Predicting Student Performance Using Data Mining Methods

Zareen Rahman (ID: 15-98635-2)

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science

Department of Computer Science
Faculty of Science and Information Technology
American International University - Bangladesh (AIUB)

December 2017

Declaration

I declare that this thesis is my original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Zareen Rahman

ID: 15-98635-2

Acknowledgements

First of all, I express my gratefulness to the Almighty Allah, without His divine blessing it would not be possible for me to complete this thesis. I sincerely like to thank my honourable thesis supervisor Mr Abdus Salam, Assistant Professor, Department of Computer Science, AIUB for his great inspiration and proper guidance throughout the whole thesis work. I am also grateful to all of my respected teachers without whom I would never reach at this educational stage. I am grateful for their time and ideas and appreciate their availability whenever I turned to them. I extend my thanks to Dr. Khandaker Tabin Hasan for his kind help and continuous inspiration. I would also like to show my gratitude to my parents for being so understanding.

Abstract

Data mining is widely used in educational field to find the problems arise in this field. Student performance is of great concern in the educational institutes where several factors may affect the performance. For prediction, the three required components are: Parameters which affect the student performance, Data mining methods and third one is data mining tool. These parameters may be psychological, personal, and environmental. Different methods and techniques of data mining were compared during the prediction of students' success, applying the data collected from the surveys conducted during the summer semester at American International University-Bangladesh, the Department of Computer Science, academic year 2017-2018, among the students of Theory of Computing and Compiler Design. The success was evaluated with the passing grade at the exam. The impact of students' socio-demographic variables, previous academic records, achieved and attitudes towards studying which can have an effect on success were all investigated. In this paper, prediction of student performance is done by applying Naïve Bayes, Multilayer Perceptron and J48 Decision Tree classification techniques using WEKA tool. By applying data mining techniques on student data, we can obtain knowledge which describes the student performance. This knowledge will help to improve the education quality, student's performance and to decrease failure rate. All these will help to improve the quality of institute. The experimental result shows the Multilayer Perceptron is the best model among the other techniques by receiving the highest accuracy value of 92%. The extracted knowledge from prediction model will be used to identify and profile the student to determine the students' level of success in the semester.

Table of Contents

##

##

##

##

List of Figures

List of Tables

Chapter 1

Introduction

1.1 INTRODUCTION

Data Mining (DM), is an approach to discover useful information from large amount of data. Data mining techniques apply various methods in order to discover and extract patterns from stored data. The pattern found will be used to solve a number of problems occurred in many fields such as education, economic, business, statistics, medicine, and sport. The large volume of data stored in those areas demands for DM approach because the resulting analysis is much more precise and accurate.

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Data mining is generally an iterative and interactive discovery process. The goal of this process is to mine patterns, associations, changes, anomalies, and statistically significant structures from large amount of data. Furthermore, the mined results should be valid, novel, useful, and understandable.

These "qualities" that are placed on the process and outcome of data mining are important for a number of reasons, and can be described as follows:

(1) Valid: It is crucial that the patterns, rules, and models that are discovered are valid not only in the data samples already examined, but are generalizable and remain valid in future new data samples. Only then can the rules and models obtained be considered meaningful.

(2) Novel: It is desirable that the patterns, rules, and models that are discovered are not already known to experts. Otherwise, they would yield very little new understanding of the data samples and the problem at hand.

(3) Useful: It is desirable that the patterns, rules, and models that are discovered allow us to take some useful action. For example, they allow us to make reliable predictions on future events.

(4) Understandable: It is desirable that the patterns, rules, and models that are discovered lead to new insight on the data samples and the problem being analyzed.

In fact, the goals of data mining are often that of achieving reliable prediction and/or that of achieving understandable description. The former answers the question “what”, while the latter the question “why”. With respect to the goal of reliable prediction, the key criteria are that of accuracy of the model in making predictions on the problem being analyzed. How the prediction decision is arrived at may not be important. With respect to the goal of understandable description, they key criteria is that of clarity and simplicity of the model describing the problem being analyzed. There is sometimes a dichotomy between these two aspects of data mining in the sense that the most accurate prediction model for a problem may not be easily understandable, and the most easily understandable model may not be highly accurate in its predictions. For example, on many analysis and prediction problems, support vector machines are reported to hold world records in accuracy [1]. However, the maximum error margin models constructed by these machines and the quadratic programming solution process of these machines are not readily understood to the non-specialists. In contrast, the decision trees constructed by tree induction classifiers such as C4.5 are readily grasped by non-specialists, even though these decision trees do not always give the most accurate predictions.

1.2 PROBLEM STATEMENT

The amount of data in educational environment maintained in electronic format has seen a dramatic increase in recent time. The data can be collected from historical and operational data reside in the databases of educational institutes. The task to manage the large amount of data and determine the relationships among variables in the data is not easy to be done.

The face value assessment of students at the point of entry can only be confirmed or dispelled by the dynamic follow-up monitoring of students’ performance during the course of study leading to serve as an indicator of the suitability and unsuitability of students before admission and during their course of study. Performance predicate is dependent upon motivation, attitudes, peer influence, curriculum and by the continued real-time monitoring of student’s performance using a simple rapid response system and as noted predicts correctly which student may need some attention or reinforcements in the course of their education [2].

The research studies revealed that various factors are responsible for scholastic failure of students, such as low socio-economic background, student’s cognitive abilities, school related factors, environment of the home, or the support given by the parents and other family members [3].

In present day’s educational system, a student’s performance in any universities is determined by the combination of internal assessment and external mark. An internal assessment is carried out by the teachers upon the student’s performance in various evaluation methods such as tests, assignments, and seminar, attendance, and extension activities. An external mark is the one that is scored by the student in semester examination. Each student has to get the minimum pass mark in internal and as well as in external examination. The current educational system does not involve any prediction about pass or fail percentage based on performance. The system does not deal with dropouts. There is no efficient method to caution the students about the deficiency in attendance. It does not identify the weak student and inform the teachers. The most likely place where data miners may initiate data mining project in higher education area is Institutional Effectiveness. Thus, some of the research questions that arise in higher education data analysis can be stated as follows:

(i) What variables or combination of variables collected can be used as predictors of students' performance final grade?

(ii) How the discovered knowledge from academic data can aid decision makers to improve decision making processes?

(iii) How to apply the kernel methods in constructing the model of student performance predictors?

1.3 RESEARCH OBJECTIVES

The main objectives in this study are to apply the data mining techniques in educational data; Naïve Bayes, Multilayer Perceptron and Decision Tree, all of the techniques will be used to construct the model of student's performance predictors based on factors which are influencing the academic performance of the students for a better approach in future. The detail of research objectives in this study as follows:

(i) To apply data mining methods: Naïve Bayes, Multilayer Perceptron and Decision Tree algorithm for predicting student performance.

(ii) To create a model of student performance predictors that can be used as decision support system002E

1.4 RESEARCH SCOPE

The scope of this research is to create a students' performance prediction model by using different socioeconomic indicators and student behavior as variable predictors and to build the classification model that classifies the performance of the undergraduate students of American International University Bangladesh (AIUB). The sample data of this research come from student academic databases and the surveyed intrinsic motivation and behavioral of undergraduate students in summer semester 2017, from two course students that are Theory of Computing and Compiler Design.

1.5 RESEARCH CHALLENGES

Some of the research challenges for this analysis can be seen in future are issues such as:

(1) Scalability: How does a data mining calculation perform if the dataset has expanded in volume and in measurements? This may require a few developments in light of effective and adequate inspecting, or an exchange off between in-memory versus plate based handling, or an approach in light of elite dispersed or parallel processing.

(2) Automation: While a data mining algorithm and its output may be readily handled by a computer genius, it is important to realize that its ultimate user is often not the developer. In order for a data mining tool to be directly usable by the ultimate user, issues of automation— especially in the sense of ease of use—must be addressed. Even for the computer genius, the use and incorporation of past knowledge into a data mining algorithm is often a hard challenge; (s)he too would like to get if data mining algorithms can be modularized in a way that facilitate the exploitation of prior knowledge.

1.6 CONCLUSION

Data mining has significance with respect to finding the examples, determining, and revelation of learning and so on. In various business areas. Data mining procedures and calculations, for example, grouping, bunching and so forth helps in finding the examples to choose the future patterns in organizations to develop. Data mining has wide application space nearly in each industry where the information is produced that is the reason data mining is viewed as a standout amongst the most imperative boondocks in database and data frameworks and a standout amongst the most encouraging interdisciplinary advancements in Information Technology.

Chapter 2

Background Study

2.1 INTRODUCTION

Education is the backbone of a nation. Bangladesh is one of the most densely populated countries in the world. There are 168,957,745 people living in Bangladesh (World Factbook, 2015). Higher education facilities are not sufficient for the students in Bangladesh. Higher educations in universities of Bangladesh are divided into two sectors namely public universities and private universities. According to the University Grant Commission (UGC) of Bangladesh, there are 37 public universities and 85 private universities in Bangladesh and 10 more universities are in process to open academic activities. Initially the private universities were established under the provision of Private University Act 1992 which was amended in 1998. The Private University Act 2010 was introduced after abolishing the previous acts. According to the annual report 2014 of UGC of Bangladesh, almost 400000 students are studying in 80 private universities. Naturally it is an interest to know the status of the students of these universities. It may also be an interest to know the factors which are influencing the results of these students [4].

Educational data has become a vital resource in this modern era, contributing much to the welfare of the society. Educational institutions are becoming more competitive because of the number of institutions growing rapidly. To stay afloat, these institutions are focusing more on improving various aspects and one important factor among them is quality learning. For providing quality education and to face new challenges, the institutions need to know about their potentials which are explicitly seen and which are hidden. The truths behind today’s educational institutions are a substantial amount of knowledge is hidden. To be competitive, the institutions should identify their own potentials hidden and implement a technique to bring it out.

In recent years, Educational Data Mining has put on a mammoth recognition within the research realm as it has become a vital need for the academic institutions to improve the quality of education. The higher education institutions has potential knowledge such as academic performance of students, administrative accounts, potential knowledge of the faculty, demographic details of the students and many other information in a hidden form. The technique behind the extraction of the hidden knowledge is Knowledge Discovery process. Recently Data mining is widely used on educational dataset.

2.2 RELATED WORK

To know the academic performance of the students, results and socio demographic characteristics of students may be used as tools. Three supervised data mining algorithms, i.e. Bayesian, Decision trees and Neural Networks which were applied by Osmanbegovic E., Suljic M. [1] on the preoperative assessment data to predict success in a course (to produce result as either passed or failed) and the performance of the learning methods were evaluated based on their predictive accuracy, ease of learning and user friendly characteristics. The researchers observed that that this methodology can be used to help students and teachers to improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of learning.

Suman Biswas, Mst. Shonamoty Khatun, Md. Yasin Ali Parh & Dr. Md. Sazzad Hossain [4] found in their study that the factors like SSC and HSC results of the student, parental academic qualification, higher family income, residential in hall, student's class attendance, study time without class period had a positive impact and the factors like students' internet use for non-academic purpose, political status, mobile phone using for non-academic purpose in the University had a negative impact of students on academic results. It was seen that 51.5% of the respondents was male and 48.6% was female. It was also seen that 70% of the respondent was Muslim and 30% was Hindu. It was found that 24.3% of the respondents came from urban area and 75.7% came from rural area. It was also found that father's occupation of 31.4%, 31.4% and 37.1% of the respondents were job, business and farmer respectively. It was revealed that mother's occupation of 91.4% of the respondents was housewife and 8.6% was job.

Easmin et al. [5] observed that mother's education had significance effect on the academic performance of the students. However, the parental educational qualification of the students was identified to have statistical significant effect on the academic performance of the students. It was also found that 66.3% of the respondents were male and 33.7% of the respondent was female.

Alam et al. [6] reported that over all CGPA of IIUC (International Islamic University Chittagong) students were 3.25 (Out of 4.00). It was also found that the regression coefficient of academic performance on its correlates varied from faculty to faculty. The results of their study revealed that age, gender, past academic track, medium of education and absence in the classes also influenced the academic performance of a student. It was seen that the result (CGPA) of 26.5% of the students was less than 3, 37.5% was 3.01-3.50, 18.5% was 3.51-3.75 and 17.5% was above 3.75. It was found that 61.0% of the respondents were male and 39.0% were female. It was also found that 95.5% of the respondents were Muslim and 4.5% were Non-Muslim. It was observed that the father's education level of 10%, 20% and 70% of the respondents were primary, secondary and higher respectively. It was also observed that the mother's education level of 22%, 41% and 36% of the respondents were primary, secondary and higher respectively. It was revealed that the father's occupation of 4.0%, 37%, 34% and 11.5% of the respondents were agriculture, business, service and teaching respectively. It was also revealed that the mother's occupation of 91.5%, 0.5%, 2.5% and 5% of the respondents was home maker, business, service and teaching respectively.

Rahman [7] found that the religion of 88.93%, 07.07%, 02.00% and 01.60% of the respondents were Islam, Hindu, Christian and Buddhist respectively. It was also investigated that 39.2%, 37.07% and 13.60% of the students were science group, commerce group and arts group respectively in HSC level.

Mutairi [8] identified that female students performed better academic performance than male counterparts. Ali et al. (2009) found that 62.2% of the respondents was female and 37.8% was male. It was also found that 11.2% of the students' fathers' highest educational level was at primary level, 61% at secondary level while 27.8% was at tertiary level. It was also seen that 12% of the students' mothers' highest educational level was at primary level, 68.7% at secondary level while 19.3% was at tertiary level.

Alaa M. El-Halees et al. [9], proposed a case study that Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. They used educational data mining to improve graduate students' performance, and overcome the problem of low grades of graduate students. In their case study they tried to extract useful knowledge from graduate students data collected from the college of Science and Technology – Khanyounis. The data include fifteen years period [1993-2007]. After preprocessing the data, they applied data mining techniques to discover association, classification, and clustering and outlier detection rules. In each of these four tasks, they presented the extracted knowledge and describe its importance in educational domain.

Furthermore, Bekele, R., Menzel, W. [10] observed that in the problem of prediction of performance, it is possible to automatically predict students‟ performance. Moreover by using extensible classification formalism such as Bayesian networks, which was employed in their research it becomes possible to easily and uniformly integrate such knowledge into the learning task. The researchers‟ experiments also show the need for methods aimed at predicting performance and exploring more learning algorithms.

Bayesian classification method was also used by Bhardwaj, K., Pal, S [11] in their work on student database to predict the students‟ grades on the basis of previous year performance. The researchers concluded that the study will help the students and the teachers to improve the grades of the student. The study also helps to identify those students which needed special attention to reduce failing ratio and taking appropriate action at right time.

Cortez P, Silva A. [12] addressed the prediction of secondary school students‟ performance in two core subjects of mathematics and Portuguese by using their past score in the previous session and other demographic factors and employed four data mining methods of Decision trees, Random Forests, neural networks and Support Vector machines approach. The results show that the prediction was achievable provided the grades of the previous session were known. This confirms that the prediction of students‟ performance is premised on past performance and hence shows that a student's performance is closely related to the performance in previous course (most likely a prerequisite course).

Mladen D., Mirjana P. B., Vanja Š. [13] described the process of knowledge discovery from databases using a practical example of a current actual problem. They developed two models based on decision tree which were successfully used to predict student success based on GPA criterion and time student needs to finish the undergraduate program (time-to-degree) criterion.

Another study undertaken by Sembiring S, Zarlis, M, Hartama, D. Ramliana S, Elvi W [14] showed that Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. The study further showed how useful data mining can be in higher education particularly to predict the final performance of student. The researchers collected data from student by using questionnaire to find the relationships between behavioral attitude of student and their academic performance. Data mining techniques were then applied. They obtained the prediction rule model using decision tree as well as implementing the rules into Support Vector Machine (SVM) algorithm to predict the students‟ final grade. Also the students were clustered into groups using kernel k-means clustering. The study expressed the strong correlation between mental condition of student and their final academic performance.

Shannaq et al. [15], applied the classification as data mining technique to predict the numbers of enrolled students by evaluating academic data from enrolled students to study the main attributes that may affect the students' loyalty (number of enrolled students). The extracted classification rules are based on the decision tree as a classification method, the extracted classification rules are studied and evaluated using different evaluation methods. It allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support.

Surjeet K, Yadav, Bharadwaj, B. Pal B. [20] concluded that 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 Student data to predict the student‟s performance in the end semester examination. The experimental results show that Classification and Regression Tree (CART) CART is the best algorithm for classification of data.

Al-Radaideh et al. [22], applied the data mining techniques, particularly classification to help in improving the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. The extracted classification rules are based on the decision tree as a classification method; the extracted classification rules are studied and evaluated. It allows students to predict the final grade in a course under study.

Baradwaj and Pal [24], applied the classification as data mining technique to evaluate student' performance, they used decision tree method for classification. The goal of their study is to extract knowledge that describes students' performance in end semester examination. They used students' data from the student' previous database including attendance, class test, Seminar and Assignment marks. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.

Chandra and Nandhini [26], applied the association rule mining analysis based on students' failed courses to identify students' failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. The extracted association rules reveal some hidden patterns of students' failed courses which could serve as a foundation stone for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate.

Ayesha et al. [24], used k-means clustering algorithm as a data mining technique to predict students' learning activities in a students' database including class quizzes, mid and final exam and assignments. This correlated information will be conveyed to the class teacher before the conduction of final exam. This study helps the teachers to reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

2.3 EDUCATIONAL DATA MINING (EDM) PROCESS

Educational Data mining has become a very useful research area for all the people who is looking to explore this sector. Data mining helps to extract the knowledge from available dataset and should be created as knowledge intelligence for the benefit of the institution. Higher education does categorize the students by their academic performance. Many factors influence the academic performance of the student. The model is mainly focused on exploring various indicators that have an effect on the academic performance of the students. The extracted information that analyzes student performance can be stored as intelligent knowledge for decision maker to improve the quality of education in institutions. The knowledge stored is used for predicting the student's performance in advance.

In recent years, there has been increasing interest in the use of DM to investigate educational field. Educational Data Mining (EDM) is concerned with developing methods and analyzing educational content to enable better understanding of students' performance. It is also important to enhance teaching and learning process. The data can be collected form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Educational Data Mining use many techniques such as Decision Trees, Naïve Bayes, Neural Networks, Naïve Bayes, K-Nearest neighbor, and many others.

Prediction models that include all personal, social, psychological and other environmental variables are necessitated for the effective prediction of the performance of the students. The prediction of student result with perfect accuracy is useful for identify the students with low academic achievements partially. It is required that the identified students can be assisted more by the teacher so that their performance is improved in future. The general data mining process is depicted in Figure 2-1 [18].

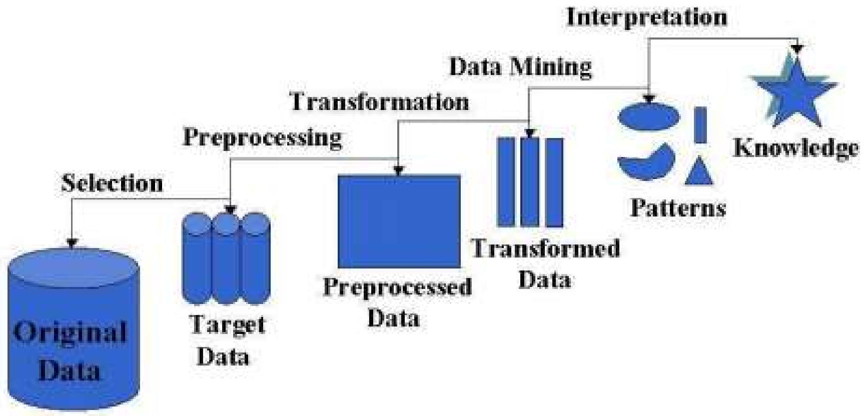


Fig 2-1: The Data Mining Process

It comprises the following steps some of which are optional depending on the problem being analyzed:

(1) Understand the application domain: An appropriate comprehension of the application area is important to welcome the data mining results wanted by the client. It is likewise essential to acclimatize and exploit accessible earlier learning to expand the possibility of achievement.

(2) Collect and make the objective dataset: Data mining depends on the accessibility of appropriate information that mirrors the hidden decent variety, request, and structure of the issue being examined. In this way, the gathering of a dataset that catches all the conceivable circumstances that are pertinent to the issue being investigated is critical.

(3) Clean and change the objective dataset: Raw information contain numerous mistakes and irregularities, for example, clamor, exceptions, and missing esteems. A critical component of this procedure is the de-duplication of information records to deliver a non-excess dataset. For instance, in gathering data from open succession databases for the expectation of protein interpretation start destinations, a similar arrangement might be recorded different circumstances in the general population grouping databases.

(4) Select highlights, diminish measurements: Even after the information have been tidied up as far as killing copies, irregularities, missing esteems, et cetera, there may at present be commotion that is immaterial to the issue being broke down. These clamor characteristics may confound resulting information mining steps, deliver insignificant guidelines and affiliations, and increment computational cost. It is in this way insightful to play out a measurement lessening or highlight choice advance to isolate those qualities that are appropriate from those that are unimportant.

(5) Apply information mining calculations: Now we are prepared to apply fitting information mining calculations—affiliation rules revelation, succession mining, grouping tree acceptance, bunching, et cetera—to investigate the information. Some of these calculations are introduced in later segments.

(6) Interpret, assess, and envision designs: After the calculations above have created their yield, it is as yet important to inspect the yield with a specific end goal to decipher and assess the different examples, principles, and models. It is just by this elucidation and assessment process that we can determine new bits of knowledge on the issue being investigated. As illustrated over, the information mining attempt includes many advances. Besides, these means require innovations from different fields. Specifically, techniques and thoughts from machine learning, insights, database frameworks, information warehousing, superior registering, and representation all have critical parts to play. In this instructional exercise, we talk about basically information mining systems applicable to Step (5) above.

2.4 TECHNIQUES USED IN DATA MINING

2.4.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination [21].

The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

2.4.2 Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality [21].

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

2.4.3 Prediction

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models [23].

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression.

2.4.4 Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value [25].

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

2.4.5 Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs [25].

Types of neural networks

- Back Propagation

2.5 APPLICATIONS OF DATA MINING IN EDUCATION SECTOR

In Higher Education There are many application areas of data mining like customer analytics, Agriculture, banking, Security Applications, Educational data mining, Mass surveillance, Privacy preserving etc. The main concerned area is about data mining applications in educational systems. Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. A key area of EDM is mining student’s performance. Another key area is mining enrollment data. Key uses of EDM include predicting student performance and studying learning in order to recommend improvements to current educational practice. EDM can be considered one of the learning sciences, as well as an area of data mining. The main applications of EDM are listed as follows:

2.5.1 Analysis and Visualization of Data

It is used to highlight useful information and support decision making. In the educational environment, for example, it can help educators and course administrators to analyze the students’ course activities and usage information to get a general view of a student’s learning. Statistics and visualization information are the two main techniques that have been most widely used for this task. Statistics is a mathematical science concerning the collection, analysis, interpretation or explanation, and presentation of data. It is relatively easy to get basic descriptive statistics from statistical software, such as SPSS. Statistical analysis of educational data (logs files/databases) can tell us things such as where students enter and exit, the most popular pages students browse, number of downloads of e-learning resources, number of different pages browsed and total time for browsing different pages. It also provides knowledge about usage summaries and reports on weekly and monthly user trends, amount of material students might go through and the order in which students study topics, patterns of studying activity, timing and sequencing of events, and the content analysis of students notes and summaries. Statistical analysis is also very useful to obtain reports assessing how many minutes student worked, number of problems he resolved and his correct percentage along with our prediction about his score and performance level. Visualization uses graphic techniques to help people to understand and analyze data. There are several studies oriented toward visualizing different educational data such as patterns of annual, seasonal, daily and hourly user behavior on online forums. Some of such investigations are statistical graphs to analyze assignments complement, questions admitted, exam score, student tracking data to analyze student's attendance, results on assignments and quizzes, weekly information regarding students and group's activities.

2.5.2 Predicting Student Performance

In this case, we estimate the unknown value of a variable that describes the student. In education, the values normally predicted are student's performance, their knowledge, score, or marks. This value can be numerical/continuous (regression task) or categorical/discrete (classification task). Regression analysis is used to find relation between a dependent variable and one or more independent variables. Classification is used to group individual items based upon quantitative characteristics inherent in the items or on training set of previously labeled items. Prediction of a student's performance is the most popular applications of DM in education. Different techniques and models are applied like neural networks, Bayesian networks, rule based systems, regression, and correlation analysis to analyze educational data. This analysis helps us to predict student's performance i.e. to predict about his success in a course and to predict about his final grade based on features extracted from logged data. Different types of rule-based systems have been applied to predict student's performance (mark prediction) in a learning environment (using fuzzy-association rules). Several regression techniques are used to predict student's marks like linear regression for predicting student's academic performance, stepwise linear regression for predicting time to be spent on a learning page, multiple linear regressions for identifying variables that could predict success in colleges' courses and for predicting exam results in distance education courses.

2.5.3 Outlier Analysis

According to Ogunde A.O., Ajibade D.A [19] Outlier can be defined as "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". Outlier detection has been used to detect and, where appropriate, remove anomalous observations from data. Outlier detection can identify system faults and fraud before they escalate with potentially catastrophic consequences.

There are three fundamental approaches for outlier detection.

- Type 1 - Determine the outliers with no prior knowledge of the data. This is essentially a learning approach analogous to unsupervised clustering. The approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers.

- Type 2 - Model both normality and abnormality. This approach is analogous to supervised classification and requires pre-labeled data, tagged as normal or abnormal.

- Type 3 - Model only normality (or in a few cases model abnormality). This is analogous to a semi supervised recognition or detection task. It may be considered semi-supervised as the normal class is taught but the algorithm learns to recognize abnormality.

2.5.4 Grouping Students

In this case groups of students [16] are created according to their customized features, personal characteristics, etc. These clusters/groups of students can be used by the instructor/developer to build a personalized learning system which can promote effective group learning. The DM techniques used in this task are classification and clustering. Different clustering algorithms that are used to group students are hierarchical agglomerative clustering, K-means and model-based clustering. A clustering algorithm is based on large generalized sequences which help to find groups of students with similar learning characteristics like hierarchical clustering algorithm which are used in intelligent e-learning systems to group students according to their individual learning style preferences.

2.5.5 Planning and Scheduling

Planning and scheduling is used to enhance the traditional educational process by planning future courses, course scheduling, planning resource allocation which helps in the admission and counseling processes, developing curriculum, etc. Different DM techniques used for this task are classification, categorization, estimation, and visualization. The main objective of using above techniques is academic planning, predicting alumni pledges and creating meaningful learning outcome typologies. Decision trees, link analysis and decision forests have been used in course planning to analyze enrollee's course preferences and course completion rates in extension education courses. Classification, prediction, association-rule analysis, clustering, etc have been compared to discover new explicit knowledge that could be useful in the decision-making process in higher learning institutions. Educational training courses have been planned through the use of cluster analysis, decision trees, and back-propagation neural networks in order to find the correlation between the course classifications of educational training. Decision trees and Bayesian models have been proposed to help management institutes to explore the probable effects of changes in recruitments, admissions and courses. [35]

2.5.6 Target Marketing

Consider a sample universe u, which represents a sample candidate set. The DM algorithm generates T called target set. It is used by marketing agent to organize promotion and marketing campaigns. The universe u consists of a database of various students' attributes like name, academic profile, region etc. By using various attributes, u is successively portioned to get t1, t2 etc. which are target sets based in the corresponding partition algorithms.

Fig 2-2 [28] shows flow of above mentioned algorithm. An optimal portioned can also be generated in this process based upon some condition.

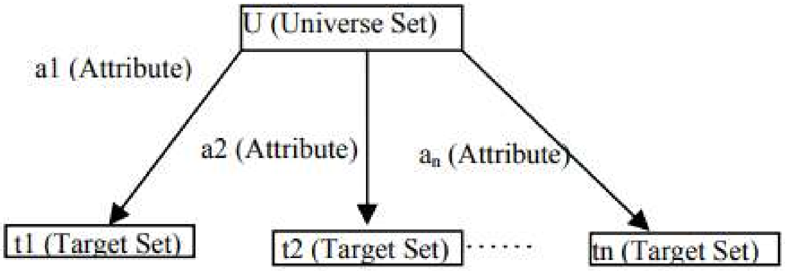


Fig 2-2: Target Marketing Based Upon Various Students Attributes

2.5.7 Enrollment Management

This term is frequently used in higher education to describe well-planned strategies and tactics to shape the enrollment of an institution [17] and meet established goals. Enrollment management is an organizational concept and a systematic set of activities designed to enable educational institutions to exert more influence over their student enrollments. Such practices often include marketing, admission policies, retention programs, and financial aid awarding. Strategies and tactics are informed by collection, analysis, and use of data to project successful outcomes. Activities that produce measurable improvements in yields are continued and/or expanded, while those activities that do not are discontinued or restructured. Competitive efforts to recruit students are a common emphasis of enrollment managers. The numbers of universities and colleges instituting offices of "enrollment management" have increased in recent years. These offices serve to provide direction and coordination of efforts of multiple offices such as admissions, financial aid, registration, and other student services. Often these offices are part of an enrollment management division.

Some of the typical aims of enrollment management include:

- Improving yields at inquiry, application, and enrollment stages.
- Increasing net revenue, usually by improving the proportion of entering students capable of paying most or all of unsubsidized tuition
- Increasing demographic diversity

- Improving retention rates
- Increasing applicant pools

2.5.8 Management and Generation of Strategic Information

Information technology (IT) has brought a revolution in business practices and serves as a significant element in business strategies. Information systems, enabled by sophisticated technology, among subsidiaries and branches or even inter-companies, can help enterprises adapt swiftly to the ever-changing business environment, providing new forms of design, manufacture, distribution and customer services. In the process of application of the IT, enterprises need an efficient and mature strategic information system (SIS). SIS also plays an important role in educational institutions. It can be applied to facilitate academic and administrative activities in educational institutions. These systems should integrate all information into a single platform to ensure that academic and administrative activities are managed systematically. In the context of teaching and research, SIS can facilitate the process of creating, sharing and diffusing information. Administrators in higher education face complex demands. Apart from administration of staff, student, revenues and all other resources that are applied to higher education, they also need to provide accurate and up to date information in all these areas [31].

Thus, this can be possible only with the aid of computer based systems. SIS information for educational group has the following subfields:

- Profitability analysis
- High yielding program of study
- Low yielding program of study
- Competitive intelligence on key parameters like enrollment etc.
- Target marketing for campaign management
- Data driven planning for growth of university
- Student System which include their registration, study records, lecture time table, exam time table etc.
- Finance system which include cashier, purchase ordering, accounts payback.

2.6 CONCLUSION

The increased use of technology in education is generating an oversized quantity of knowledge each day that has become a target for several researchers round the world; the sphere of academic data processing is growing quickly and has the advantage of containing new algorithms and techniques developed in several data processing areas and machine learning. Data mining of academic data (EDM) helps produce development strategies for the extraction of attention-grabbing, explainable, useful, and novel info, which may cause higher understanding of scholars and therefore the settings during which they learn. EDM may be utilized in many alternative areas together with distinguishing at-risk students, distinguishing priorities for the training desires of various teams of scholars, increasing graduation rates, effectively assessing institutional performance, maximizing field resources, and optimizing subject program renewal. This paper surveyed the foremost relevant studies dole out within the field of EDM together with information utilized in bound studies and therefore the methodologies utilized. It conjointly outlined the foremost common tasks utilized in EDM moreover as people who are the foremost promising for the long run.

Chapter 3

Proposed Methodology

3.1 INTRODUCTION

Educational data mining has vast amount of data that has to be organized in a consistent manner. To organize, analyze and classify students details data mining algorithm is been used based on academic records. The data collected was from Theory of Computing and Compiler Design in Computer Science department of American International University Bangladesh. A questionnaire was also distributed to students to collect data about the other factors considered in the prediction such as the students' socioeconomic, academic factors and motivation to study. The data collected from the result sheet was entered into weka tool for analysis. But it simply specifies the current scenarios whereas no future prediction is available and variables used for analysis are only based on demographic and academic records.

3.2 METHODOLOGY

In this study, primary data were collected from the undergraduate students of American International University Bangladesh, using survey on students. Socio-economic, demographic and educational information were collected of 300 students from Theory of Computing and Compiler Design of Computer Science and Engineering Department. WEKA tool was used to analyze the data. Some tables and cross tables were prepared and some proportions and percentages were calculated to study the distribution of the respondents. Some cross tables were used to study the association of the academic performance of the students among their background characteristics. Naïve Bayes, Multilayer Perceptron and J48 algorithm were used to test the significance of the association between academic performances of the students among their academic and socio economic factors. The methodology starts from the problem definition, then data collection from questionnaire and Students Database. Attribute selection, Nominal conversion, file conversion and WEKA tool implementation. Comparative analysis of efficient classification algorithm is done to predict student's performance by creation of student model.

Student model is designed for the prediction of the outcome of the student based on the framework given below in Figure 3-1. This system provides an efficient analysis on student performance by data collection and result prediction.

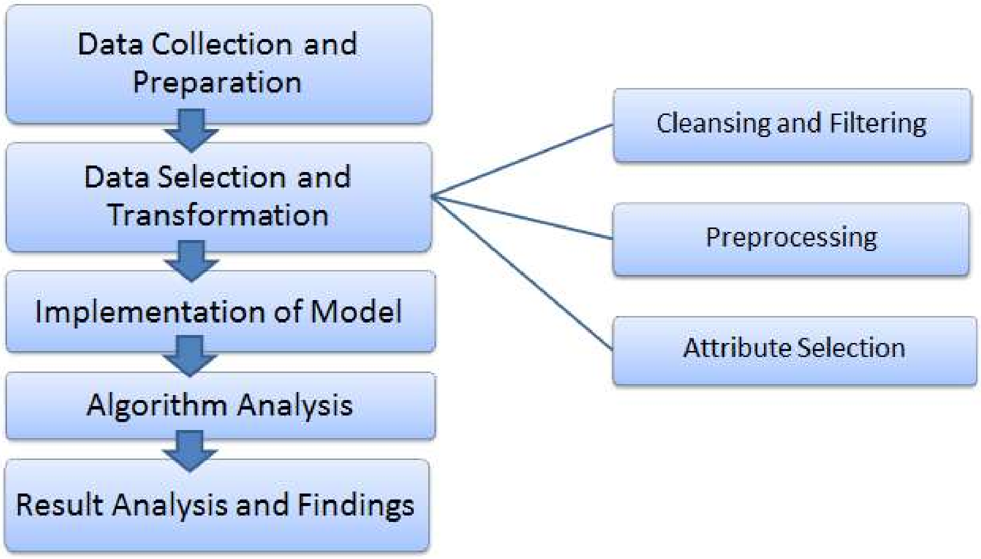


Fig 3-1: Steps to Predict Student Outcome

3.3 DATA COLLECTION AND PREPARATION

The datasets of about 250 students are collected from Theory of Computing and Compiler Design of CSE department in American International University. In this process, a questionnaire form is used to collect the real data from the students that describe the relationship between their socio-economic factor, learning behavior and their academic performance. The variables for judging the learning and academic behavior of students used in the questionnaire are student demographic details, Attendance, CGPA, Interest in this subject, Group study or consultation status and Final grade in last semester. These data's are thereby recorded in excel sheets for analysis. Data sets about 300 students were collected by 2 weeks. Among the dataset around 250 are been used as training dataset to design student model.

3.4 DATA SELECTION AND TRANSFORMATION

In this stage only the data required for data mining are selected. A few derived variables were selected. From the available database, some of the information for the variables is collected. The data collected from Feedback forms and database .initially attribute selection is done. In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. The process of attribute selection deals with selecting the most appropriate attributes for classifying the data sets. By the analysis among the 24 attributes, attributes of higher ranking are used for classifying the training dataset.

The attributes are:

Table 3-1*:* Attributes Used In Data Processing

Variables (Factor)	**Description**	**Attributes**
Gender	Student is Male or Female	Male, Female
Current CGPA	Current CGPA of the student	Number
Quiz Marks	Quiz Marks of the Subject	Number
Midterm Marks	Midterm Mark of the Subject	A+, A, B+, B, C+, C, D+, D, F
Overall Attendance	Overall Attendance Performance	Good, Average, Bad
Interest in this Subject	Interest level for this Subject	High, Medium, Low
Average Study Hours Per Week	Average Study Hours Per Week	0-3, 4-6, 7-10, More than 10
Attend Class in Time	Attend Class in Time	Always, Sometimes, Never
Group Study	Group Study outside of class	Always, Sometimes, Often, Very Few Times, Never
Consultation	Consultation with course teacher for any problem	Always, Sometimes, Often, Very Few Times, Never
Time Spent on Mobile	Time spent on Mobile per day	0-3, 4-6, 7-10, More than 10
Live with Parents	Live with Parents	Yes, No
Working Status	Working Status, If No then 0, if Yes then total hours per week	Number
Final Marks (Dependent Variable)	Final Marks dependent on other variables	A+, A, B+, B, C+, C, D+, D, F

All the predictor and response variables which were derived from the database are given in Table 3-1. On attribute selection the analysis is done for school and college dataset separately based on certain conditions as given below in figure and finally analyzing the performance using final condition on both the records of college and school based on conditions. Data collected from students as feedback and from database. The profile of students is defined based on the academic and demographic details of students. The students' academic background is measured using the entry requirements to be fulfilled to get entry into the university/college. In this stage the only the data required for data mining are selected. A few derived variables were selected. From the available database, some of the information for the variables is collected. The data collected from Feedback forms and database are entered in excel sheets and converted to ARFF format for further processing in WEKA tool.

3.5 IMPLEMENTATION OF MODEL

The main objective is to explore if it is possible to predict the performance of the student (output) based on the various explanatory (input) variables which are retained in the model. The classification model was built using several different algorithms and each of them using different classification techniques.

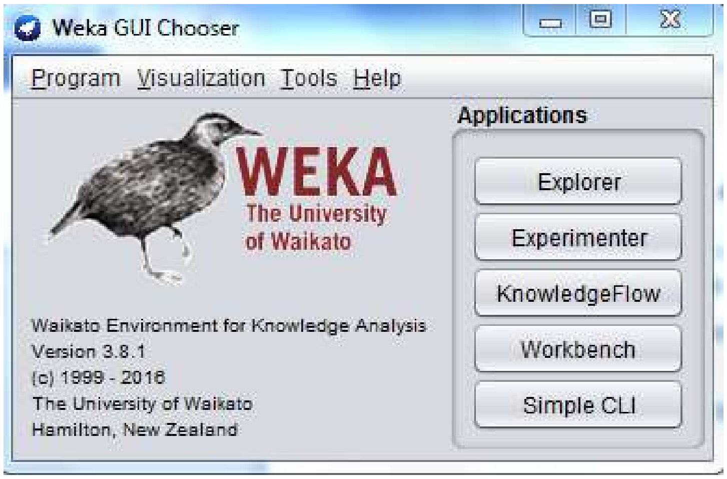
The WEKA Explorer application is used at this stage. The implementation of the dataset is done using a data mining tool WEKA. WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA stands for Waikato Environment for Knowledge Analysis. From the above data, the student.xl file is converted to ARFF (Attribute Relation File Format) for data analysis WEKA explorer. The ARFF file i.e. student.arff is opened in WEKA console. The classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. The algorithm used for classification is Naive Bayes, Multilayer Perception (MLP), and J48.

After pre-processing attribute selection is done using select attribute option in WEKA tool to identify the attributes which has higher rank of contribution to the analysis. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. This predictive model provides way to predict the student's future learning outcome.

A. Applying Training Data in WEKA Tool

WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA stands for Waikato Environment for Knowledge Analysis. From the above data, student.arff file is created, and then this file is loaded into WEKA explorer for processing.

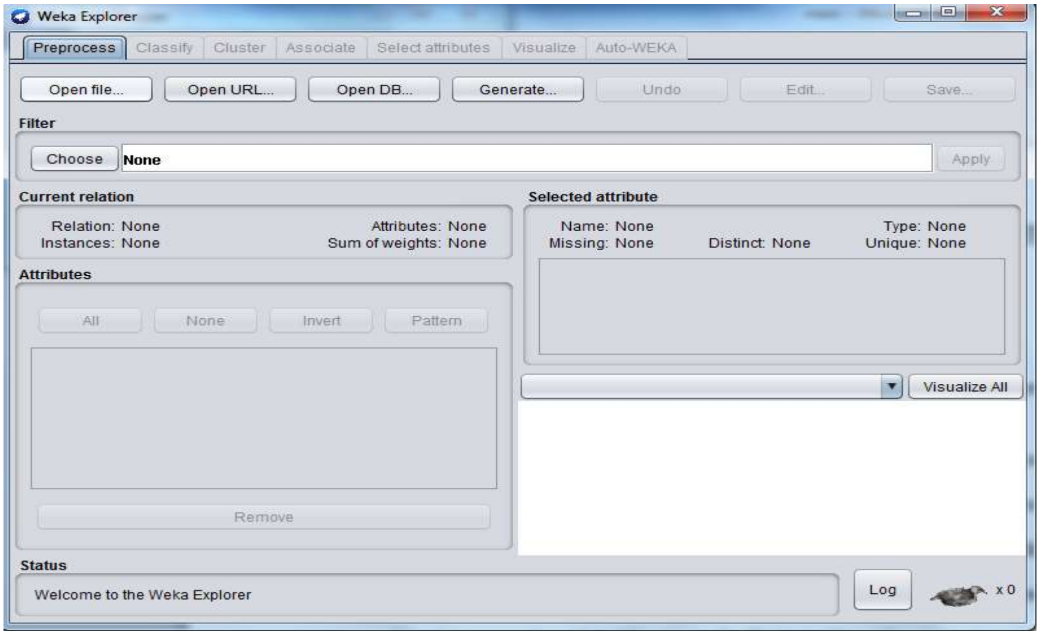
• Choose "WEKA 3.8" from Programs. The first interface that appears looks like the one given below:



Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand

Fig 3-2: Weka Tool

• **Explorer:** An environment for exploring data. It supports data preprocessing, attribute selection, learning and visualization, the screen is given below:



Filter: None

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

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

Welcome to the Weka Explorer

Fig 3-3: Weka Explorer

• Get to the WEKA Explorer environment and load the dataset using the Preprocess mode.

• Get to the Classify mode (by clicking on the Classify tab) as shown below:

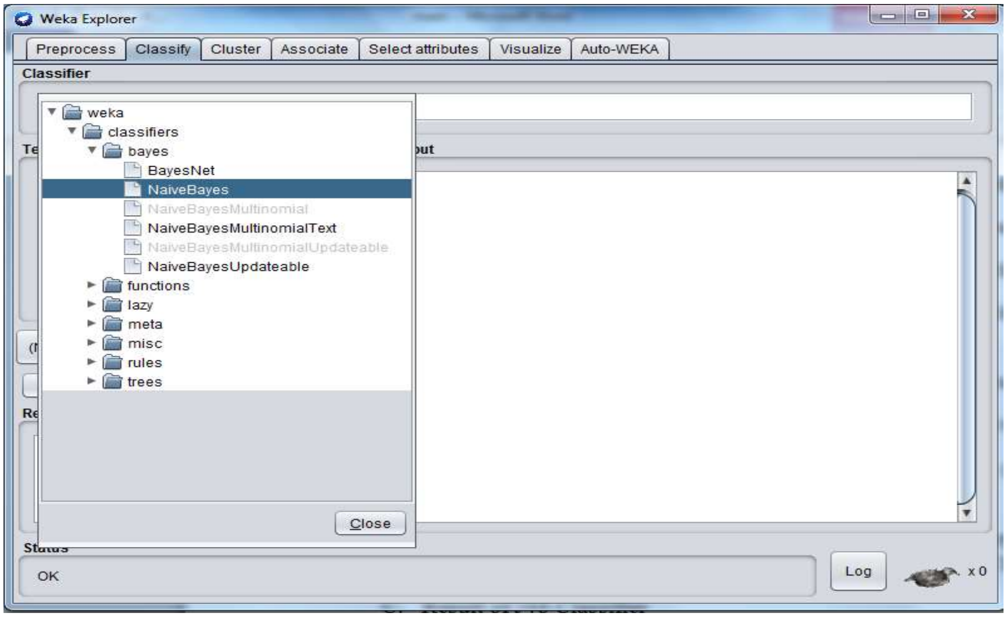


Fig 3-4: Using Naïve Bayes Algorithm to Analyze Dataset

• Use training set means that using the training set (the file loaded in Preprocess) for testing.

B. Comparative identification of efficient Classification algorithm

• Initially the datasets are filtered with attributes of higher rank for classification based on select attribute option.

• The datasets are thereby tested with various classification algorithms

• Classifiers simulated are: ***Naïve bayes***, ***Multilayer Perception*** and ***J48 Tree***.

• Running a Test Click on the Choose button and choose a classifier (the default is ZeroR – the majority predictor).

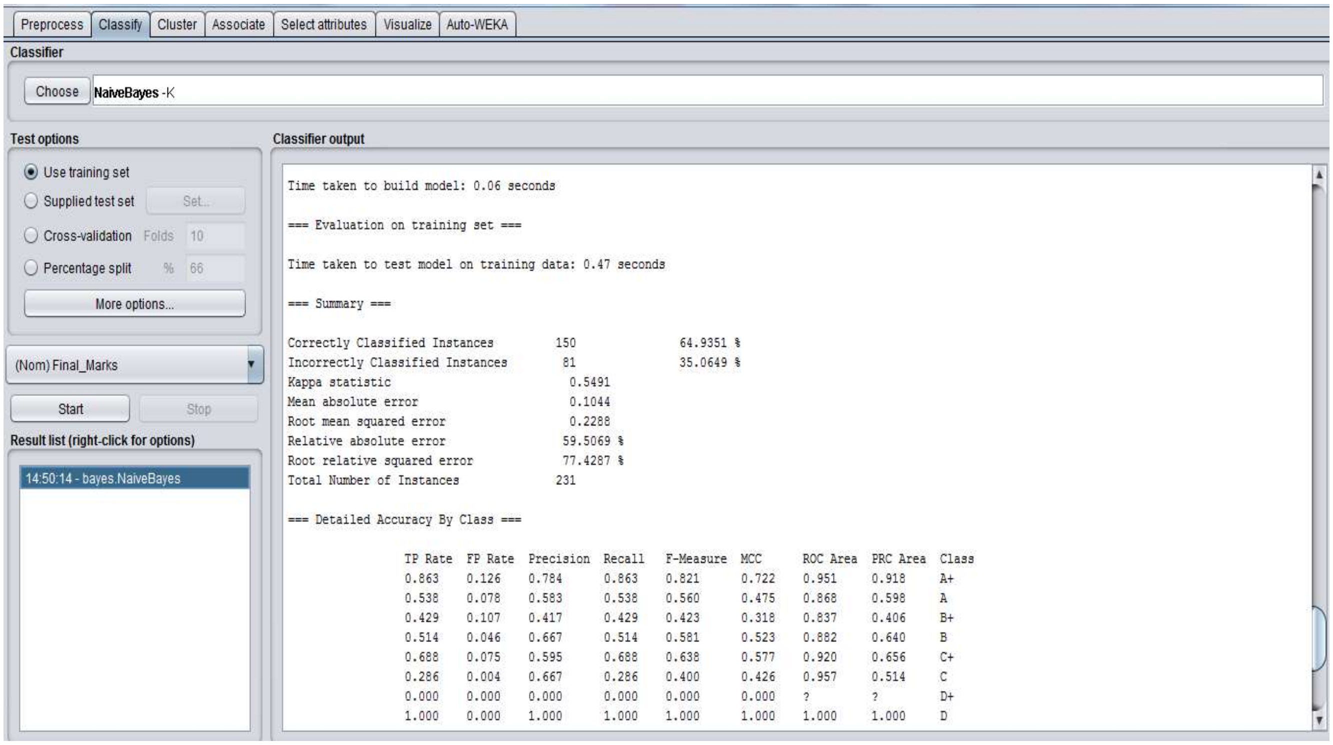
• After selecting a classifier and setting its parameters (we can always start with the defaults), click on OK and then on Start. We can get the output from the classifier in the Classifier output window.

C. *Result of Naïve Bayes Algorithm*

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.

The Naive Bayes classification algorithm includes the probability-threshold parameter ZeroProba. The value of the probability-threshold parameter is used if one of the above mentioned dimensions of the cube is empty. A dimension is empty, if a training-data record with the combination of input-field value and target value does not exist [29].

On evaluating the data set under Naïve Bayes classifier, the result generated is as shown below which shows the correctly classified instance by 65%. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. These predictive models provide ways to predict the percentage of accuracy of result.



NaiveBayes -K

● Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Final_Marks

14:50:14 - bayes.NaiveBayes

```
Time taken to build model: 0.06 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.47 seconds

=== Summary ===

Correctly Classified Instances         150               64.9351 %
Incorrectly Classified Instances        81               35.0649 %
Kappa statistic                          0.5491
Mean absolute error                      0.1044
Root mean squared error                  0.2288
Relative absolute error                 59.5069 %
Root relative squared error             77.4287 %
Total Number of Instances              231

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.863    0.126    0.784      0.863    0.821      0.722    0.951     0.918     A+
                 0.538    0.078    0.583      0.538    0.560      0.475    0.868     0.598     A
                 0.429    0.107    0.417      0.429    0.423      0.318    0.837     0.406     B+
                 0.514    0.046    0.667      0.514    0.581      0.523    0.882     0.640     B
                 0.688    0.075    0.595      0.688    0.638      0.577    0.920     0.656     C+
                 0.286    0.004    0.667      0.286    0.400      0.426    0.957     0.514     C
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         D+
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     D
```

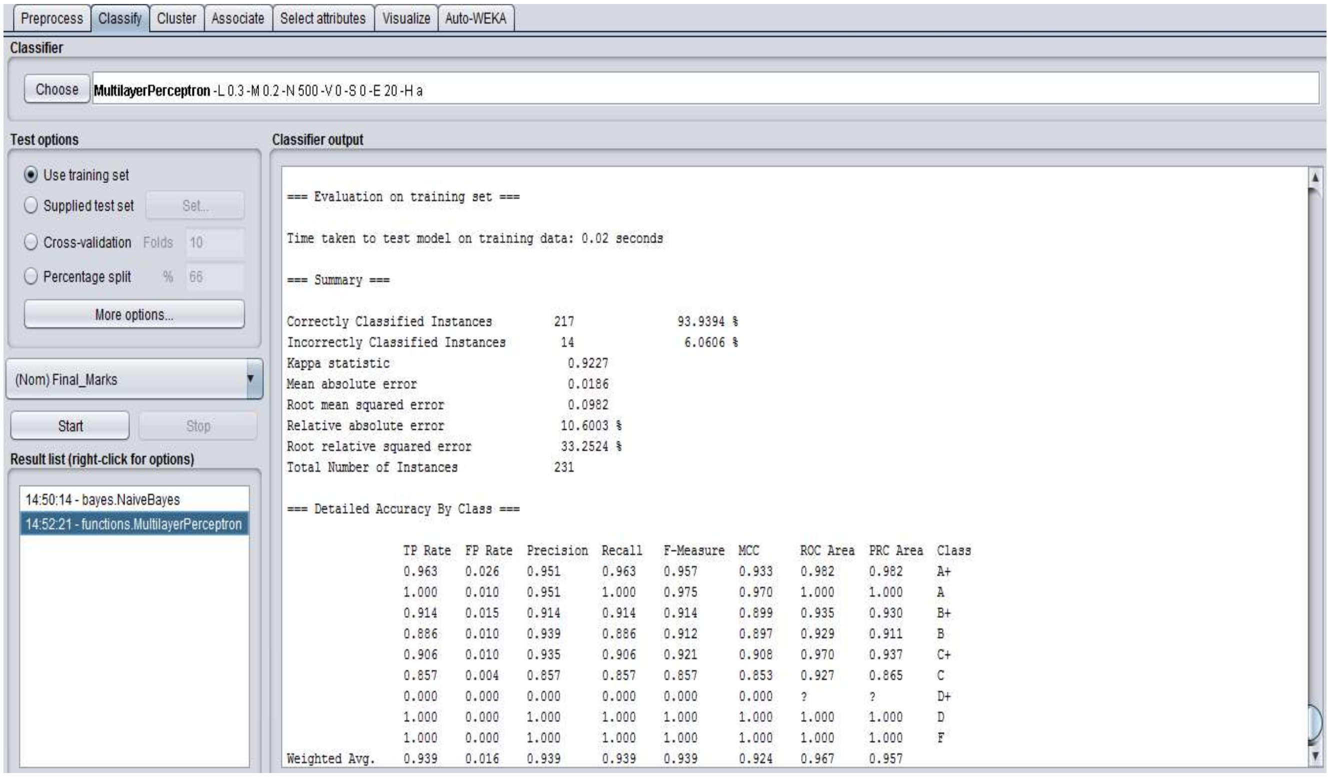
Fig 3-5: Analysis result of Naïve Bayes Algorithm

D. Result of Multilayer Perceptron Algorithm

A **Multilayer Perceptron** (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.

The supervised learning problem of the MLP can be solved with the *back-propagation algorithm*. The algorithm consists of two steps. In the *forward pass*, the predicted outputs corresponding to the given inputs are evaluated as in Equation. In the *backward pass*, partial derivatives of the cost function with respect to the different parameters are propagated back through the network. The chain rule of differentiation gives very similar computational rules for the backward pass as the ones in the forward pass. The network weights can then be adapted using any gradient-based optimization algorithm. The whole process is iterated until the weights have converged [33].

On evaluating the data set under Multilayer Perceptron, the result generated is as shown below which shows the correctly classified instance by 94%.



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Final_Marks

14:50:14 - bayes.NaiveBayes
14:52:21 - functions.MultilayerPerceptron

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         217               93.9394 %
Incorrectly Classified Instances        14                6.0606 %
Kappa statistic                          0.9227
Mean absolute error                      0.0186
Root mean squared error                  0.0982
Relative absolute error                 10.6003 %
Root relative squared error             33.2524 %
Total Number of Instances              231

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.026    0.951      0.963    0.957      0.933    0.982     0.982     A+
                 1.000    0.010    0.951      1.000    0.975      0.970    1.000     1.000     A
                 0.914    0.015    0.914      0.914    0.914      0.899    0.935     0.930     B+
                 0.886    0.010    0.939      0.886    0.912      0.897    0.929     0.911     B
                 0.906    0.010    0.935      0.906    0.921      0.908    0.970     0.937     C+
                 0.857    0.004    0.857      0.857    0.857      0.853    0.927     0.865     C
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         D+
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     D
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     F
Weighted Avg.    0.939    0.016    0.939      0.939    0.939      0.924    0.967     0.957
```

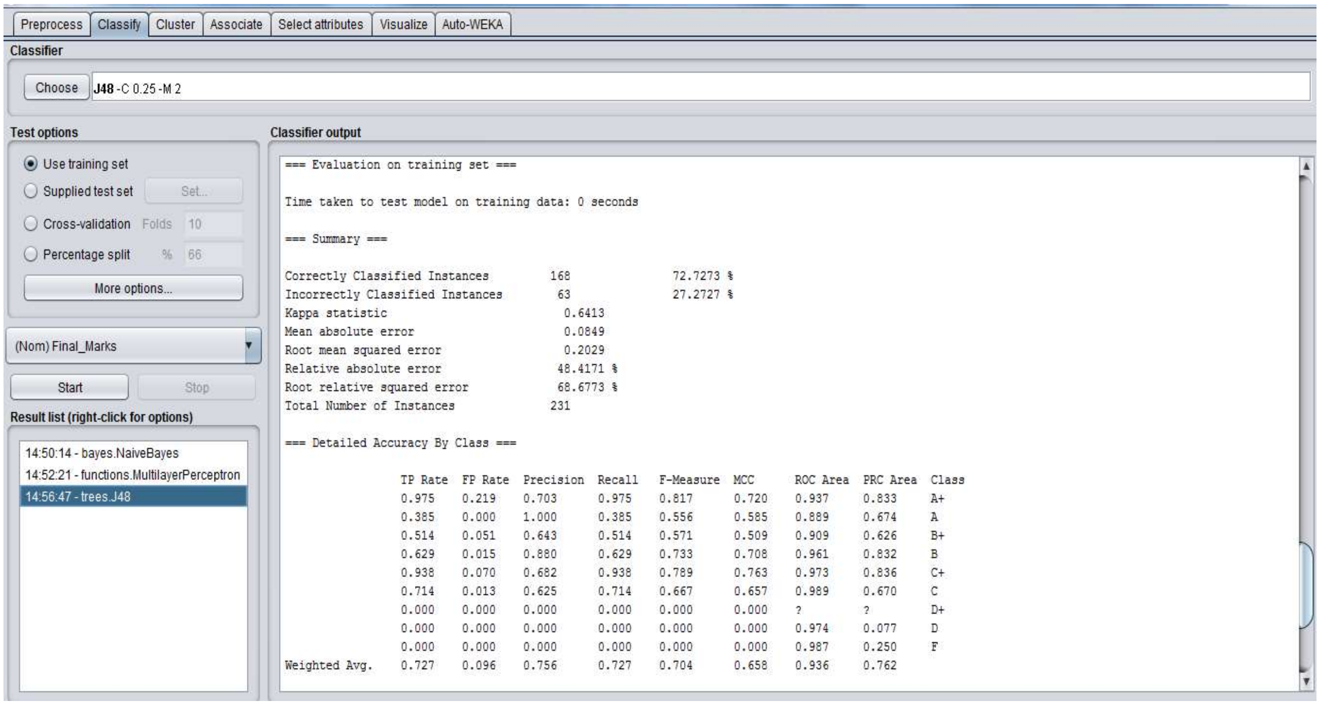
Fig 3-6: Analysis result of Multilayer Perceptron Algorithm

E. Result of J48 Algorithm:

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible [33].

This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

On evaluating the data set under J48 Algorithm, the result generated is as shown below which shows the correctly classified instance by 72%.



J48 -C 0.25 -M 2

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Final_Marks

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         168               72.7273 %
Incorrectly Classified Instances        63               27.2727 %
Kappa statistic                          0.6413
Mean absolute error                      0.0849
Root mean squared error                  0.2029
Relative absolute error                 48.4171 %
Root relative squared error             68.6773 %
Total Number of Instances              231

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975    0.219    0.703      0.975    0.817      0.720    0.937     0.833     A+
                 0.385    0.000    1.000      0.385    0.556      0.585    0.889     0.674     A
                 0.514    0.051    0.643      0.514    0.571      0.509    0.909     0.626     B+
                 0.629    0.015    0.880      0.629    0.733      0.708    0.961     0.832     B
                 0.938    0.070    0.682      0.938    0.789      0.763    0.973     0.836     C+
                 0.714    0.013    0.625      0.714    0.667      0.657    0.989     0.670     C
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         D+
                 0.000    0.000    0.000      0.000    0.000      0.000    0.974     0.077     D
                 0.000    0.000    0.000      0.000    0.000      0.000    0.987     0.250     F
Weighted Avg.    0.727    0.096    0.756      0.727    0.704      0.658    0.936     0.762
```

Fig 3-7: Analysis result of J48 Algorithm

F. *Decision Tree:*

A decision tree is a supervised classification technique that builds a top-down tree-like model from a given dataset attributes. The decision tree is a predictive modeling technique used for predicting, classifying, or categorizing a given data object based on the previously generated model using a training dataset with the same features (attributes). The structure of the generated tree includes a root node, internal nodes, and leaf (terminal) nodes. The root node is the first node in the decision tree which have no incoming edges, and one or more outgoing edges; an internal node is a middle node in the decision tree which have one incoming edge, and one or more outgoing edges; the leaf node is the last node in the decision tree structure which represents the final suggested (predicted) class (label) of a data object [34].

After analyzing the data set we have loaded the decision tree based on the final result, the view of decision tree model is given below:

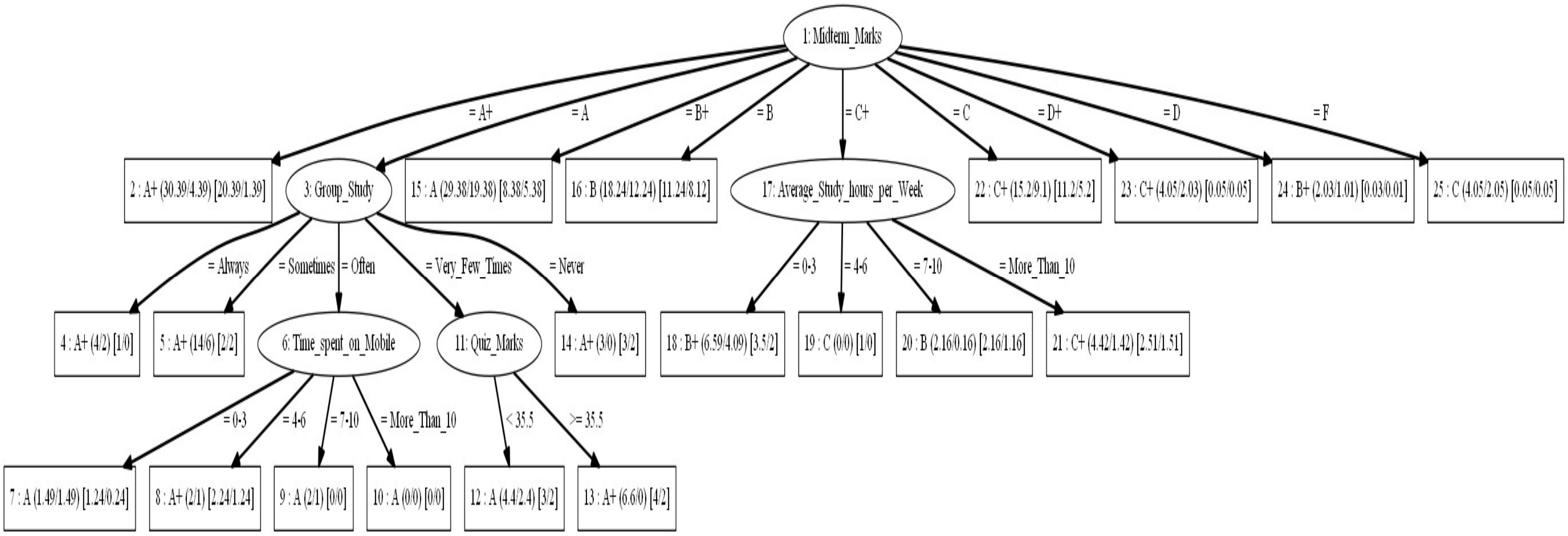


Fig 3-8: J48 Tree Visualization

Chapter 4

Result Analysis

4.1 RESULT

In this section, multiple decision tree techniques and algorithms were reviewed, and their performances and accuracies were tested and validated. As a final analysis, it was obviously noticed that some algorithms worked better with the dataset than others.

By the simulation of dataset with the various classifiers the accuracy of correctly classified instances is as below in Table 4-1.

Table 4-1: Accuracy of each Algorithm

Classification Algorithm	**Accuracy**
Naïve Bayes	65%
Multilayer Perceptron	94%
J48	72%

As Multilayer Perceptron algorithm classifies the instance with maximum accuracy, it is used in designing the student model to predict students' performance by analyzing training data and test data, thereby predicting students' performance as provided grades.

After checking the J48 decision tree, it shows that Midterm marks, Quiz marks, Time spent on Mobile, Group Study, and Study hours per week have influenced the academic performances of a student. Among these variables most important variable for all the faculties turned out to be Midterm and Group Study variable. Time spent on mobile also varies the result, it comes out that as low as the student uses his mobile, he can concentrate on his study. Another factor was average time hours per week, that is also changes the factor of final terms.

4.2 RECOMMENDATION

Following the research results it may be recommended that student’s relation with each other in every class should be improved for regular group study and students should be more aware about their midterm result because it effects the final marks in high percentage.

Students mobile usage should reduce from past time, university can take some strict rules to avoid using mobile phone for a long time. Our analysis yielded a significant negative relationship between total time spent using smartphones and academic performance, after controlling for known predictors of performance such as self-efficacy and past academic results. Moreover, if we consider usage during class time only (as opposed to during free time and weekends), the effect was almost twice as high. The magnitude of the effect found is alarming. Thus, this study brings new evidence of the potential harm of excessive smartphone use and should be useful for educators to be aware of the impact of technology on students' performance.

Students with higher and above average grades indicated that their teachers treated them with respect. This showed that these students interacted with teachers better than students with average and low grades. Almost all students thought that student-teacher relationship could affect their academic attainment. This was not supported by the finding which showed no association between the student-teacher relationship and the students’ grades. This implies that their grades were not dependent on their views towards student-teacher relationship. This could be explained by the influence of the student-teacher relationship on the psychosocial condition of the student, rather than their academic achievement. In addition, a majority of students irrespective of their grades observed that attendance of the students at lectures could have a bearing on student-teacher relationship. Students with higher and above average grades felt that the teacher's performance could also be affected by the students’ attendance. This implies that students had a part to play in the forging of the student-teacher relationship.

Most students independent of their academic grades thought that interactive teaching was the best method of teaching on the grounds that it kept them alert, interested, active and facilitated and enhanced their understanding. Students could do without a teacher who merely gave information and left students to work it out by themselves [27]. The relationship between the teacher and students was influenced significantly by the help teachers gave students outside lectures. It was assumed that previous or senior students’ opinions had an effect on the junior students’ views towards their teachers. Surprisingly, our study showed that all students were influenced by the opinions of the students ahead of them of their teachers, but found no significant association with their grades. This can be explained by the fact that the experience of the senior students was more trustworthy than their own experiences, which is more common in medical colleges. Also average study hours should increase, so teacher can take initiative to provide regular assignments and quizzes for students to keep them busy in their regular study.

Chapter 5

Conclusion and Future Work

5.1 CONCLUSION

Educational data mining mainly focus to analyze the education system of an institution. The model focuses on analyzing the prediction accuracy of the student's performance .The dataset that comprises of all academic and personal factors of the students. This model can be useful in the educational system like Universities and Colleges. By this model we can know the academic status of the students in advance and can concentrate on students to improve their academic results and placements. Thereby improve their standards and reputations. As a result the quality of education can be improved. The results of the data mining algorithms for the classification of the students based on the attributes selected reveals that the prediction rates are not uniform among the algorithms. The range of prediction varies from (65-92%).Thereby by comparative analysis of classification algorithms (such as Naïve Bayes, Multilayer Perceptron, Decision tree, J48) using WEKA tool, it is proven that the attributes chosen from the original dataset have high influence using Multilayer Perceptron with an accuracy of 92% under analysis and used for predicting test data set for future outcome as grades. The work can be further extended out by designing the student model analyzing records of students extra-curricular skills and provide a suggestions on communication and technical skill development by which students can be built in professional aspect of talents. Data analysis plays an important role for any type of decision support irrespective of type of industry. Data warehousing and data mining methods are used for data analysis are explained in detail. Main core of this paper is to review role of data mining techniques in education system.

5.2 FUTURE WORK

Educational Data Mining has been started as an upcoming research area, again the number of specific tools specially developed for applying Data Mining algorithms in educational data/environments are emerging day by day. Data mining techniques in educational organizations help us to learn student performance, student behavior, carefully designing course curriculum, to motivate students and to group student depending upon various parameters.

It is observed that recent Data Mining tools are too complex to use and their features go well beyond the scope of what an educator may want to do. One possible solution is to develop the tools that use a default algorithm for each task and parameter-free Data Mining algorithms to simplify the configuration and execution for non-expert users. Secondly, the Data Mining tool has to be integrated into the e-learning environment so that results obtained with Data Mining techniques could be easily and directly applied. Moreover, recent tools for mining data pertaining to a specific course/framework may be useful to their developers only. There are nothing like general tools or reusing tools that can be applied to any educational system. Therefore, a standardization of input data and output model are needed. Data mining techniques are useful in selection revenue analysis, student marketing, predicting student performance, planning of courses and result analysis. So, it has a wider area of applications for the higher education sector.

Bibliography

1. Osmanbegovic E., Suljic M. “Data mining approach for predicting student performance” Economic Review- Journal of Economics and Business. Volume 10 (2012)

2. Sajadin Sembiring (2012) ‘An Application Of Predicting Student Performance Using Kernel K-Means And Smooth Support Vector Machine’ - Universiti Malaysia Pahang

3. El-Halees, A. (2008) ‘Mining Students Data to Analyze Learning Behavior: A Case Study’, The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.

4. Suman Biswas, Mst. Shonamoty Khatun, Md. Yasin Ali Parh & Dr. Md. Sazzad Hossain (2016) ‘Factors That Affect The Academic Results: A Case Study Of Islamic University, Kushtia, Bangladesh’- Global Journal Of HUMAN-SOCIAL SCIENCE: G Linguistics & Education Volume 16 Issue 1 Version 1.0 Year 2016

5. Easmin, S., Hossain, M.A. and Das, P.C. (2015). Effect of Socioeconomic Background on the Academic Performance of the Students: A Study on Undergraduate Students of Bangladesh, The Cost and Management 43(4): 28-36.

6. Alam, M.M., Billah, M.A. and Alam M.S. (2014). Factors Affecting Academic Performance of Under-graduate Students at International Islamic University Chittagong (IIUC), Bangladesh. Journal of Education and Practice, 5(39): 143-154.

7. Rahman, M.M. (2012). Students’ Matriculation Factors for Higher Education in Private Universities of Bangladesh, Global Disclosure of Economics and Business, 1(1): 54-64.

8. Mutairi, A.A. (2011). Factors Affecting Business Students' Performance in Arab Open University: The Case of Kuwait, International Journal of Business and Management, 6(5): 146-155.

9. Mohammed M. Abu Tair, Alaa M. El-Halees (2012) 'Mining Educational Data to Improve Students' Performance: A Case Study' International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.

10. Bekele, R., Menzel, W. "A bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students". Journal of Information Science (2013).

11. Bhardwaj, K., Pal, S "Data Mining: A prediction for performance improvement using classification". International Journal of Computer Science and Information Security. Volume 9 (2011).

12. Cortez P, Silva A. Using data mining to predict Secondary school student performance. Journal of information science Volume 2(6). (2013).

13. Mladen D., Mirjana P. B., Vanja Š., "Improving University Operations with Data Mining: Predicting Student Performance", International Journal of Social, Behavioral, Educational, Economic and Management Engineering Volume 8(4), 2014.

14. Sembiring S, Zarlis, M, Hartama, D. Ramliana S, Elvi W. "Prediction of student academic performance by an application of data mining techniques." International Conference on Management and Artificial Intelligence IPEDR Volume.6, (2011).

15. 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.

16. Meltem, D. "Gender difference in academic performance in a large public university in Turkey". Economic Research center working papers in economics. 4(17). Pp. 22-23, (2004).

17. Abubakar, R. B. and Oguguo, O. D. "Age and gender as predictors of academic achievements of College mathematics and science students." Proceedings of the International Conference of teaching, learning and change. International Association of Teaching and learning. (2011).

18. Nnamani, C. N, Dikko, H. G and Kinta, L. M. “Impact of students‟ financial strength on their academic performance: Kaduna Polytechnic experience”. African Research Review 8(1), (2014).

19. Ogunde A.O., Ajibade D.A. “A data Mining System for Predicting University Students F=Graduation Grade Using ID3 Decision Tree approach”, Journal of Computer Science and Information Technology, Volume 2(1) (2014).

20. Ryan S.J.D. Baker, Kalina Yacef. “The State of Educational Data Mining in 2009: A Review and Future Visions”, Journal of Educational Mining, Volume 1(2009).

21. Undavia, J. N., Dolia, P. M.; Shah, N. P. “Prediction of Graduate Students for Master Degree based on Their Past Performance using Decision Tree in Weka Environment”. International Journal of Computer Applications; Volume 74 (21), (2013).

22. Surjeet K, Yadav, Bharadwaj, B. Pal B.” Data Mining Applications: A comparative Study for Predicting Student‟s performance.” International journal of innovative technology & creative engineering. Volume 1(12). (2012).

23. Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) ‘Mining Student Data Using Decision Trees’, The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

24. Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) ‘Data Mining Model for Higher Education System’, European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

25. Baradwaj, B. and Pal, S. (2011) ‘Mining Educational Data to Analyze Student s’ Performance’, International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

26. Chandra, E. and Nandhini, K. (2010) ‘Knowledge Mining from Student Data’, European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

27. El-Halees, A. (2008) ‘Mining Students Data to Analyze Learning Behavior: A Case Study’, The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.

28. Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

29. Kumar, V. and Chadha, A. (2011) ‘An Empirical Study of the Applications of Data Mining Techniques in Higher Education’, International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84.

30. Mansur, M. O., Sap, M. and Noor, M. (2005) ‘Outlier Detection Technique in Data Mining: A Research Perspective’, In Postgraduate Annual Research Seminar.

31. Romero, C. and Ventura, S. (2007) ‘Educational data mining: A Survey from 1995 to 2005’, Expert Systems with Applications (33), pp. 135-146.

32. Crist´obal Romero, Member, IEEE, and Sebasti´an Ventura, Senior Member, IEEE, “Educational Data Mining: A Review of the State of the Art” VOL. 40, NO. 6, NOVEMBER 2010.

33. Parneet Kaura, Manpreet Singhb ,Gurpreet Singh Josanc “Classification and Prediction based Data Mining Algorithms to Predict Slow Learners in Education Sector” Science Direct Procedia Computer Science 57 (2015) 500 – 508 2015 (ICRTC- 2015).

34. Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri, “Data Mining Models for Student Careers”, Science Direct - Expert Systems with Applications 42 (2015) 5508–5521.

35. Nurbiha A Shukora , Zaidatun Tasira, Henny Vander Meijden, “An Examination of Online Learning Effectiveness using Data Mining” ,Science Direct - Procedia - Social and Behavioral Sciences 172(2015) 555 – 562.