PREDICTING STUDENT PERFORMANCE TO REDUCE DROPOUT USING J48 DECISION TREE ALGORITHM

PREPARED BY

MD. ZAHIDUL HASAN
ID: 113-25-231

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

Supervised By

Dr. Akhter Hossain
Professor and Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2019

APPROVAL

This Thesis titled **"Predicting Student Performance to Reduce Dropout using J48 Decision Tree Algorithm"**, submitted by Md. Zahidul Hasan, ID No: 113-25-231 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 December 6, 2019.

BOARD OF EXAMINERS

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

Chairman

Dr. Md. Ismail Jabiullah
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

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

Internal Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

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

Supervised by:

Dr. Akhter Hossain
Professor and Head
Department of CSE
Daffodil International University

Submitted by:

(Md. Zahidul Hasan)
ID: 113-25-231
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty ALLAH for His divine blessing makes me possible to complete the final year thesis successfully.

I really grateful and wish my profound my indebtedness to **Dr. Akhter Hossain, Professor and Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my 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 drafts and correcting them at all stages have made it possible to complete this thesis.

I would like to express my heartiest gratitude to other faculty members and the staff of the CSE department of Daffodil International University.

I would like to thank my entire course mate at Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents, my wife, and my two daughters.

ABSTRACT

Student dropout is a major problem that faces most of the Private University of Bangladesh. In Bangladesh, a huge number of student studies in private universities. But, almost one-third of students are dropped out in the first year of their study. Various reasons are identified behind this problem, including the medium of study, assessment process and lack of knowledge in semester-based education. In this research, the researcher uses the Educational Data Mining technique to predict the students' final grade after the Midterm Examination. This early result prediction can help the students, teachers and the university authority to take necessary action to reduce students' dropout. A lot of Education Data Mining tools are available. This research proposes a classification model particularly a decision tree algorithm to predict the future grades of the students in **Introduction to Computer**, a first-semester course for undergraduate university students. Popular data mining software, WEKA is used for model construction and evaluation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

A huge amount of facts is collected each day from varied sources, like academic institute, business organization, economics, geography, sports, health and medicine, etc. Such data are very important source of information that can help the organization for smooth operation. It is extremely impossible for human to discover the hidden knowledge from those data without powerful tools. That’s why, special techniques are required to discover potential information from those data to make proper decisions. A popular term, data mining can be used to extract relevant information from large and complex databases.

According to Wikipedia [1], “Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.” In data mining, different machine learning algorithms are used to identify hidden but potentially important patterns in data. The main purpose of data mining is to find patterns in data that allow us to take fast and accurate decision based on historical data. Data mining process can be applied on database, Datawarehouse, Datamart, the Web, other information repositories data, or streamed data.

Data mining techniques are extremely helpful for knowledge analysis and predictions. One of the most used data mining technique is classification, that assigns items in a collection to target categories or classes. A variety of classification algorithm like, Decision tree, Neural network, Bayesian network, etc. are used in data mining. The primary goal of classification algorithm is to accurately predict future tendency based on the past pattern.

In this research paper, the researcher applied J48 Decision tree algorithm to early predict the student’s final grade of a particular course (Introduction to Computers) after the mid-term examination. One of the popular machine learning /data mining software, WEKA has been used for model construction and evaluation.

1.2 MOTIVATION

All the work described in this thesis was conducted at the Northern University Bangladesh (NUB), Business Campus, Dhanmondi, Dhaka. The reasons for choosing this topic and conducting the research work at this university were, (1) I have served more than 11 years for that university as a faculty member of CSE department and, thus, have a good understanding of the actual enrollment and dropout situation of first year students; (2) Most of the time, I have been conducted classes on Introduction to Computer, Basic Mathematics and Computer Applications in Business course of First year students; (3) Beside that, I have been actively involved as a course supervisor for the First year students. (4) Moreover, I have been also conducted a research work on learning behavior of first year students. All the above mentions reasons argue most powerfully for the topic and location of the research work conducted.

1.3 RATIONALE OF THE STUDY

In this research paper, the researcher applied J48 Decision tree algorithm to early predict the student's final grade. This study will help the teachers, students, management of the university and other stackholders to take proper action to reduce dropout of first year students.

This study reveals the following benefits.

- From this study the readers will come to know about the factors that directly related to performance at university level students.
- This study helps the students and the teachers to find out the gaps and plan to improve.
- This also, helps the university authority to control student dropout rate.

1.4 RESEARCH QUESTIONS

From this study, the researcher wants to measure the impact of the following on the academic performance of undergraduate students:

- Parents' educational qualifications
- Parents' social-economic status
- Students' previous academic result
- Student's class attendance, class performance and mid-term exam marks

1.5 EXPECTED OUTPUT

The researcher expects a strong and positive relationship on academic performance and the followings: parents' educational qualifications, parents' social-economic status, students' previous academic result, student's class attendance, class performance, mid-term exam marks.

CHAPTER 2

BACKGROUND

2.1 INTRODUCTION

A lot of factors that can influence the academic performance of a student. In this study, the researcher analyzes admission test scores, past academic records, social-economic status of parents, parents educational background, current academic performance, etc. to predict the grade of the student that can help to reduce the dropout rate.

2.2 RELATED WORKS

Good amount of data mining researches have been conducted on Educational domain. Some of them are briefly describe below for better understanding of the readers.

In 2005, Khan [2] conducted a study on “Scholastic Achievement of Higher Secondary Students in Science Stream – for that study, 400 students (200 boys and 200 Girls) selected from senior secondary school of A.M.U., Aligarh-India, to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The scores obtained on different variables were factor-analyzed to get a smaller number of meaningful variables or factors to establish the predictive validity of these predictors. Factors responsible for success in science stream were identified. The prognostic value of the predictors was compared for high achievers and low achievers in order to identify the factors which differentiate them”.

In 2007, Galit [3] conducted a study on “A sample study on applying data mining research techniques in educational science: developing a more meaning of data – in that study, Data was analyzed using descriptive statistics (t- test and analysis of variance), and the data mining techniques of decision tree, dependency networks and clustering to predict the results and to warn students at risk before their final exams”.

In 2006, Al-Radaideh, et al [4] conducted a study on “Mining Student Data Using Decision Trees – That paper was an attempted to use the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. For this purpose, the CRISP framework for data mining is used for mining student related academic data. The classification rule generation process is based on the decision tree as a classification method where the generated rules are studied and evaluated. A system that facilitates the use of the generated rules is built which allows students to predict the final grade in a course under study.”

In 2011, Pandey and Pal [5] conducted a study on “Mining Educational Data to Analyze Students' Performance – In that research, the classification task was used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. By this task we extract knowledge that describes students' performance in end semester examination. It helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling. Keywords-Educational Data Mining (EDM); Classification; Knowledge Discovery in Database (KDD); ID3 Algorithm.”

In 2012, Monika Goyal and Rajan Vohra [6] conducted a study on “Importance of Data Mining in Higher Education System - In that research, data mining techniques was applied to improve the efficiency of the higher education institution. Data mining techniques such as association, clustering, decision tree would help to improve students’ performance, selection of courses, to measure their retention rate, their life cycle management, and the grant fund management of an institution if applied to higher education processes.”

In 2012, Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal [7] conducted a study on “Mining Education Data to Predict Student's Retention: A comparative Study – That paper presented a data mining project to generate predictive models for student retention management. Given new records of incoming students, these predictive models can produce short accurate prediction lists identifying students who tend to need the support from the student retention program most. That paper examined the quality of the predictive models generated by the machine learning algorithms. The results show that some of the machines learning algorithms are able to establish effective predictive models from the existing student retention data."

In 2013, K.Shanmuga Priya and A.V.Senthil Kumar [8] conducted a study on "Educational Research Survey on Evaluation ff Students and Staff Performance using Various Data Mining Techniques- there applied a Classification Technique in Data Mining to improve the student's performance and help to achieve the goal by extracting the discovery of knowledge from the end semester mark."

In 2015, Ahmad et al [9] "designed a framework to predict the academic performance of the first-year bachelor students of computer science course. The dataset contained 8 years of data starting from July 2006-07 to July 2013-14. The data contained students' records including demographics, previous academic records and family background. Three different classifiers Rule-Based classifiers, Decision Tree and Naïve Bayes are applied to find the academic performance of students. The experiments showed that Rule-Based classifier was the best among the other classifiers and its accuracy was found as 71.3%. The first-year students' level of success was predicted by the model."

In 2016, Sumitha et. al. [10] "developed a model to predict student's future learning outcomes using senior students' dataset. They compared the data mining classification algorithms and found that the J48 algorithm was best suited for such a job based on their data."

In 2016, Amjad Abu Saa [11] conducted a study on "Educational Data Mining & Students' Performance Prediction- that study was built a qualitative model to analyze student performance based on students' personal and social factors. He explored theoretically various factors of the students' performance in the field of higher education."

In 2017, Khasanah et. al. [12] conducted a study on "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques- to examined high influence attributes to predict student performance. Feature selection may be used before classification for such a job. They used Bayesian Network and Decision Tree algorithms for classification and prediction of student performance."

2.3 SCOPE OF THE PROBLEM

This study was conducted in Northern University Bangladesh, Business Campus, Dhanmondi, Dhaka, using a questionnaire method to collect data of first year students who are taking Introduction to Computer course in between 2016 to 2018 session.

2.4 CHALLENGES

In this research the researcher used classification technique in data mining to predict future performance based on past data. There is an uncertainty, because prediction may be wrong in some cases. For better prediction a large dataset is required for model creation, but in this study, the researcher used only 82 records for model creation. Beside that, Some data are confidential for the organization, that's why the authority didn't allow to access those data.

2.5 A REVIEW ON DATA MINING

2.5.1 DATA MINING DEFINITION

Accoeding to Oxford Dictionary "Data miming is the practice of examining large pre-existing databases in order to generate new information."

According to The Economic Times [13] "Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD)."

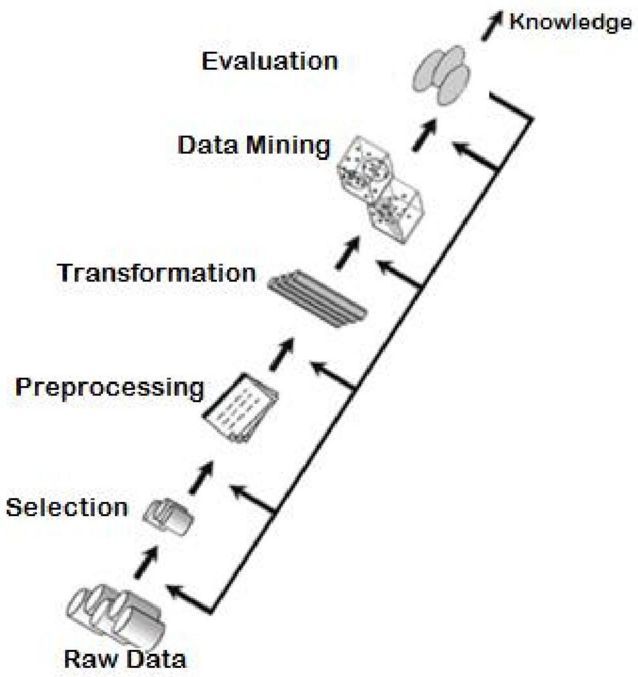


Figure 2.5.1: The steps of extracting knowledge from data

2.5.2 DATA MINING PROCESS

In data mining, the knowledge discovery process involves the sequence of the following steps:

1. Select the proper source data
2. Prepare the source data by preprocessing
3. Transforme the prepossessed data into suitable format that is required for data mining
4. Apply appropriate data mining algorithm to extract interesting data patterns
5. Examine the patterns to identify relevant information
6. Use mined knowledge to achive the goal

2.5.3 DATA PREPROCESSING

Data pre-processing is the series of steps to prepare the data set before apply any data mining algorithm. According to Wikipedia[14] -“Data preprocessing is an important step in the data mining process. The phrase garbage in, garbage out is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology.”

Figure 2.5.3 shows the steps involved in data preprocessing:

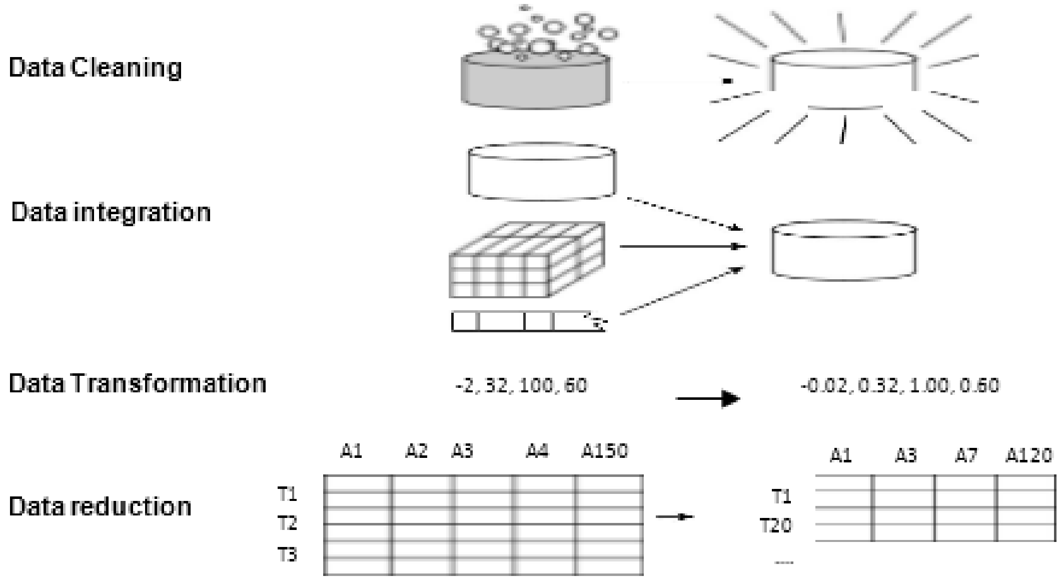


Figure 2.5.3: The series of steps involve in data preprocessing

2.6 EDUCATIONAL DATA MINING (EDM)

Data mining, which deals with educational data is referred to as Educational Data Mining (EDM). In higher education, academic planners can take benefits by using Educational data mining. Because, EDM can help in the decision making process by analyzing educational databases to enhance students' performance, to understand students' behavior, to guide instructors, to improve teaching quality, etc. EDM can also used to identify the low performers who require support by analyzing students' learning process.

2.7 VARIOUS DATA MINING TECHNIQUES

Various data mining methods are available in the market to carryout data mining jobs. The most common methods are Sequential Patterns, Artificial Intelligence, Neural Networks, Decision Trees, Regression Analysis, Association Rules, Classification, Clustering, etc.

Now it is time to brief discussion on the data mining techniques for better understanding.

2.7.1 REGRESSION

A regression analysis is used for identifying and analyzing the relationship among variables. In regression, the main goal is to determine the value of the dependent variable (output /outcome variable) on the basis of independent variables (input variables / predictors. Linear regression is the simple type of regression that used for numeric prediction. Regression analysis is usually used for forecasting and prediction.

2.7.2 ASSOCIATION RULE

According to Wikipedia [15], "Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness."

Association rule can used in shopping cart analysis to identify frequently purchase products of customers.

2.7.3 CLASSIFICATION

According to Oracle help center [16] “Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.”

Classification is a two steps process, training phase and teasing phase. In training, classification begins with a data set in which the class labels are known. The classifier model is created by applying a classification algorithm on the training data set. In testing, accuracy of the classifier is compute by suppling a new test data set in which the class lables are unknown.

2.7.4 CLUSTERING

Clustering is an unsupervised Data Mining techniques with no predefined classes. It helps to discover the natural grouping from the data set. In clustering, data are divides into groups of similar objects. The accuracy of cluster depends on method applied.

2.7.5 SEQUENTIAL PATTERNS

To determine similar patterns, regular events or trends over a business period, we can use the sequential pattern analysis technique. Using sequential analysis, businesses can discover the set of items that customers buy together and recommend customers to buy it with attactive price, based on past purchasing frequency.

2.7.6 DECISION TREES

Decision tree is a supervised machine learning technique where the data is continuously fragmented according to a certain parameter. The decision tree has two entities, decision nodes and leaves. The decision nodes are where the data is split and the leaves are the final outcomes or result. It is frequently used in operations research.

2.7.7 NEURAL NETWORKS

A neural network (NN) is a network or circuit of neurons. “Neural network is a set of connected input/output units where each connection has a weight present with it. In the learning phase, the network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples.” In data mining, neural network used for prediction or forecasting.

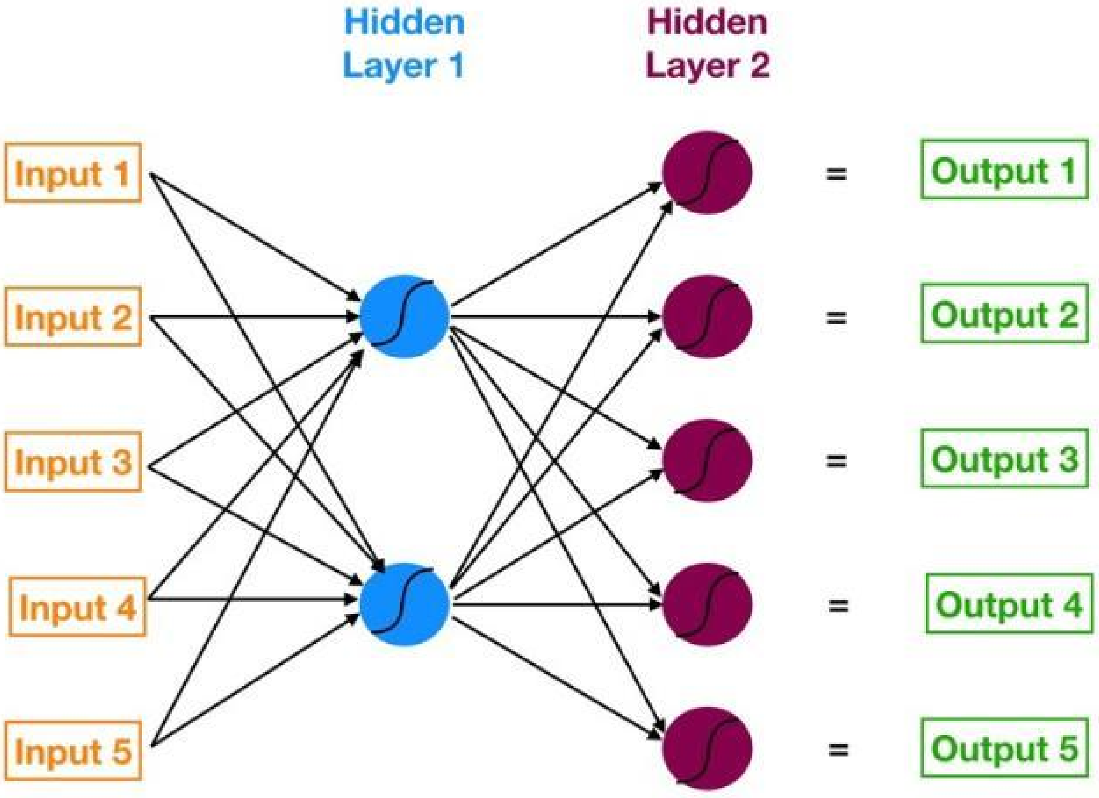


Figure 2.7.7 Neural network with hidden layers

CHAPTER 3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

The specific techniques that are used to identify, select, process, and analyze information on a research topic is called Research methodology. It shows the way how the study was conducted.

3.2 RESEARCH DESIGN

A popular data mining technique classification using J48 Decision tree algorithm was used to conduct this research. In this research, the students' past academic results, class performance, class test mark, mid-term mark, parents' social economic status, parent's academic qualification, etc. were analyzed to predict the semester final grade of a particular course.

3.3 POPULATION

630 students of Northern University Bangladesh (NUB) who were admitted in between 2016 to 2018 session in BBA Program considered as target population. The respondents in this study were first semester undergraduate students of BBA Program.

3.4 SAMPLE SIZE AND SAMPLING TECHNIQUE

132 students of BBA program of NUB were selected as a sample from 630 students form 2016 to 2018 session. All 132 respondents were first semester student of BBA Program. Random sampleing technique was follow to choose the students.

3.5 RESEARCH INSTRUMENTS

The questionnaires technique was primarily used for data collection from the students. Students' class performance records of Introduction to Computer course was retrived form NUB ERP system.

Microsoft Excel 2016 was used for performing basic operation, like shorting, calculation, merging, etc. on the collected data. Finally used Data mining software,WEKA was used for data analysis and result production.

CHAPTER 4

PROPOSED MODEL, EXPERIMENTAL RESULTS AND DISCUSSION

4.1 INTRODUCTION

The CRISP-DM methodology for data mining [17] was adopted for built classification model. It consists of five steps: (1) Gathering the relevant data, (2) preparing the data, (3) build the classifer model using appropriate algorithm, (4) evaluate the model, (5) use the model for future prediction. These steps are presented in the next subsections.

4.2 CRISP-DM (CROSS-INDUSTRY STANDARD PROCESS for DATA MINING)

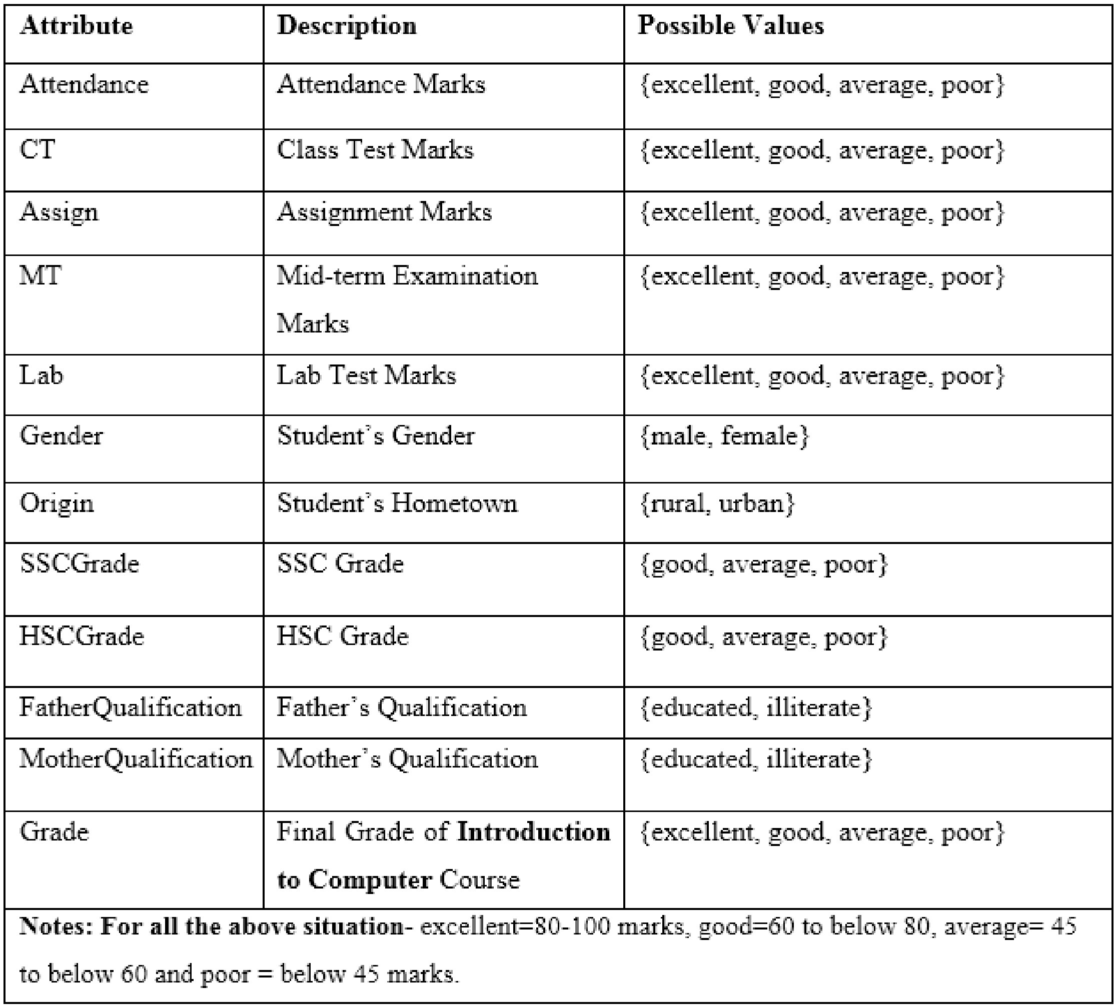
4.2.1 GATHERING THE RELEVANT DATA

132 students of BBA program of NUB were selected as a sample from 630 students form 2016 to 2018 session.

The required data were collected from the students of Department of Business Administration (BBA), Northern University Bangladesh who took Introduction to Computer course from 2016 to 2018 session. Initially, 31 attributes have been collected by questionnaire. During data preprocessing, irrelevant attributes have been removed from the dataset. Finally, only 12 relevent attributes and one class attribute have been considered for this study. Details about the attributes were presented in Table 1. Final grade of the students in the Introduction to Computer course was selected as class attribute. In this study, out of 132 records there are 42 female and 90 male students' records are available.

Total 132 records are spliting into two set (a) training set (contains 82 records) and (b) testing set (contains 50 records).

Table 4.2.1: Attributes description (with Domain values)



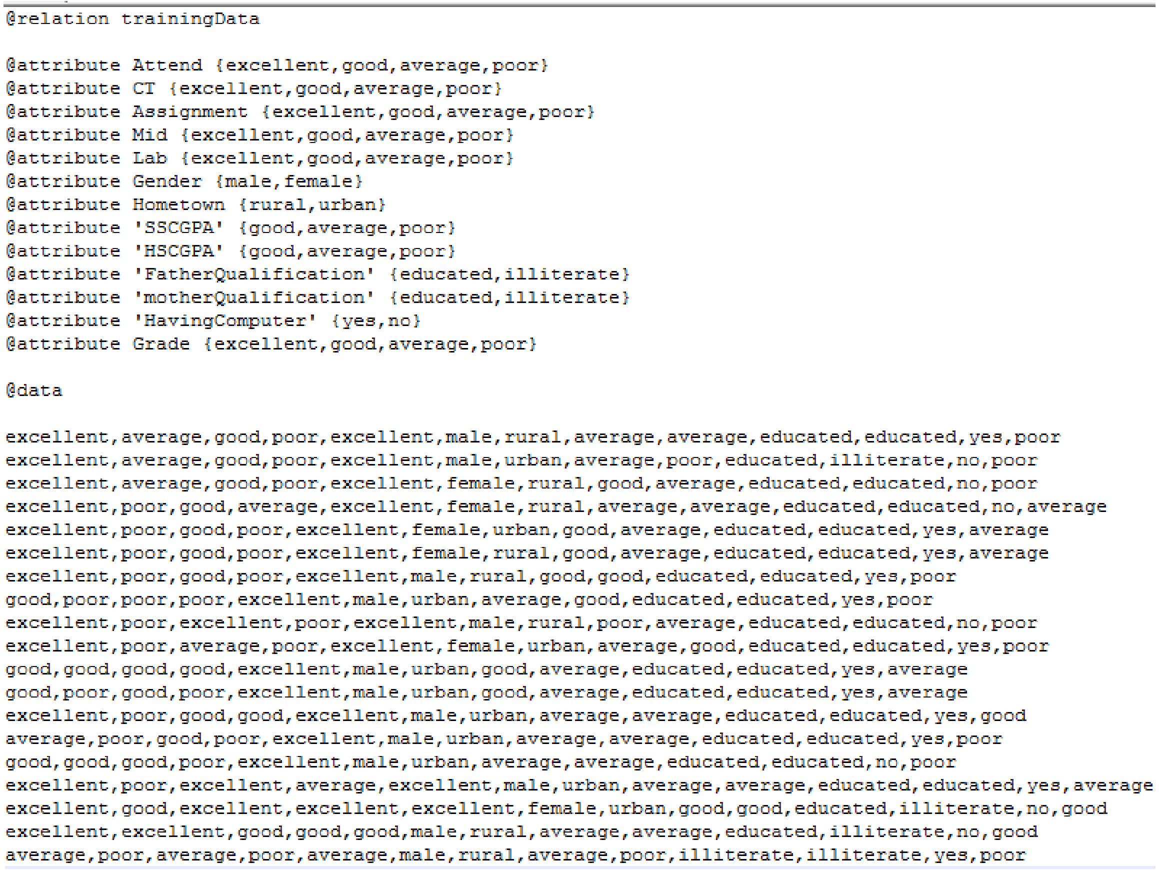
Attribute	Description	Possible Values
Attendance	Attendance Marks	{excellent, good, average, poor}
CT	Class Test Marks	{excellent, good, average, poor}
Assign	Assignment Marks	{excellent, good, average, poor}
MT	Mid-term Examination Marks	{excellent, good, average, poor}
Lab	Lab Test Marks	{excellent, good, average, poor}
Gender	Student's Gender	{male, female}
Origin	Student's Hometown	{rural, urban}
SSCGrade	SSC Grade	{good, average, poor}
HSCGrade	HSC Grade	{good, average, poor}
FatherQualification	Father's Qualification	{educated, illiterate}
MotherQualification	Mother's Qualification	{educated, illiterate}
Grade	Final Grade of **Introduction to Computer** Course	{excellent, good, average, poor}
Notes: For all the above situation- excellent=80-100 marks, good=60 to below 80, average= 45 to below 60 and poor = below 45 marks.		

4.2.2 PREPARING THE DATA

In this stage, all the collected data were entered into Microsoft Excel program. Then, closely examined the entire dataset for the missing values and/or inconsistent values. Next, the datasets are cleaned by removing the various inconsistent values and/or filling out the missing values with appropriate values. Then the excel file was saved as .CSV (Comma Separated Values) format, so that it can be use with WEKA data mining software. After loded the data set with WEKA, all the irrelevant attributes are removed from the system using Preprocess Tab of WEKA explorer.

According to Wikipedia [18] “WEKA is open source software for data mining under the GNU General public license. It is developed at the University of Waikato in New Zealand. “WEKA” stands for the Waikato Environment for knowledge analysis and it is freely available at http://www.cs.waikato.ac.nz/ml/weka. The system is written using object-oriented language java. Weka provides the implementation of state-of-the-art data mining and machine learning algorithm. User can perform association, filtering, classification, clustering, visualization, regression, etc. by using weka tool.”

The following two figures (Figure 4.2.2.1 and Figure 4.2.2.2) are snap-short of training dataset and test dataset in .arff format. The training dataset is used for model creation whereas the testing dataset is used for predict the result.



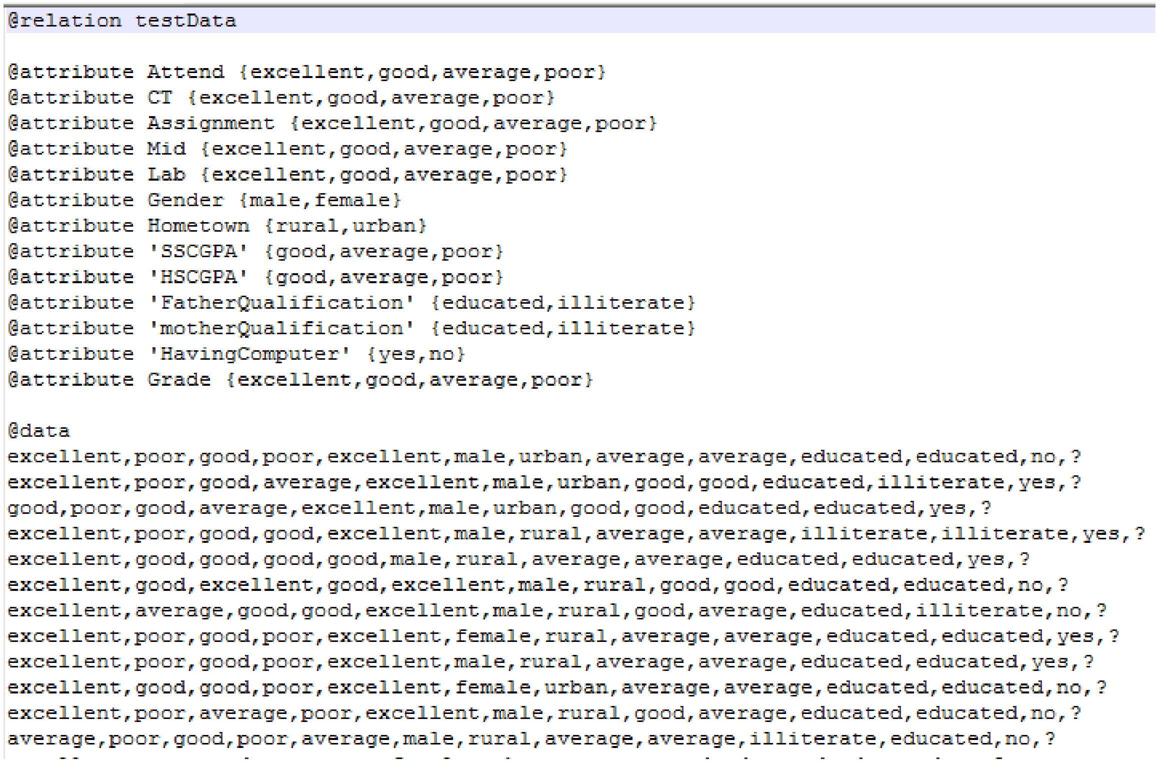
```
@relation trainingData

@attribute Attend {excellent,good,average,poor}
@attribute CT {excellent,good,average,poor}
@attribute Assignment {excellent,good,average,poor}
@attribute Mid {excellent,good,average,poor}
@attribute Lab {excellent,good,average,poor}
@attribute Gender {male,female}
@attribute Hometown {rural,urban}
@attribute 'SSCGPA' {good,average,poor}
@attribute 'HSCGPA' {good,average,poor}
@attribute 'FatherQualification' {educated,illiterate}
@attribute 'motherQualification' {educated,illiterate}
@attribute 'HavingComputer' {yes,no}
@attribute Grade {excellent,good,average,poor}

@data

excellent,average,good,poor,excellent,male,rural,average,average,educated,educated,yes,poor
excellent,average,good,poor,excellent,male,urban,average,poor,educated,illiterate,no,poor
excellent,average,good,poor,excellent,female,rural,good,average,educated,educated,no,poor
excellent,poor,good,average,excellent,female,rural,average,average,educated,educated,no,average
excellent,poor,good,poor,excellent,female,urban,good,average,educated,educated,yes,average
excellent,poor,good,poor,excellent,female,rural,good,average,educated,educated,yes,average
excellent,poor,good,poor,excellent,male,rural,good,good,educated,educated,yes,poor
good,poor,poor,poor,excellent,male,urban,average,good,educated,educated,yes,poor
excellent,poor,excellent,poor,excellent,male,rural,poor,average,educated,educated,no,poor
excellent,poor,average,poor,excellent,female,urban,average,good,educated,educated,yes,poor
good,good,good,good,excellent,male,urban,good,average,educated,educated,yes,average
good,poor,good,poor,excellent,male,urban,good,average,educated,educated,yes,average
excellent,poor,good,good,excellent,male,urban,average,average,educated,educated,yes,good
average,poor,good,poor,excellent,male,urban,average,average,educated,educated,yes,poor
good,good,good,poor,excellent,male,urban,average,average,educated,educated,no,poor
excellent,poor,excellent,average,excellent,male,urban,average,average,educated,educated,yes,average
excellent,good,excellent,excellent,excellent,female,urban,good,good,educated,illiterate,no,good
excellent,excellent,good,good,good,male,rural,average,average,educated,illiterate,no,good
average,poor,average,poor,average,male,rural,average,poor,illiterate,illiterate,yes,poor
```

Figure 4.2.2.1: A snapshot of Training Dataset that were used for create classifier model



```
@relation testData

@attribute Attend {excellent,good,average,poor}
@attribute CT {excellent,good,average,poor}
@attribute Assignment {excellent,good,average,poor}
@attribute Mid {excellent,good,average,poor}
@attribute Lab {excellent,good,average,poor}
@attribute Gender {male,female}
@attribute Hometown {rural,urban}
@attribute 'SSCGPA' {good,average,poor}
@attribute 'HSCGPA' {good,average,poor}
@attribute 'FatherQualification' {educated,illiterate}
@attribute 'motherQualification' {educated,illiterate}
@attribute 'HavingComputer' {yes,no}
@attribute Grade {excellent,good,average,poor}

@data
excellent,poor,good,poor,excellent,male,urban,average,average,educated,educated,no,?
excellent,poor,good,average,excellent,male,urban,good,good,educated,illiterate,yes,?
good,poor,good,average,excellent,male,urban,good,good,educated,educated,yes,?
excellent,poor,good,good,excellent,male,rural,average,average,illiterate,illiterate,yes,?
excellent,good,good,good,good,male,rural,average,average,educated,educated,yes,?
excellent,good,excellent,good,excellent,male,rural,good,good,educated,educated,no,?
excellent,average,good,good,excellent,male,rural,good,average,educated,illiterate,no,?
excellent,poor,good,poor,excellent,female,rural,average,average,educated,educated,yes,?
excellent,poor,good,poor,excellent,male,rural,average,average,educated,educated,yes,?
excellent,good,good,poor,excellent,female,urban,average,average,educated,educated,no,?
excellent,poor,average,poor,excellent,male,rural,good,average,educated,educated,no,?
average,poor,good,poor,average,male,rural,average,average,illiterate,educated,no,?
```

Figure 4.2.2.2: A snapshot of test Dataset that were used for result prediction

4.2.3 BUILDING AND EVALUATING THE CLASSIFIER MODEL USING J48 ALGORITHM

In this stage, the classifier was built using the J48 decision tree method to predict student's grade before the semester final Examination. The decision tree is a graphical technique for doing logical modeling. It is a flow chart like tree structure that represents the various conditions and the subsequent possible actions.

The decision tree method is fast and it can easily converted to classification rules. The decision tree method is constructed using the information gain metrix which determines the most useful attribute. The information gain depends on the entropy measure. The gain ratio is used to rank attributes and to build the decision tree where each attribute is located according to its gain ratio.

For classifier model construction the J48 (C4.5 algorithm implemented in WEKA) decision tree algorithm has been used, it was published by Ross Quinlan in 1993.

In this research, the attribute MT (The Mid-term Marks) has the highest gain ratio, that's why MT became the root of the decision tree. The process is repeated for the remaining attributes to built the complete decision tree. Popular machine learning and data mining software WEKA was used to built J48 decision tree classifier.

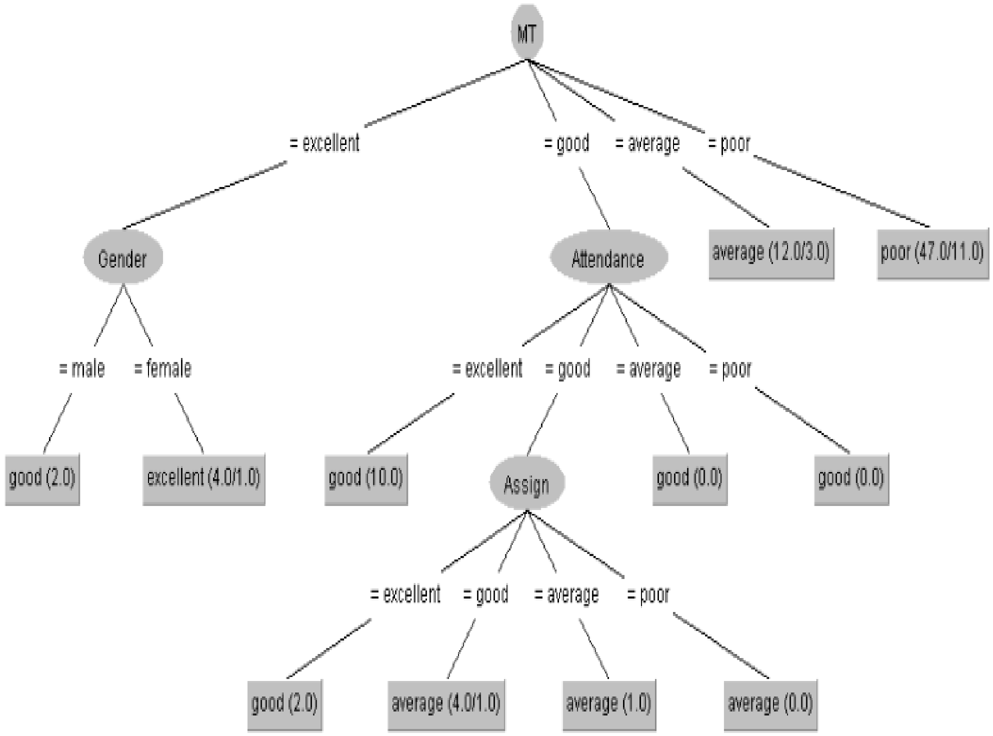
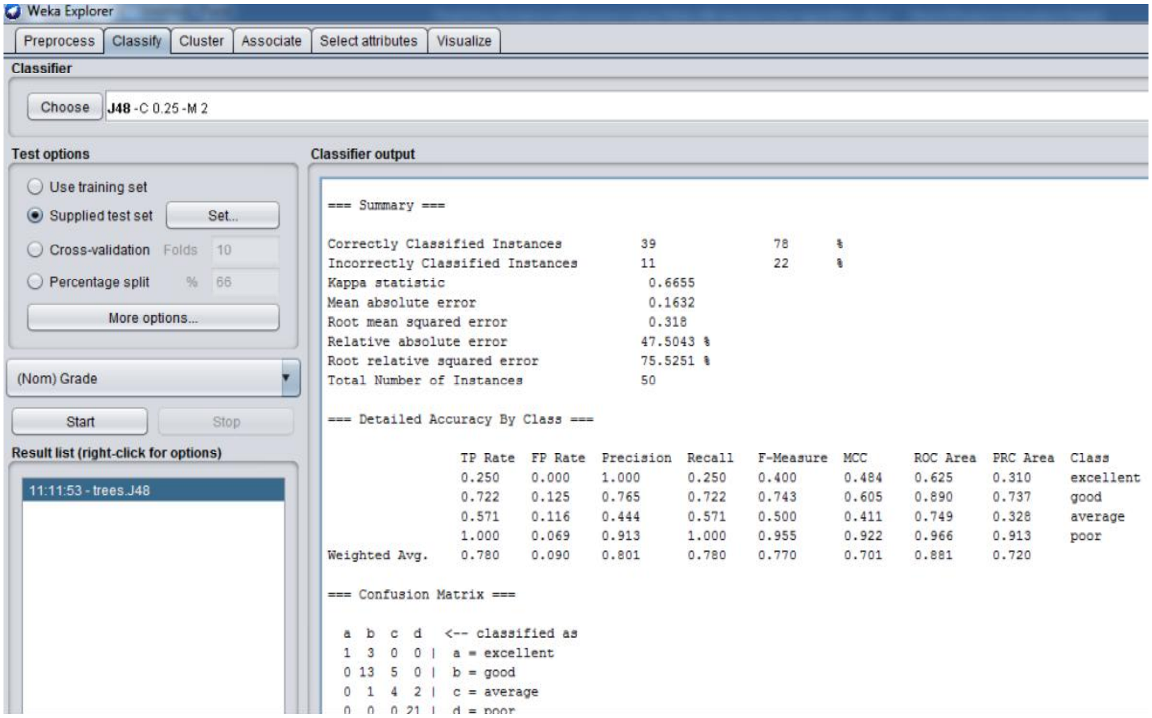


Figure 4.2.3.1: Decision Tree Construction



J48 -C 0.25 -M 2

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

(Nom) Grade

11:11:53 - trees.J48

```
=== Summary ===

Correctly Classified Instances          39               78      %
Incorrectly Classified Instances        11               22      %
Kappa statistic                          0.6655
Mean absolute error                      0.1632
Root mean squared error                  0.318
Relative absolute error                 47.5043 %
Root relative squared error             75.5251 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.250    0.000    1.000      0.250    0.400      0.484    0.625     0.310     excellent
                 0.722    0.125    0.765      0.722    0.743      0.605    0.890     0.737     good
                 0.571    0.116    0.444      0.571    0.500      0.411    0.749     0.328     average
                 1.000    0.069    0.913      1.000    0.955      0.922    0.966     0.913     poor
Weighted Avg.    0.780    0.090    0.801      0.780    0.770      0.701    0.881     0.720

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  1  3  0  0 |  a = excellent
  0 13  5  0 |  b = good
  0  1  4  2 |  c = average
  0  0  0 21 |  d = poor
```

Figure 4.2.3.2: Summary of result using the J48 algorithm

4.2.4 USEING THE MODEL FOR FUTURE PREDICTION

To evaluate the classifire model, a separate data set consists of 50 records is supplied to test the model accuracy. J48 decision tree algorithm was applied for classification.

The results of the J48 algorithm depicted in Table 4.5. The overall accuracy of this model is 78%. In this classifier, 39 students' final grade are correctly predicted out of 50 students.

Table 4.2.4.1: Showing the accuracy of J48 decision tree algorithm.

Decision Tree	No. of Correctly classified instances	No. of Incorrectly classified instances	Accuracy %
J48	39	11	78

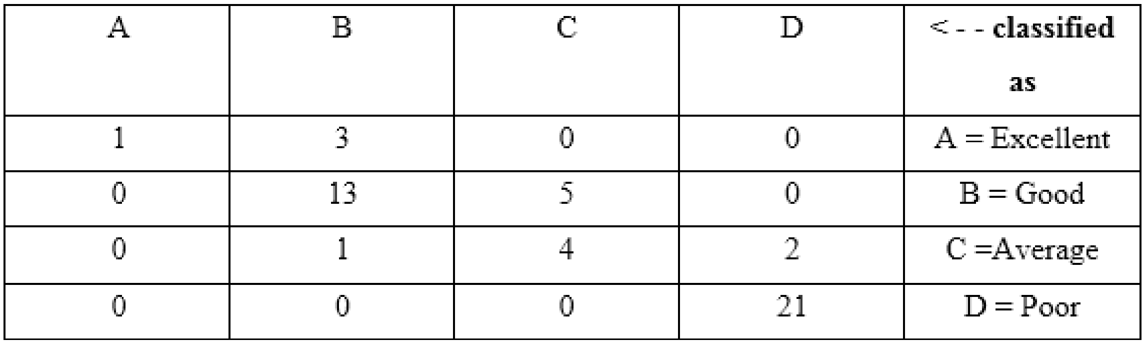
4.3 CONFUSION MATRIX USING J48 ALGORITHM

According to Wikipedia [19], “A confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).”

The confusion matrix shows the summary of the prediction result. So, it can say that, the confusion matrix shows when your model is confused to makes predictions.

Table 4.3.1 shows the confusion matrix of the J48 algorithm on the test data set.

Table 4.3.1: Confusion matrix for the J48 algorithm.



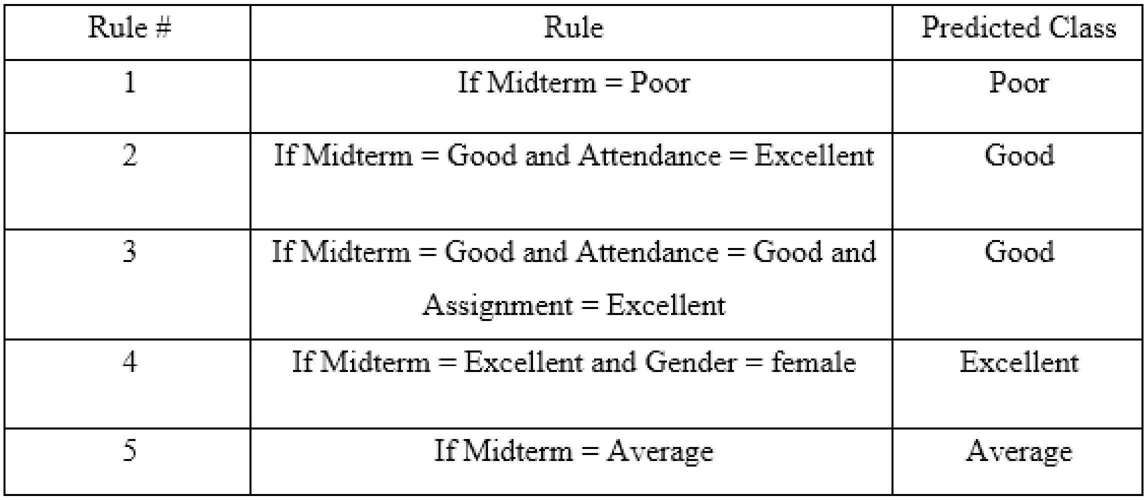
A	B	C	D	< - - **classified as**
1	3	0	0	A = Excellent
0	13	5	0	B = Good
0	1	4	2	C =Average
0	0	0	21	D = Poor

The diagonal line from top left to bottom-right of the matrix shows the correctly classified values (1+13+4+21) =39. All the other cases, it makes a mistake and unable to correctly classify the data.

4.4 CLASSIFICATION RULES GENERATED BY J48 ALGORITHM

Based on the decision tree, we can easily extract the classification rules by tracing from root to each leaf node in the tree. Table 4.4.1 shows the all possible rules and the predicted class of the decision tree.

Table 4.4.1: Set of Classification Rules extracted form decision tree.



Rule #	Rule	Predicted Class
1	If Midterm = Poor	Poor
2	If Midterm = Good and Attendance = Excellent	Good
3	If Midterm = Good and Attendance = Good and Assignment = Excellent	Good
4	If Midterm = Excellent and Gender = female	Excellent
5	If Midterm = Average	Average

4.5 MAJOR FINDINGS OF THIS STUDY

Finally, it is time to compare the expected result with the actual result. This study shows a classification technique to prediction of grade of a student. From the above analysis, it is clear that, student final grade of a particular course in only depends on the current performance of that student (i.e. Class Attendance, Assignment and midterm marks of that course). There is little / no impact of the previous result, socio-economical condition of parent, parent education. So, a student with poor results in SSC and HSC, he/she has also equal opportunity to do better.

Then, the extracted knowledge can help the course teacher to make decision for the new students. It also help the course teacher to identify low performer students.

So, if the respective course teacher takes necessary actions to help the low performing student, so they can overcome their problems. As a result, the dropout rate can be decrease.

CHAPTER 5:

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

This study proposed a decision tree based classification model for forecasting final grade of **Introduction to Computer** course. The J48 decision tree algorithm was selected for model construction. The model obtained an accuracy of 78%. This early perdiction of students' grade can help the university authority to findout low performing students. This study can also help them to take proper preventive measure to prevent students from failure.

5.2 FUTURE WORK

This study was only based on the data of a single course (Introduction to Computer) to predict the student's final grade. Also, the sample size is very limited for data mining job. In future, this research can be enhanced by including other related factors to predict the final CGPA of the students using first year data.

REFERENCES

[1] Data mining <<https://en.wikipedia.org/wiki/Data_mining >>, last accessed on 12-10-2019 at 09:00 AM.

[2] Z. N. Khan, “Scholastic achievement of higher secondary students in science stream”, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[3] Galit.et.al, “Examining online learning processes based on log files analysis: a case study”. Research, Reflection and Innovations in Integrating ICT in Education 2007.

[4] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, “Mining student data using decision trees”, International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[5] U. K. Pandey, and S. Pal, “A Data mining view on classroom teaching language”, (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[6] Monika Goyal, Rajan Vohra2, Applications of Data Mining in Higher Education, 2012.

[7] Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal, Mining Education Data to Predict Student’s Retention: A Comparative Study, 2012.

[8] K.Shanmuga Priya, A.V.Senthil Kumar, Improving the Student’s Performance Using Educational Data Mining, 2013.

[9] Ahmad, F., N.H. Ismail, and A. Abdulaziz, *The Prediction of Students’ Academic Performance Using Classification Data Mining Techniques.* Applied Mathematical Sciences, 2015. **9**(129): p. 12.

[10] Sumitha, R. and E.S. Vinothkumar, *Prediction of Students Outcome Using Data Mining Techniques.* International Journal of Scientific Engineering and Applied Science (IJSEAS), 2016. **2**(6): p. 8.

[11] Saa, A.A., *Educational Data Mining & Students’ Performance Prediction.* (IJACSA) International Journal of Advanced Computer Science and Applications, 2016. **7**(5): p. 9.

[12] Khasanah, A.U. and Harwati, *A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques.* IOP Conf. Series: Materials Science and Engineering, 2017. **215**(012036): p. 7.

[13] Defination of Data Mining << https://economictimes.indiatimes.com/definition/data-mining >>, last accessed on 11-10-2019 at 09:00 AM.

[14] Data pre-processing << https://en.wikipedia.org/wiki/Data_pre-processing>>>>, last accessed on 11-10-2019 at 10:00 AM.

[15] Association rule learning <<https://en.wikipedia.org/wiki/Association_rule_learning>>, last accessed on 11-10-2019 at 11:00 AM.

[16] <<https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004>>, last accessed on 11-10-2019 at 11:10 AM.

[17] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

[18] WEKA, Machine learning << https://en.wikipedia.org/wiki/Weka_(machine_learning) >>, last accessed on 01-09-2019 at 11:00 AM.

[19] Confusion matrix << https://en.wikipedia.org/wiki/Confusion_matrix >>, last accessed on 01-11-2019 at 11:00 AM.