assignments to improve their learning. Educators can utilize them to get more expectation input, to distinguish understudies in danger and guide them to enable them to succeed, to recognize the most much of the time committed errors and to arrange the substance of the site in an effective manner. What's more, administrators can utilize them to settle on a choice which courses to offer, which graduated class are probably going to contribute more to the establishment and so forth.

However, it is better to start with a pilot study before going into the depth of this study. The purpose of doing the pilot study is to investigate the appropriateness of the research questions with the objectives of this study. Next, the study will explain the search strategy for conducting pilot study.

1.4 Expected Output

From this research study, we get some particular solution First one is among the various related and co-related attributes of a student’s performance which are the most important attributes used in predicting student’s performance. Process of feature selection is most crucial for that Second one is to select the prediction method based on the selection of feature for predicting performance. It is prior to select the appropriate techniques then feature in educational data mining. Predicting results varies though apply same feature in the different classification and clustering techniques of data mining. This study also shows a comparative study among the multiple classifiers techniques.

1.5 Acquainting of this Research

In this Research we try to discuss all about Educational data mining process for the critical students management by different classifier. We try to touch every step of EDM shortly but strongly with Cristal clear concept. Every part of this research, we face new challenges. We solve those problems professionally. For every new environment, we meet new scenario and new environment. Chapter 1 is an introductory portion. We discussed introduction of the research and in Chapter 2 beholds the literature about Educational performance measure related work. In chapter 3 have research methodology and adopted model for this research. In chapter 4 we described experimental analysis, descriptive analysis and their summary finally in Chapter 6 is the last chapter of this research. Conclusion and future works are define here.

Chapter 2: Background

2.1 Introduction

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary territory centering upon systems for extricating valuable learning from information. The continuous quick development of online information because of the Internet and the boundless utilization of databases have made a tremendous requirement for KDD procedures. The test of separating information from information attracts upon research insights, databases, design acknowledgment, AI, information representation, improvement, and superior processing, to convey propelled business knowledge and web disclosure arrangements. What's more, as of late there are expanding research interests in Educational Data Mining (EDM) In addition to this, recently there are increasing research interests in Educational Data Mining (EDM). EDM is a field that exploits statistical, machine-learning, and data-mining algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational research issues. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn[1]. Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which often need to be determined by properties in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data. The main objective of educational institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students’ performance. The found learning can be utilized to offer a supportive and helpful proposals to the scholastic organizers in advanced education organizations to upgrade their basic leadership process, to improve understudies 'scholarly execution and trim down disappointment rate, to all the more likely comprehend students‟ conduct, to help educators, to improve instructing and numerous different benefits.[2],[3].

Educational Data mining can be implemented in many techniques such as decision trees, neural networks, k-nearest Neighbor, Naive Bayes, support vector machines and many others. using these methods many kinds of knowledge can be discovered such as association rules, classification, clustering, pruning the data. The main objective of this paper is to predict the student academic performance and make a comparative study on

Bayesian network classifiers, through that we compute which classifier predicts more students when compared to other classifiers. In this paper, student's information like Previous Semester Performance, Attendance, Seminar , Assignment marks, Internal marks, and whether the student has attend any Co-curricular Activities are collected from students to predict the performance at the end of the semester examination.

2.2 Related Works

In this section, we briefly discuss the works which is similar techniques as our approach but serve for different purposes.

Nguyen et al. [4] compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes. These predictions are most useful for identifying and assisting failing students, and better determine scholarships. As a result, the decision tree classifier provided better accuracy in comparison with the Bayesian network classifier.

Al-Radaideh et al. [5] proposed to use data mining classification techniques to enhance the quality of the higher educational system by evaluating students' data that may affect the students' performance in courses. They used the CRISP framework for data mining to mine students' related academic data. A classification model was built using the decision tree method. They used three different classification methods ID3, C4.5 and the NaïveBayes. The results indicated that the decision tree model had better prediction accuracy than the other models. As a result, a system was built to facilitate the usage of the generated rules that students need to predict the final grade in the C++ undergraduate course.

Cesar et al. [6] proposed the use of a recommendation system based on data mining techniques to help students to make decisions related to their academic track. The system provided support for students to better choose how many and which courses to enroll on. As a result, the authors developed a system that is capable to predict the failure or success of a student in any course using a classifier obtained from the analysis of a set of historical data related to the academic field of other students who took the same course in the past.

Muslihan et al. [7] have compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. Students' data were collected from the data of the National Defence University of Malaysia (NDUM). As a result, the technique that gives accurate prediction and classification was chosen as the best model. Using the proposed model, the pattern that influences the student's academic performance was identified.

Han and Kamber [8] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

Bharadwaj and Pal [9] conducted study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University,

Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching,

mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance

Pandey and Pal [10] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Ramaswami and Bhaskaran [11] have constructed a predictive model called CHAID with 7-class response variable by using highly influencing predictive variables obtained through featureselection so as to evaluate the academic achievement of students at higher secondary schools in India. Data were collected from different schools of Tamilnada, 772 students' records were used for CHAID prediction model construction. As a result, set of rules were extracted from the CHAID prediction model and the efficiency was found. The accuracy of the present model was compared with other models and it has been found to be satisfactory.

Shannaq et al. [12], connected the arrangement as information mining procedure to anticipate the quantities of selected understudies by assessing scholarly information from enlisted understudies to think about the principle traits that may influence the understudies' faithfulness (number of enlisted understudies). The extricated characterization rules depend on the choice tree as a grouping strategy; the separated arrangement rules are examined and assessed utilizing diverse assessment strategies. It enables the University the board to plan fundamental assets for the recently selected understudies and demonstrates at a beginning time which kind of understudies will conceivably be enlisted and what territories to amass upon in advanced education frameworks for help.

2.3 Research Summary

To summarize, the field of data mining is concerned with finding new patterns in large amounts of data. Widely used in business, data mining also has several application areas in education. One of the most useful DM tasks in online learning is the ability to classify a student's retention and attrition. There are different educational objectives for using classification, such as: exploring students' reactions towards a certain instructional strategy, discovering students' characteristics and so forth (Romero et al., 2008). The aim of the study is to examine dropout prediction with the help of student personal characteristics using data mining techniques in an online program. The main focus of this study is not only to examine the effects of student personal characteristics on dropout, but also, to attempt performing various data mining techniques for such analyses. Four different classifiers, which were based on k-Nearest Neighbor (k-NN), Decision Trees (DT), Naive Bayes (NB) and Neural Networks (NN), were trained and tested using 10-fold cross validation. The prediction performances of k-NN, DT, NB and NN classifiers were compared by showing the prediction results and plotting ROC (Receiver Operating Characteristics graph of sensitivity, y-axis, versus specificity, x-axis) curves. Also, Genetic Algorithm (GA) was applied to find an optimal set of feature weights that are the most important factors in dropout prediction. As such, we provide a good sample and an illustration of using these various techniques for the analysis of students' dropouts based on the surveyed data, which has potential to contribute to existing literature in this context.

2.4 Scope of the Problem

Data mining is a tremendously vast area for research that includes employing different techniques and algorithms for pattern finding. Educational data mining is one of them emerging domain which increases the research interest of the researchers now a day. Scope of the EDM is everywhere related to educational institution to improve educational results and explain educational strategies for further decision making. In the field of EDM data mining has been using to analyze the early students' performance, attitude and behavior prediction for decrease the ratio of student's dropout and enhancing their skills through applying tweaks and tricks in their weak point.

2.5 Challenges

After summarize the technical issue concerned with EDM and the challenges associated with this research are as follows-

- Getting good quality datasets with needed attributes.
- Preprocessing of the datasets for the effective analysis.
- Develop algorithms that minimize the amount of data needed.
- Define how to measure the efficiency of algorithms output.
- Choosing a proper classifier for the highest accuracy.
- Getting enough data to trained them properly.
- Legal privacy Issue
- Define how to choose the algorithms parameters.

Chapter 3: Research Methodology

3.1 Introduction

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data [13].

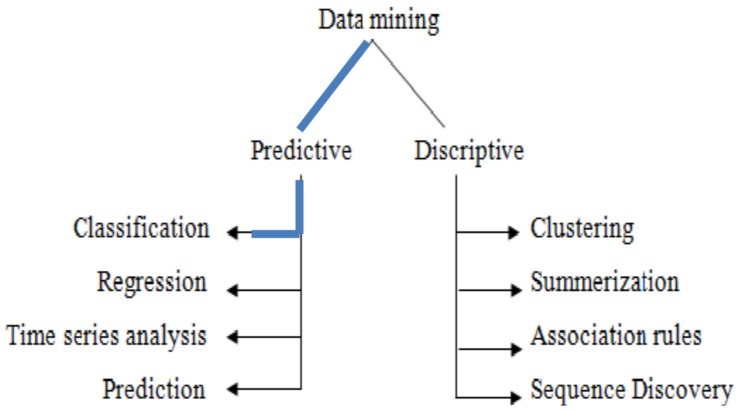


Figure 3.1: Types of Data mining

3.1.1 Mining Methodology

It refers to the following kinds of issues [15]–

- Mining different kinds of knowledge in databases.
- Interactive mining of knowledge at multiple levels of abstraction.
- Incorporation of background knowledge.
- Data mining query and ad hoc data mining.
- Presentation and visualization of data mining results.

- Handling noisy or incomplete data.
- Pattern evaluation.

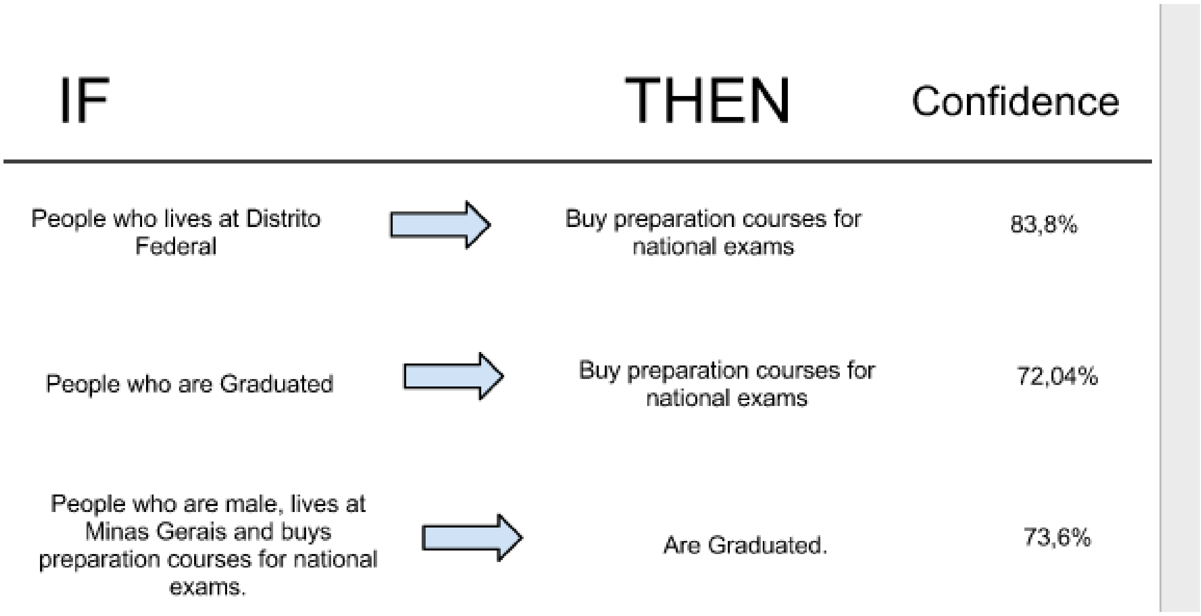
3.2 Key Techniques

Several core techniques that are used in data mining describe the type of mining and data recovery operation. Unfortunately, the different companies and solutions do not always share terms, which can add to the confusion and apparent complexity.

Let's look at some key techniques and examples of how to use different tools to build the data mining.

3.2.1 Association

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, we make a simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, you might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.
Building association or relation-based data mining tools can be achieved simply with different tools. Information flow that is used in association-



IF		THEN	Confidence
People who lives at Distrito Federal	⇨	Buy preparation courses for national exams	83,8%
People who are Graduated	⇨	Buy preparation courses for national exams	72,04%
People who are male, lives at Minas Gerais and buys preparation courses for national exams.	⇨	Are Graduated.	73,6%

Figure 3.2: Association Example [16]

3.2.2 Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that “given all records of employees who left the company; predict who will probably leave the company in a future period.” In this case, we divide the records of employees into two groups that named “leave” and “stay”. And then we can ask our data mining software to classify the employees into separate groups [16].

3.2.3 Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.

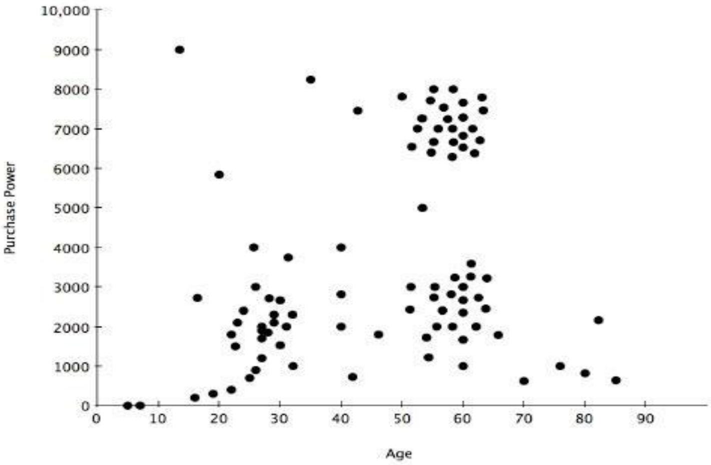


Figure 3.3: Clustering Techniques [16]

In the example, we can identify two clusters, one around the US$2,000/20-30 age group, and another at the US$7,000-8,000/50-65 age group. In this case, we've both hypothesized and proved our hypothesis with a simple graph that we can create using any suitable graphing software for a quick manual view [16].

3.2.4 Prediction

Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances, you can make a prediction about an event. Our research objective is to predict the performance of critical students.

Using the credit card authorization, for example, you might combine decision tree analysis of individual past transactions with classification and historical pattern matches to identify whether a transaction is fraudulent. Making a match between the purchase of flights to the US and transactions in the US, it is likely that the transaction is valid.

3.3 Adopted Approach

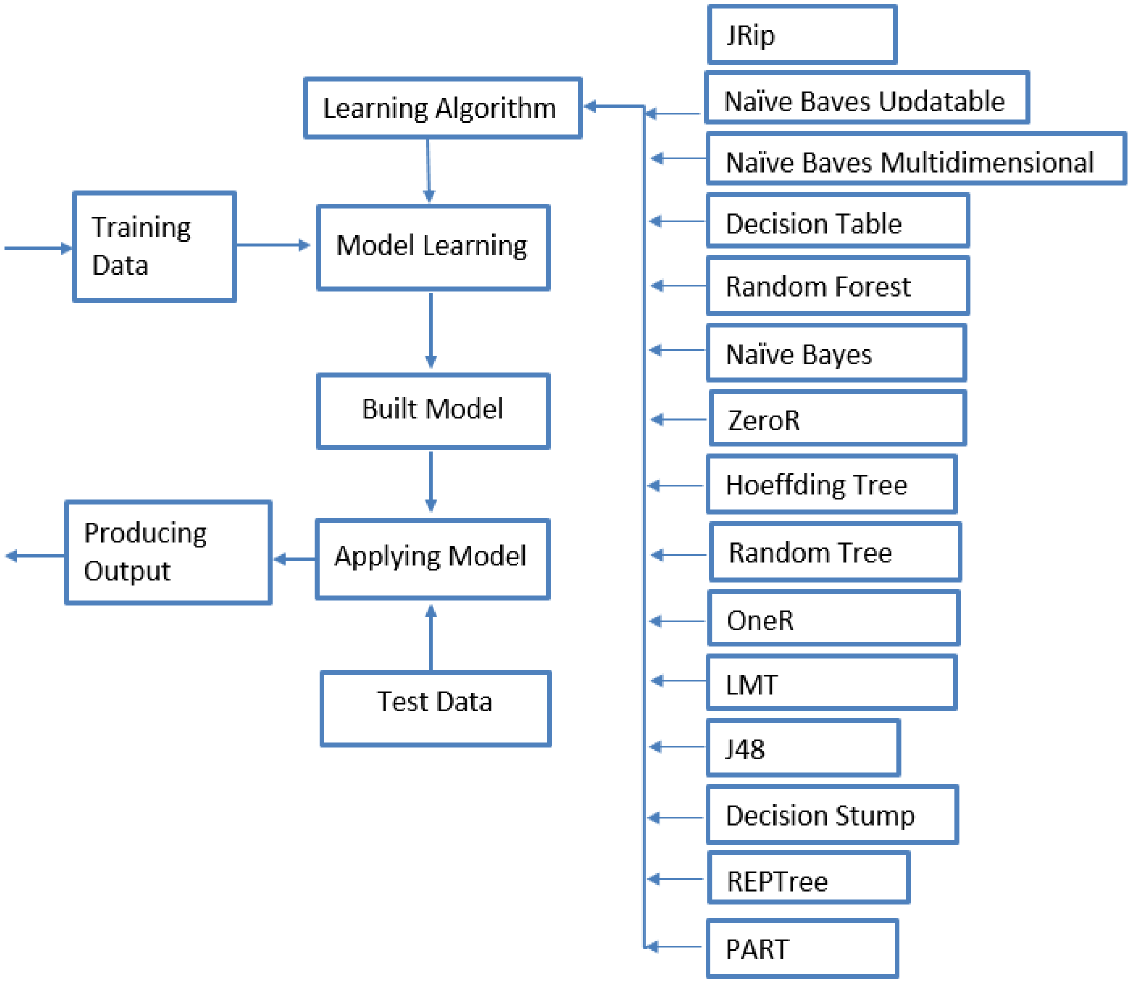


Figure 3.4: Adopted Approach for recognizing critical Student’s

The model shown in figure 3.3 which is adopted in this research is described below its different components briefly. The main objectives of the above given approach to built a model by applying some learning algorithm on the training data set. This is known as classifiers in data mining. After applying the learning algorithm on training data three will have some outputs like accuracy, precision, f-measure, TP rate, FP rate, ROC value etc. All the result produces with the calculating of what number of data’s been performed computational calculations perfectly or not that is defined in predicted and actual class as confusion matrix.

A confusion matrix is a technique that often used to summarizing the performance of a classification model on a set of test data for which the true values are known. The confusion matrix shows the ways in which your classification model is confused when it makes predictions.

Although a confusion matrix provides the information needed to determine how well a classification model performs, summarizing this information with a single number would make it more convenient to compare the performance of different models.

This can be done using a performance metric such as accuracy, which is defined as follows:
Confusion matrix is table with 4 different combinations of predicted and actual class. Predicted classes are represented in the columns of the matrix on the hand the actual classes are in the rows of the matrix.

Different term of confusion matrix is described below-

- Positive (P): Observation is positive.
- Negative (N): Observation is not positive.
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

Now the performance of all classifiers will be discussed from the table 4.2 which are experimentally evaluated through the data mining tools Weka.

Classification Accuracy: way of calculating the rate of accuracy is given below by the relation-

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP\ +TN\ +FP+FN}$$

Recall: Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given below by the relation-

$$\textbf{\textit{Recall}} = \frac{TP}{TP+FN}$$

Precision: To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High precision indicates an examples labeled as the as positive is indeed positive (small number of FP).

Precision is given below by the relation-

$$\textbf{\textit{Precision}}= \frac{TP}{TP+FP}$$

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F-measure: Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$\textbf{\textit{F – measure}}= \frac{2 *Recall *Precision}{Recall +Precision}$$

A couple other terms are also worth mentioning: Null Error Rate, Cohen's Kappa, F Score, ROC Curve.

3.4 Data Collection Procedure

In this step only those fields were selected which were required for data mining. The data are collected from the regular students who had studied in Bachelor course in different level and different terms in the department of CSE. The data set used in this study was obtained from Daffodil international University, 2017.For this research purpose we collected the data in two ways one is from the exam control section where they produces the report in every semester of the university students those who are not performing well and becoming irregular in their studies and the other ways by the academic advisor of students means respected faculty members. The data is stored in a database: MS Excel. We have used MS Excel because is the world's most popular database. Using MS-excel we preprocessed the uncategorized data into categorize them. Data we have used most of their types is numerical which is easy to preprocessed than other types. Whenever preprocessed is done now the next step is to make the data set for Weka readable format which is might also part of preprocess. Since the data mining software used to generate association rules accepts data only in .arff format, we have first converted the data on MS Excel file into comma separated text format and then to .arff format.

3.5 Implementation Requirements

The research subjects of this paper are datasets of critical students of undergraduate course of a university. The programming platform needed to analyze the data is Weka data mining tool, MS excel. The various tools for data mining them some are like-DBMiner, Rapidminer and Weka. We choose Weka because of easier than other for applying the adopted algorithm what we thought. Additionally for exploration for the curve drawing illustrator is used. To perform efficient and effectively to avoid time complexity GPU enabled computing system is needed.

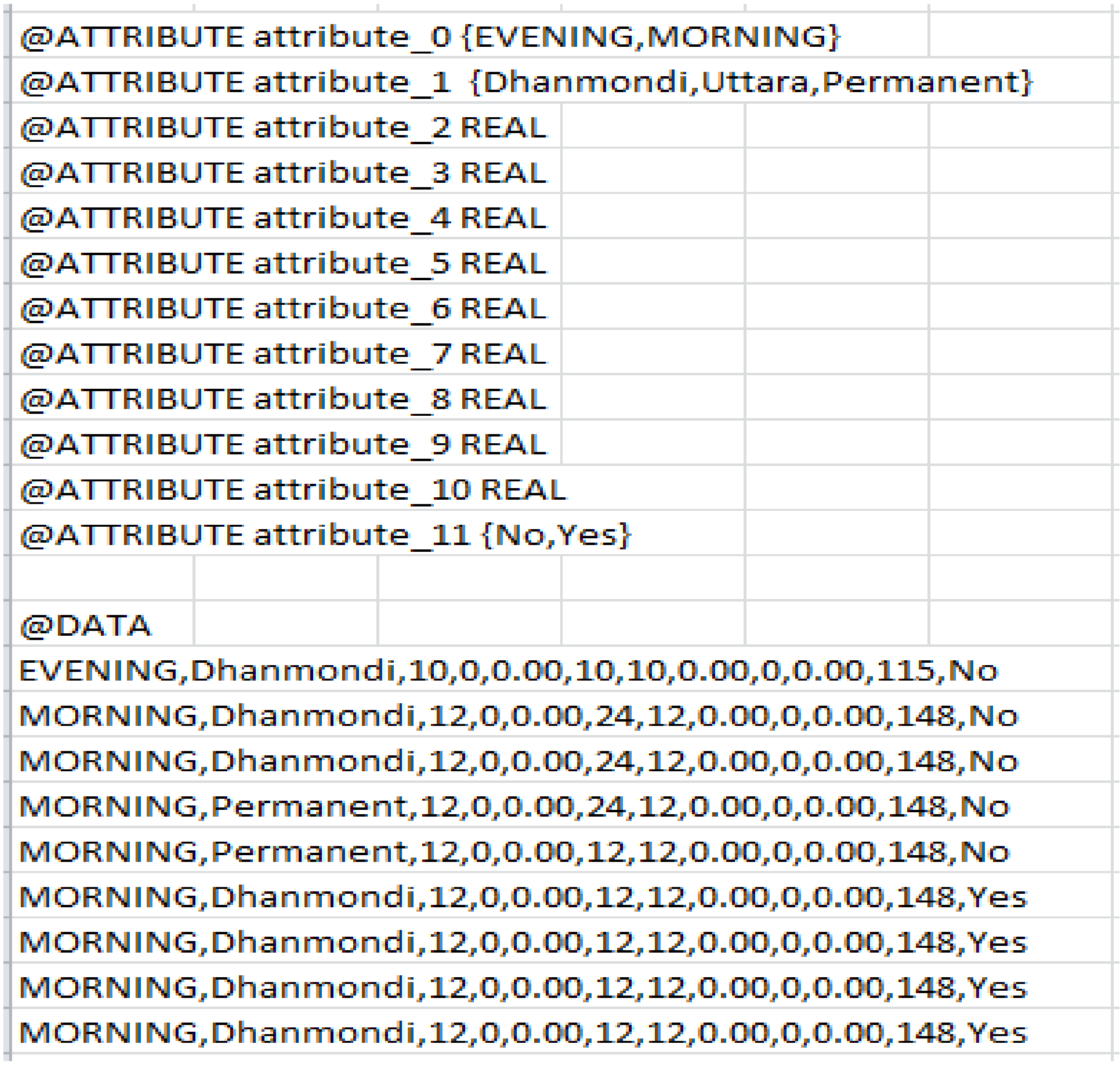
Chapter 4: Experimental Results and Discussion

4.1 Introduction

Weka is an open source software issued under the GNU General Public License. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [12].

After Starting WEKA, a GUI chooser pop up and lets you choose five ways to work with WEKA and your data. For all the examples in this project, we will choose only the Explorer option. This option is more than sufficient for everything we need to do in this project.

Building the Data Set for Weka: To load data into WEKA, we have to put it into a format that will be understood. Weka's preferred method for loading data is in the Attribute-Relation File Format (ARFF), where you can define the type of data being loaded, then supply the data itself. In the file, you define each column and what each column contains.



```
@ATTRIBUTE attribute_0 {EVENING,MORNING}
@ATTRIBUTE attribute_1  {Dhanmondi,Uttara,Permanent}
@ATTRIBUTE attribute_2 REAL
@ATTRIBUTE attribute_3 REAL
@ATTRIBUTE attribute_4 REAL
@ATTRIBUTE attribute_5 REAL
@ATTRIBUTE attribute_6 REAL
@ATTRIBUTE attribute_7 REAL
@ATTRIBUTE attribute_8 REAL
@ATTRIBUTE attribute_9 REAL
@ATTRIBUTE attribute_10 REAL
@ATTRIBUTE attribute_11 {No,Yes}

@DATA
EVENING,Dhanmondi,10,0,0.00,10,10,0.00,0,0.00,115,No
MORNING,Dhanmondi,12,0,0.00,24,12,0.00,0,0.00,148,No
MORNING,Dhanmondi,12,0,0.00,24,12,0.00,0,0.00,148,No
MORNING,Permanent,12,0,0.00,24,12,0.00,0,0.00,148,No
MORNING,Permanent,12,0,0.00,12,12,0.00,0,0.00,148,No
MORNING,Dhanmondi,12,0,0.00,12,12,0.00,0,0.00,148,Yes
MORNING,Dhanmondi,12,0,0.00,12,12,0.00,0,0.00,148,Yes
MORNING,Dhanmondi,12,0,0.00,12,12,0.00,0,0.00,148,Yes
MORNING,Dhanmondi,12,0,0.00,12,12,0.00,0,0.00,148,Yes
```

Figure 4.1: Critical Student’s data set

4.2 Experimental Results and Descriptive Analysis

The result of confusion matrix is found after performing each classifier which is visualized in tabular form in respect to the table 4.1. This is shown below –

Table 4.1: Confusion matrix for each classifier

Classifier	Matrix				Classifier	Matrix			
Naïve Bayes			Predicted class		Random Forest			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1409	115		Actual Class	Class=1	1508	16
		Class=0	21	66			Class=0	47	40
Naïve Bayes Multidimensional			Predicted class		LMT			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1524	0		Actual Class	Class=1	1508	16
		Class=0	87	0			Class=0	47	40
Naïve Bayes Updatable			Predicted class		Hoeffding Tree			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1409	115		Actual Class	Class=1	1524	0
		Class=0	21	66			Class=0	87	0
Decision table			Predicted class		J48			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1498	26		Actual Class	Class=1	1496	43
		Class=0	54	33			Class=0	44	28
JRip			Predicted class		Random Tree			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1480	44		Actual Class	Class=1	1481	43
		Class=0	35	52			Class=0	45	42
OneR			Predicted class		Decision Stump			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1499	25		Actual Class	Class=1	1524	0
		Class=0	59	28			Class=0	87	0
PART			Predicted class		REPTree			Predicted Class	
			Class=1	Class=0				Class=1	Class=0
	Actual Class	Class=1	1484	40		Actual Class	Class=1		
		Class=0	44	43			Class=0		
ZeroR			Predicted class						
			Class=1	Class=0					
	Actual Class	Class=1	1524	0					
		Class=0	87	0					

In the Table 4.1 we see the confusion matrix is given for each classifier which applied in this research distinctly.

In the field of data mining specially the problem related to the statistical classification, a confusion matrix plays vital role to visualizing and understanding the model applied for classifications better effectiveness, better performance exactly we want. Since we described the confusion matrix in the section Deployed approach chapter 3.

Above shown confusion matrix in Table 4.1 for a two-class classification problem predicted class and actual class. Confusion matrix for Naive Bayes classification shows that True positive is 1444 as pointing the cell actual class yes (1) and predicted yes (1) on the other hand and True negative is 444 as pointing actual class (0) and predicted (0) finally we can say that out of 1611 cases the classifier predicted critical yes 1430 students and no predicted 181 students. In reality 1524 students among 1611 possible to becoming critical and 87 students not. As this way other classifiers also calculated and so on.

TABLE 4.2:- Comparison of experimentally evaluated classifiers.

Classifier	*Accuracy*	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*ROC Area*	*Time taken to build model (ms)*
Naïve Bayes	91.55%	0.916	0.232	0.952	0.916	0.947	20
Nave Bayes Multidimensional	94.59%	0.946	0.946	0.932	0.946	0.487	20
Naïve Bayes updatable	91.55%	0.916	0.232	0.952	0.916	0.947	10
Decision table	95.03%	0.950	0.588	0.943	0.950	0.937	220
JRip	95.09%	0.951	0.382	0.953	0.951	0.802	190
OneR	94.78%	0.948	0.942	0.939	0.948	0.653	12
PART	94.78%	0.948	0.480	0.947	0.948	0.836	30
ZeroR	94.59%	0.946	0.946	0.922	0.946	0.487	10
LMT	96.08%	0.961	0.512	0.956	0.961	0.966	550
HoeffdingTree	94.59%	0.946	0.946	0.930	0.946	0.487	60
J48	95.53%	0.955	0.479	0.952	0.955	0.883	10
RandomTree	94.53%	0.945	0.491	0.945	0.945	0.727	10
DecisionStump	94.59%	0.946	0.946	0.927	0.946	0.794	20
REPTree	95.03%	0.950	0.556	0.945	0.950	0.893	20
RandomForest	96.08%	0.961	0.512	0.956	0.961	0.966	310

Now the performance of all the classifier will be discussed from the table 4.2 which are generated with the help of Weka data mining tool. Performing each classifier on the described data set section 4.1 we can find the accuracy, TP rate, RP rate, precision, Recall, ROC area and total time to build model. Now time to describe them individually through visual form as well brief description to better understanding of each classifier in different perspective and their experimental results.

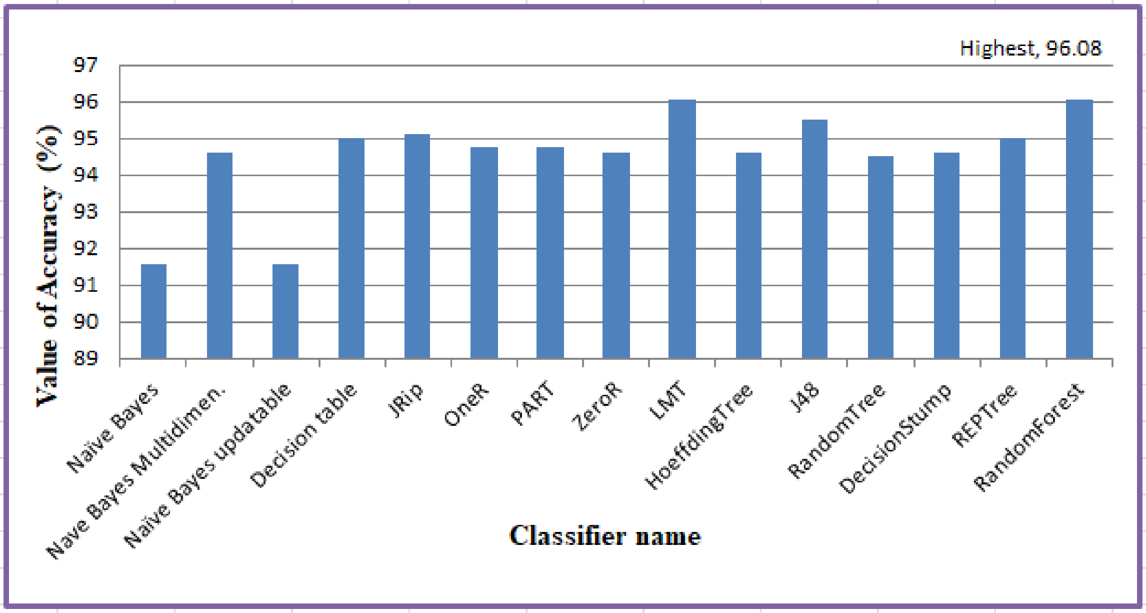


Figure 4.2: Comparison of Accuracy for each classifier

Accuracy describes that how often is the classifier correct for considering overall true positive and false positive for all instances. From the comparison table it is clearly shows that different classifier have different accuracy though applied them on the same instances whereas lowest accuracy is 91.55 for Naïve Bayes, Naïve Bayes updatable. Average accuracy is 94.53 for J48, Random tree and the highest accuracy is for Random forest. Other classifier accuracy falls in between the lowest and average or in between the average and highest accuracy.

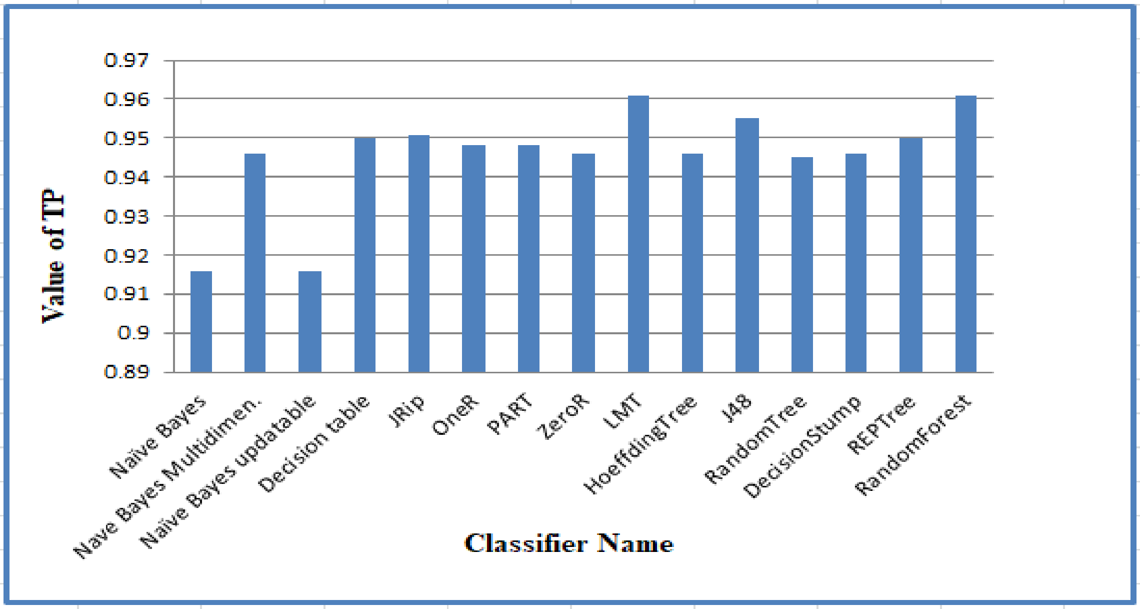


Figure 4.3: Comparison of True Positive (TP) rate for each classifier

In the above figure we see the true positive rate of each classifier which means that when actually a student is critical how often the model accurately predicted that also critical means yes. That's the way of measure TP rate of a classifier. Before moving forward, look on one thing that is LMT and Random forest had the highest accuracy which also lead the highest true positive rate.

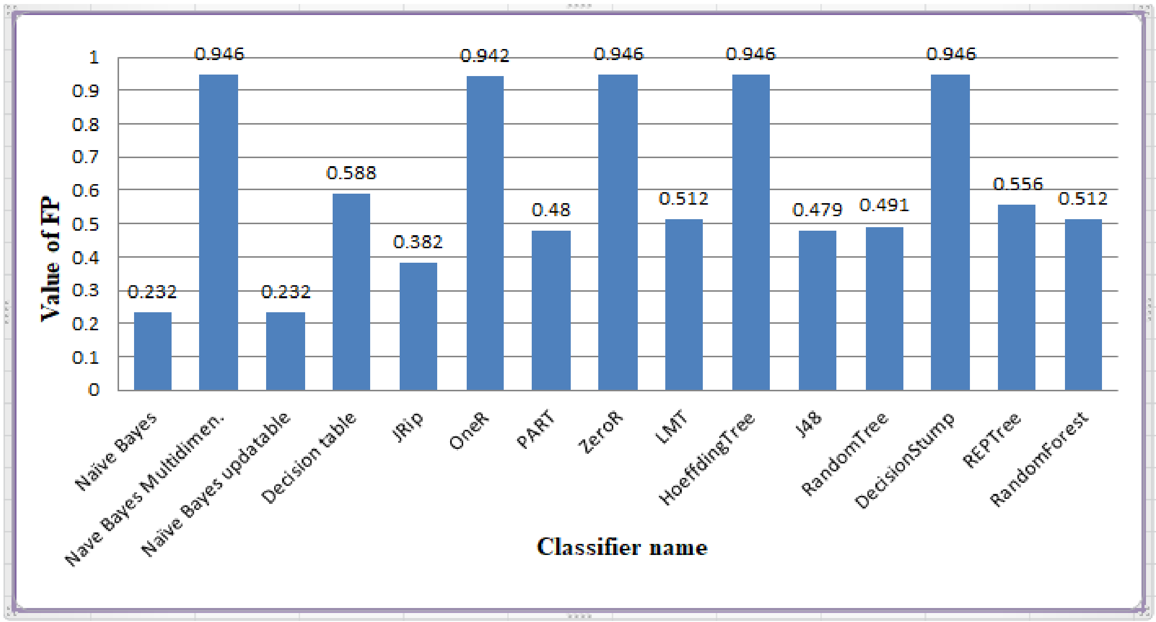


Figure 4.4: Comparison of False Positive (FP) rate for each classifier

In this figure we see the false positive rate of each classifier. False positive is type I error. which means that when actually a student is critical but the model inaccurately predicted no. lowest value of FP lead Naïve Bayes and highest value of FP lead couple of classifier Naïve Bayes multidimensional, ZeroR, and DecisionStump.

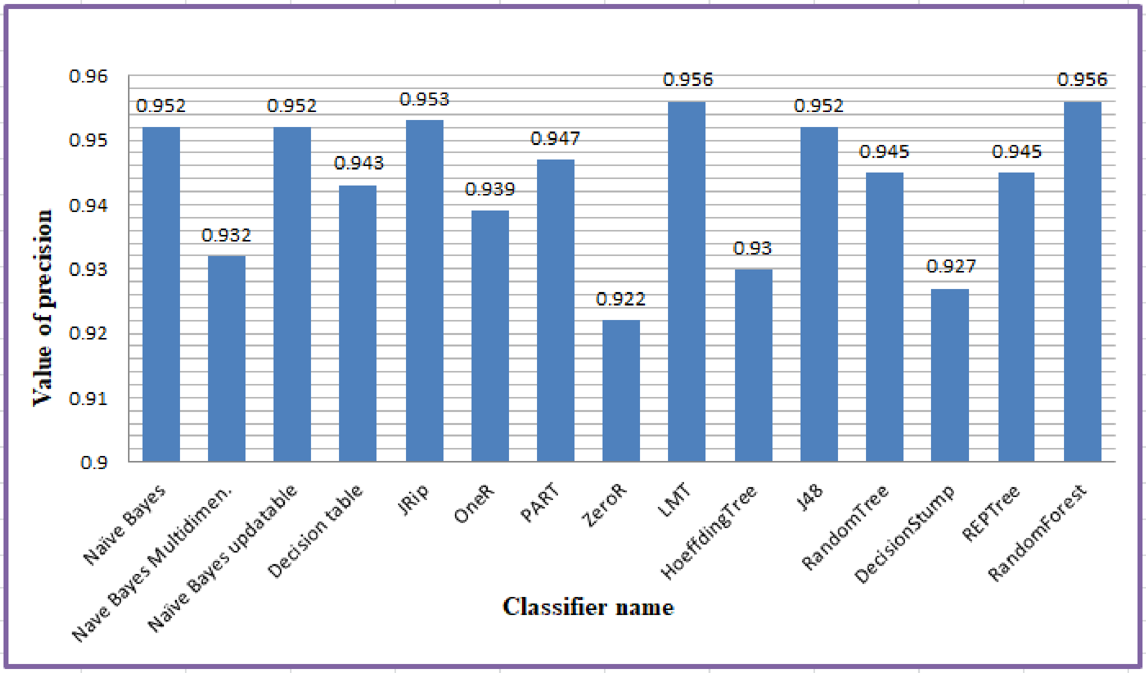


Figure 4.5: Comparison of Precision for each classifier

Precision describes that how often the classifier correct for considering only the true positive means predicted yes. Precision value should be highest as much possible. In comparison table highest precision value is 0.956 for several classifier, others are all most close to the highest precision value. Value of precision and recall is needed to calculate the F-measure. Sometimes it is difficult we have lowest precision and highest recall values or vice-versa, closer value of them is impressively preferable to calculate f-measure.

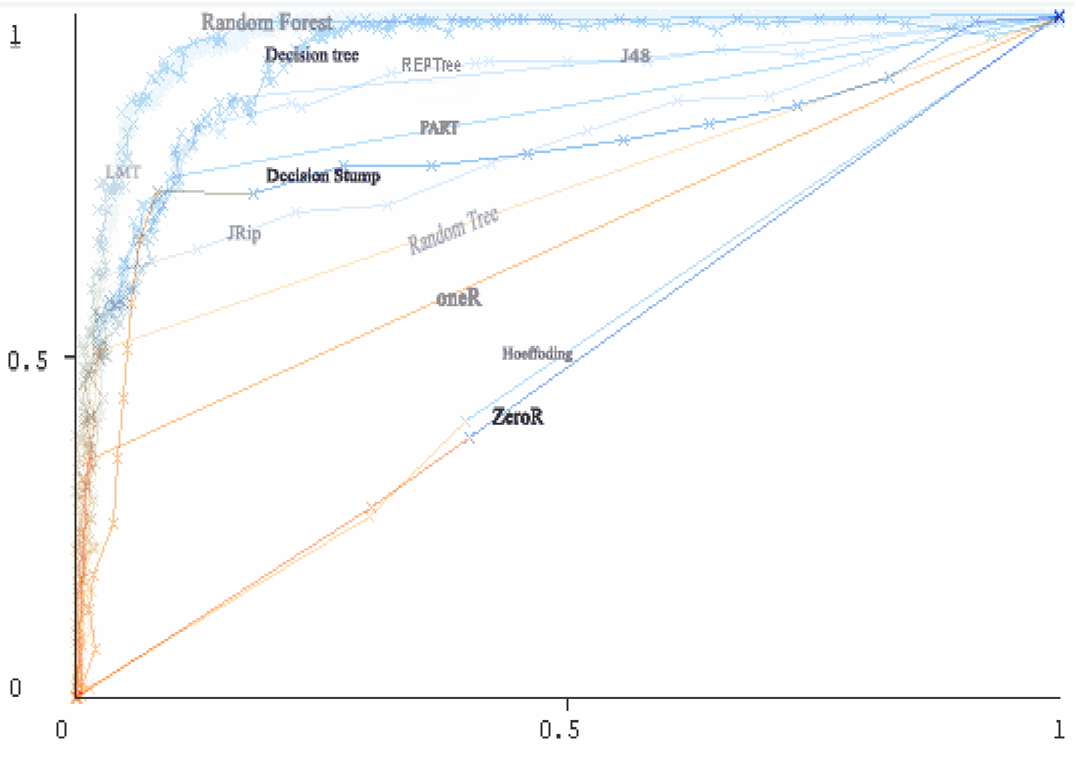


Figure 4.6: ROC area curve

In the above ROC (receiver operating characteristic curve) curve graph shows the performance of a classification model at all classification thresholds.

By the using the threshold value we might compute TPR and FPR for the threshold equal to 0.5, we apply the model to each example, get the score, and, if the score if higher than or equal to 0.5, we might predict the positive class. On the other hand we can predict the negative class. Though we described the TPR and FPR distinctly in the above section visually and descriptively we can also plot them by using ROC curve.

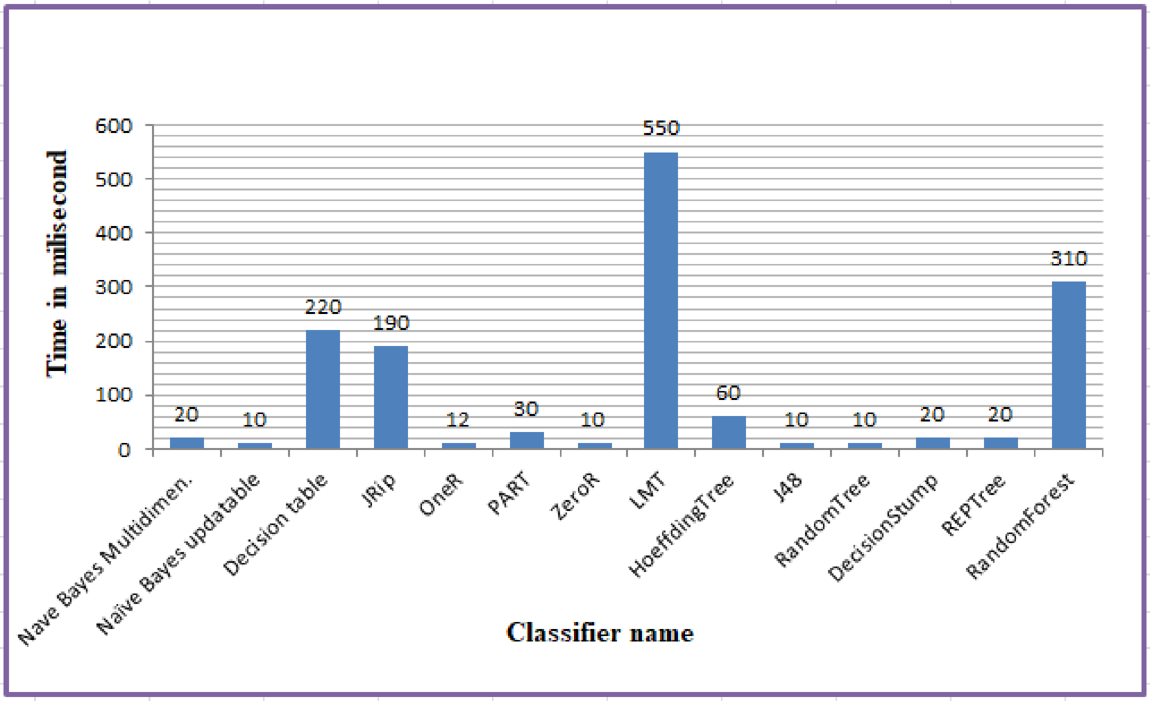


Figure 4.7: Comparison of Time complexity for each classifier

Another thing which is most important for any computation experiment is time complexity means time taken to functionally fully complete the assigned model. That is shown in the unit as millisecond (ms). The lowest time taken 10 ms for random tree, J48, Naïve Bayes updatable on the other side highest time taken 550 ms for LMT.

4.3 Summary

Almost of all the classifiers perform quite well by giving impressive accuracy and effectively classified the instances. But still there are rooms for improvement by tweaking some parameters and increasing the quantity of the data sets. The computation time for this also big issue highly computing machine is best for less time complexity. Though we are not looking for the best as well as that is quite difficult too because accuracy may vary in different types of data and data set. But at present situation if we look on the experimental analysis table highest accuracy rate is 96.08 percent for Random forest but its time complexity is high on the other side accuracy of Random Tree is near to highest 94.53 as well as taken time to build model only 10ms. Finally on the basis of couple of parameters Random Tree performed well in this research other than all.

Chapter 5: Conclusion, Recommendation and Implication for Future Research

5.1 Conclusion

One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for prediction of critical student's management of a reputed private university in different level and term. This research model is successfully identifying the students who are likely to dropout or not. These students can be considered for proper counseling so as to improve their result and come back to their regular track. This finding is a preliminary research in this area and we think it is a good starting point for researchers in the region to establish a research track related to using data mining to enhance college/university education. Finally we can say, in this dissertation we will explain the contributions we have made the fields of EDM for the students, educational institution and educators also.

5.2 Recommendations

For identifying critical students Educational data mining can response number of queries from the adopted model and attained from student data such as-

- ✓ Are the students at risk.
- ✓ Chances of placement of student.
- ✓ The students who likely to drop the course or degree.
- ✓ Quality level of student in the particular domain.
- ✓ To attract students which course should have to offer by educational institution.

5.3 Implication for Further Study

The application of EDM (Educational Data mining) in educational performance has been a true hope for an intelligent future. The limited resources and traditional prediction way we have cannot ensure the sustainable future of our education sector. This research should be further enhanced as a future work by considering data from several other universities including private and public. Collect more instances to build the model. Other attributes could also be added to the data set for further enhancing the generated model. Furthermore, some other classification models could be tested in this domain.

REFERENCES

[1] About educational data mining http://www.educationaldatamining.org, Last accessed on 06 April, 2018.

[2] Mohammed M.Abu Tair,Alaa M.El-Hales," Mining Educational Data to Improve Student"s Performance: A Case study", International Journal of Information and Communication Technology Research(ICT Jounal), 2012.

[3] Surjeet Kumar Yadav, Brijesh Bharadwaj,Saurabh Pal, "Data Mining Applications:A Comparative study for Predicting Students Performance," International Journal of Innoviative Technology & Creative Engineering, 2011.

[4] Nguyen N., Paul J., and Peter H., A Comparative Analysis of Techniques for Predicting Academic Performance. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12, 2007.

[5] Al-Radaideh Q., Al-Shawakfa E., and AI-Najjar M., Mining Student Data using Decision Trees, In Proceedings of the International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[6] Cesar V., Javier B., liela S., and Alvaro O., Recommendation in Higher Education Using Data Mining Techniques, In Proceedings of the Educational Data Mining Conference, 2009.

[7] Muslihah W., Yuhanim Y., Norshahriah W., Mohd Rizal M., Nor Fatimah A., and Hoo Y. S., Predicting NDUM Student's Academic Performance Using Data Mining Techniques, In Proceedings of the Second International Conference on Computer and Electrical Engineering, IEEE computer society, 2009.

[8] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[9] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[10] U. K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN: 0975-9646, 2011.

[11] Ramaswami M., and Bhaskaran R., CHAID Based Performance Prediction Model in Educational Data Mining, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, 2010.

[12] Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) 'Student Relationship in Higher Education Using Data Mining Techniques', Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.

[13] There is a huge amount of data, htps://www.tutorialspoint.com/data_mining/dm_quick_guide.htm, last accessed on 23 September 2019.

[14] Data mining is widely used, https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm, last accessed on 23 September 2018.

[15] Data mining is one of, https://www.flatworldsolutions.com/data-management/articles/data-mining-future-trends.php, last accessed on 10 January 2019.

[16] There are several major, http://www.zentut.com/data-mining/data-mining-techniques/, last accessed on 10 January 2019.