STUDENT RESULT PREDICTION USING DATA MINING TECHNIQUE

BY

MD.MARUF ISLAM
ID: 143-15-4426

AND

MD. HABIBUR RAHMAN
ID: 143-15-4413

This Report Presented to the Department of CSE of Daffodil International University in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Afsara Tasneem Misha
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
November 3, 2018

APPROVAL

This Project titled "**Student Result Prediction Using Data Mining Technique**" submitted by Md. Maruf Islam, ID: 143-15-4426 and Md. Habibur Rahman, ID: 143-15-4413 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 B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on 03 November 2018

BOARD OF EXAMINERS

_____

(Name) **Chairman**
Designation
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

(Name) **Internal Examiner**
Designation
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

(Name) **External Examiner**
Designation
Department of CSE
Jahangirnagar University

DECLARATION

We hereby declare that this research-based project has been done by us under the supervision of **Ms. Afsara Tasneem Misha, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

_____

Ms. Afsara Tasneem Misha
Lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

(Md. Maruf Islam)
ID: - 143-15-4426
Department of CSE
Daffodil International University

_____

(Md. Habibur Rahman)
ID: - 143-15-4413
Department of CSE
Daffodil International University

ACKNOWLEDGMENT

First of all, we would like to express our heartiest thanks and gratefulness to Almighty Allah for His divine blessing makes us possible to complete our research-based final year project successfully.

We really grateful and wish our profound our indebtedness to **Ms. Afsara Tasneem Misha**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of data mining influenced us to carry out this thesis. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

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

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

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

The intent of the research paper is to presents a brief review of the use of a large amount of student data that was kept by the educational institutions. Using the data mining technique student result can be predicted from these data. The purpose of this research paper is to provide an overview of how an educational institution can predict the student result, what tools are used on this purpose, what kind of data is needed for the prediction, how students can visualize their current progress and also can see their predicted result. Finally, this paper will propose a recommended system for the students and the institutions to improve the student progress depending on some specific data such as, predict the student final grade depending on the student's assignment, presentation, attendance, quiz, mid, final mark on a particular subject and suggested some predictive modeling such as Decision tree, Naïve Bayes. Compare the predictive modeling with each other and also find out the impact the various variable in student result prediction.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURES **PAGE**

LIST OF TABLES

TABLES **PAGE**

CHAPTER 1
Introduction

1.1 Introduction

Data mining in the sector of education is relatively new. It is mainly used in the business world to find some hidden data using some algorithm for business purpose. Nowadays, Data mining is used in the educational sector. The purpose of using data mining in the educational sector is to develop a model. With the help of the model, student success can be analyzed or predicted. From the concern of educational institutions, their goal is to make improvement of the quality of education. Students result is a higher indicator for any educational institution. Student success is mainly the institution's success. The process of the creation of human capital is a continuous process of analysis. So, the result prediction of student's is essential for an educational institution. In the student grade prediction system, it can predict that which student is going to get which grade. If we can predict student grade, then we know how many students are going to be achieved a good result and how many can't. We also can take the necessary steps to minified the number of students who do not get the good result. Thus, data mining is helpful in the educational sector. With the help of data mining, we can solve the problem of predicting a student's academic grade.

1.2 Motivation

Let's assume, a situation where a new student admitted to an educational institute. It takes some time for the student to cope up with the situation. In the meantime, there are some students those are not careful about their study. So, they can't do well in the examination. But when they realize their fault, their exam is gone. If we have a system that can predict the student result before the final examination, then it will be helpful for the students. Again, for an institution, it has many students so, it is not possible to check the student result manually and take necessary steps. But if the system can notify the authority of the institution about the percentage of a student's grade that they are going to be achieved then the authority can take the necessary steps. Students also see their predicted result. Then they also become careful about the examination.

Objects

- Predict the student result.
- Show the predicted result to the student.
- Analysis of the different model for predicting student result.
- Find out the most significant attributes that are important for student result prediction.
- Compare the prediction result using algorithms.
- Visualize the comparison of result that we get from analysis using a various type of model

1.3 Rationale of the Study

Data mining is basically a computing method of processing data which is now successfully applied in the educational sector. With the help of data mining algorithm, we can find a hidden pattern or knowledge that extracted from the huge amount of data. Without only predicting the student result, we can also extract much knowledge from the data. We also find a student economic condition from student data then analyze it with other data. We can analyze the behavior of the students from the data. We can find any difference in terms of behavior between a good student or a bad student. We can find out the reason behind the student success if we have the data of student's study time, habit etc. Educational data mining is widely used to improve student performance. Not only educational sector, these type of data mining technique used in customer behavior analysis, the market value of any product etc.

1.4 Research Questions

A research question fundamental and critical part of any research project. It needs to be centered, cleared, focused and trend to center to the analysis. The research questions are always related to the research topic and may be answered directly through the analysis of data. The research questions that we are trying to answer throughout the paper are given below.

- The first and foremost question is which data we want to use to complete our analysis, mainly the attribute?

- The next question is how to prepare the data set from the raw data set that fit to the analysis?
- Find out the algorithm which will offer the best result for the analysis?
- How to compare the one algorithmic program with another algorithmic program?
- How can we become skilled to choose the right tool for the data set?

1.5 Expected Outcome

- Best model that describes the prediction system.
- Grade predict without the final term result.
- Discover the reason behind the bad result.
- Define the attribute on which the prediction result depends.

1.6 Report layout

This research paper consists of six chapters. The chapters are Introduction, Literature review, Research methodology, Analysis of the Results and Discussions, Conclusion and future scope of the study

Chapter 1, Introduction; try to discuss the motivation behind the research work, simplify the questions that arise during the analysis, pointed out the expected outcome of the analysis.

Chapter 2, Literature review; introduction, discuss the related work similar to the project, discuss the research summary, find out the scope of the problem during the analysis, challenges that faced during the research work.

Chapter 3, Research methodology; introduction, briefly discuss the KDD process, data collection, data selection,data preprocessing,data mining algorithms, use weka tool for classification.

Chapter 4, Analysis of the Results and Discussions; result analysis, statistical analysis of the numeric attributes, number of instances of a class attribute, comparison of different algorithms using cross-validation, a training set, and testing set, discuss challenges that we face.

Chapter 5, Conclusion and Future scope; this chapter provides the summary result that we get from the previous chapter and in future scope section we discuss what we can do with this research in the future.

CHAPTER 2

Literature Review

2.1 Introduction

In the higher education system, the student result information is stored in the database. There are various types of student data stored in the database. To properly use of this huge amount of data on the other hand extract knowledge from these data, we can use a new technology called data mining.

Data mining is a process of sorting the data which is actually needed to extract hidden patterns from a large database. It also can be said that data mining is a way of finding the patterns and relations among the data in a large database[3].There is various type of techniques to extract the hidden pattern from the data. Educational data mining is a new and advanced technique that used in the educational institution to extract knowledge from the various type of information. It can be applied to the data that are related to the educational system. Romero and Venture [1] find out the cycle of applying data mining in educational systems.

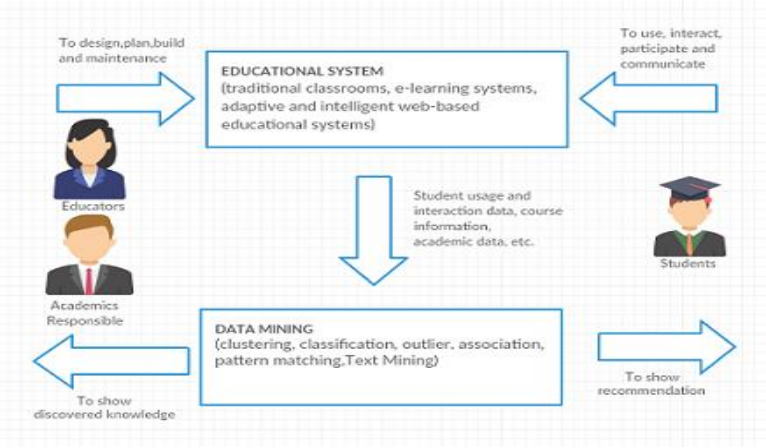

Figure 2.1: Cycle of data mining in the educational sector

The data can be collected through various ways such as from the database of the institution, from the students, from the teachers. These data consist of a various type of student information such as their personal information or academic performance. Data mining uses many formulae to extract a knowledgeable pattern such as decision tree, neural networks, naïve Bayes, K-Nearest algorithm etc.

For any educational institution, student performance is a major concern. They try to figure out the reason behind the bad result of the student. They also try to make a performance prediction that can predict student result. So, that they can manage the students and take proper steps for the students who do not achieve a good result.In this paper, we try to predict the student result based on their attendance, presentation, assignment, quiz, mid-term, final mark and predict their final grade on a particular subject.

2.2 Related work

Devesia, T.P, and Hedge [4] said that increasing the number of students who were dropping out from an institute is affecting the reputation of that institute. They find that the system was incapable of analyzing the student data. The system only stores and access the information.

Ktona, Xhaja, and Ninka [5] describe that data mining is a special area of computer science that is used widely in education. At the same time, it can provide some solution to increase the quality of education. Generally, data mining is a combination of some tools of the statistic with DBMS and AI.

Mishra, Kumar, and Gupta in [4] used different classification techniques to build a performance model based on student's social integration, various emotional skills, academic integration. Two algorithms are applied in this process named Decision tree-J48 and Random Tree to predict third-year student semester performance.

Jacob, et al. in [6] said that the C5.0 classification has one hundred percent accuracy in classification. He showed the comparison will be done between the proposed algorithm and the benchmark algorithm like the Decision Tree and Naïve Bayes.

Varun Kumar and Anupama Chadda [8] used one of the data mining technique named association rule mining. They enhancing the quality of students' performances at Post Graduation Level.

2.3 Research Summary

In the research summary, we clearly say that the research will be recommended a system that can be extremely helpful for the student because the system will predict the student result for a specific subject. Such as, in this research we take data structure as the subject. We collect the data such as assignment, mid-term, final, presentation, attendance, quiz mark on that subject. From this data, we can predict the final grade of the student. The authority can be also used this system to find out the specific student for special consideration.

2.4 Scope of the problem

- Format changing problem of the data.
- Divide the dataset into training and testing set.
- Finding an algorithm that suits best to the dataset.
- Analyze the different type of attribute with the class attribute.

2.5 Challenges

Challenges that we face in this section

- To collect the data for our experiment
- The data are not well maintained and not suitable for the experiment

CHAPTER 3

Research Methodology

3.1 Introduction

The methodology is a way of collecting data or information and it contains several concepts and theory. The full process of data mining based on Knowledge Discovery and Database (KDD). The KDD process is illustrated in figure 3.1 [9].

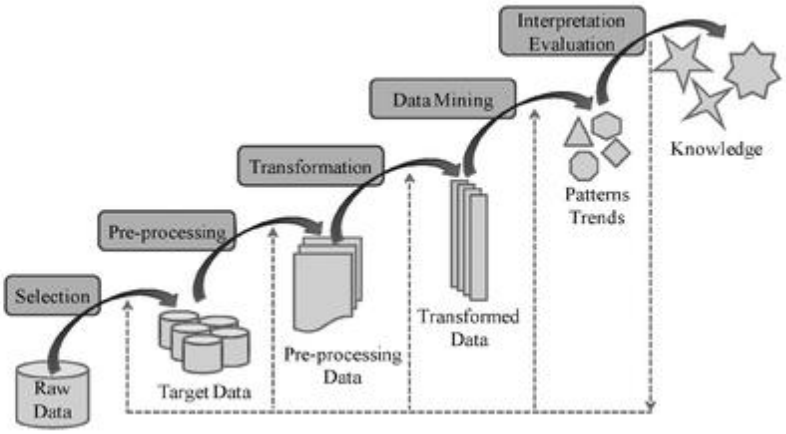

Figure 3.1: KDD process for data mining

The methodology that we are using in the research paper is KDD. It contains several parts such as data selection, data pre-processing, data transformation, some data mining techniques for finding the pattern, then extract knowledge. So that, we can say that KDD is a process where first select the data from a raw database, then pre-process the data to applying the data mining rule on the dataset, depending on the result of data mining we can analyze a pattern and from the pattern we can extract some knowledge. By using the knowledge, we can get the information that we want throughout our process.

3.2 Steps of the KDD process

There are several steps in the KDD process. Thus, we conduct our research depending on the KDD process. So, we try to discuss the process according to our research work.

Data selection

For any data mining research, the main element is data. Without the data, no experiment can be done. From the data, we can get valuable information. So we collected the 260 students data on a course named data-structure from our respected supervisor. We conduct our research on those data.

Finally, we collect the raw data that we need for our research work. Then we eliminate the attribute that we don't need for the experiment. For example, in the raw datasets that illustrated in figure 3.2, we have several attributes those are student name, attendance, three quiz mark, an average of the quiz mark, presentation, assignment, mid-term, mid-term improvement, Up to mid-term, absent checker, final, grand total, grade, grade-point. But for our research, we can not take all the value for the experiment. We choose the only value that is related to our research and we also have to find the class value. In our case, the final grade is the class. But we achieve the class value after the data pre-processing step.

Student Id	Name	ATTM	Q1	Q2	Q3	AVG (Quiz)	Presn	Assign	MT	MT Im	UTM	Abs	WH	Final	GT	GRD	GP
141-15-3353	Md. Shakil Shahriar Khan	2.72	1.5	0	0	0.50	0	0	9.5	0	12.72	☑	☐	0	13	I	0
151-15-4890	Pollob Biswas	0.78	0	0	0	0.00	0	0	0	0	0.78	☑	☐	0	1	I	0
151-15-5052	Md. Rezoanul Karim Ullash	6.22	6.5	12	11.5	10.00	6.3	3.8	15.5	0	41.82	☐	☐	27.5	70	A-	3.5
152-15-5525	Md. Aktaruzzaman Bhuiyan Rasel	1.17	0	0	0	0.00	0	0	0	0	1.17	☑	☐	0	2	I	0
153-15-6674	Nurnoshin Sultana	3.11	8	8	11	9.00	6.4	3.2	12.5	0	34.21	☐	☐	36	71	A-	3.5
161-15-6883	Umme Sharmin Ritu	2.33	2	0	0	0.67	6.2	0	12.5	0	21.7	☐	☐	24	46	C	2.25
161-15-7082	Asrafull karim	3.5	2	10	11	7.67	5	0	8	0	24.17	☐	☐	25	50	C+	2.5
161-15-7621	Md. Asaduzzaman	3.5	6	3	0	3.00	0	0	8.5	0	15	☑	☐	0	15	I	0
161-15-7631	Golam Mahmud Roman	3.5	0	5	0	1.67	0	0	17	0	22.17	☐	☐	27	50	C+	2.5
171-15-8557	Redwan Hossain	7	12	9.5	13	11.50	7	4.2	24.5	0	54.2	☐	☐	32	87	A+	4
171-15-8585	Md. Rifat Ahmed	6.61	7	11.5	14	10.83	6.3	3	11	0	37.74	☐	☐	26.5	65	B+	3.25
171-15-8607	Md. Ibrahim Hossen	7	4	9	13	8.67	6.5	3	19.5	0	44.67	☐	☐	34.5	80	A+	4
171-15-8625	Md. Ramijul Islam	6.22	8	9	15	10.67	6.2	3.5	13.5	0	40.09	☐	☐	29	70	A-	3.5
171-15-8658	Udoy Deb Das	7	6	0	12	6.00	6.8	3.5	7.5	0	30.8	☐	☐	34.5	66	B+	3.25
171-15-8673	Ayeasha Akter Liza	6.61	8	9	10	9.00	6	2.5	10.5	0	34.61	☐	☐	29.5	65	B+	3.25
171-15-8708	Ragib Muid	6.22	6	7	13	8.67	6	2.8	9	0	32.69	☐	☐	11.5	45	C	2.25
171-15-8735	Nayma Sultana Nipu	6.22	10	10	12	10.67	7.4	4	24.5	0	52.79	☐	☐	35	88	A+	4
171-15-8737	Redwan Sharafat Kabir	6.61	9.5	5	14	9.50	7.2	4	19.5	0	46.81	☐	☐	22.5	70	A-	3.5
171-15-8743	Tasmima Sultana Reya	7	8.5	12	12	10.83	7	3.8	16.5	0	45.13	☐	☐	35	81	A+	4
171-15-8751	Ahmed Shihab Muhibullah	7	6	8	12	8.67	7.1	4	7.5	16.5	43.27	☐	☐	40	84	A+	4
171-15-8787	Md. Abdullah Al Mamun	7	9	15	13	12.33	6.7	4.2	17.5	0	47.73	☐	☐	36.5	85	A+	4
171-15-8795	Alamin Hossain	7	11.5	9	13	11.17	6.9	4	18	0	47.07	☐	☐	38	86	A+	4
171-15-8808	Syed Rakesh Uddin	6.61	10	11	13	11.33	6.8	4	15.5	0	44.24	☐	☐	31	76	A	3.75
171-15-8813	Faria Nishat Khan	6.61	8	15	12	11.67	6.9	4	20.5	0	49.68	☐	☐	32	82	A+	4

Figure 3.2: Raw datasets for the experiment

For the experiment, we need only 7 attributes. Those are attendance, the average mark of the quiz, presentation, assignment, mid-term, final mark and grade.

After selection the data we find a target dataset that illustrated in figure 3.3.

	A	B	C	D	E	F	G
1	Attendance	Presentation	Assignment	Quiz	Mid Term	Final	Grade
2	0.32	0	0	0	0	0	I
3	0	0	0	0	0	0	I
4	1.27	0	3	0	13	0	I
5	2.23	0	3	0	6.5	18	F
6	6.05	6.2	2	13.83	12.5	34	A
7	6.68	6.8	4	11.5	21.5	31.5	A+
8	7	6	3.5	12.67	14	33	A
9	6.36	5.8	3	8	8	23	B-
10	5.09	6.6	4	9	7	32.5	B+
11	7	5.6	3	10.17	11	24	B
12	5.41	6.4	3	8	15	33	A-
13	3.5	0	0	0	17.5	0	I
14	4.14	0	0	3.5	14.5	0	I
15	6.36	6.3	3.5	9.33	13.5	31	A-
16	3.5	6.2	3	9.67	20	28.5	A-
17	5.73	6.7	1	9.17	4.5	29	B-
18	6.68	6.8	4	11.33	15.5	35	A+
19	6.68	5.9	3	10	13	25.5	B+
20	6.36	5.5	3	11.83	5	23.5	B-

Figure 3.3: Target data set for the experiment

Pre-processing of the Data

Pre-processing is a crucial part of the KDD process. In this phrase, we have to find out if there is any missing attribute in the data set. In our target dataset, there is no missing value.We selected the grade attribute as the class value of our prediction model. For this, we divided the class value into five categories. We can see that from table 3.1.

Table 3.1: Converting the grade attribute into the categories

Grade	Category
A+	Excellent
A, A-	Good
B, B+	Average
B-, C+, C, D	Below Average
F, I	Fail

After the pre-processing phrase, we get a data set that we can use for data mining phrase. Figure 3.4 illustrated the data set after pre-processing phrase.

	A	B	C	D	E	F	G
1	Attendance	Presentation	Assignment	Quiz	Mid	Final	Grade
2	0.32	0	0	0	0	0	Fail
3	0	0	0	0	0	0	Fail
4	1.27	0	3	0	13	0	Fail
5	2.23	0	3	0	6.5	18	Fail
6	6.05	6.2	2	13.83	12.5	34	Good
7	6.68	6.8	4	11.5	21.5	31.5	Excellent
8	7	6	3.5	12.67	14	33	Good
9	6.36	5.8	3	8	8	23	BelowAverage
10	5.09	6.6	4	9	7	32.5	Average
11	7	5.6	3	10.17	11	24	Average
12	5.41	6.4	3	8	15	33	Good
13	3.5	0	0	0	17.5	0	Fail
14	4.14	0	0	3.5	14.5	0	Fail
15	6.36	6.3	3.5	9.33	13.5	31	Good
16	3.5	6.2	3	9.67	20	28.5	Good
17	5.73	6.7	1	9.17	4.5	29	BelowAverage
18	6.68	6.8	4	11.33	15.5	35	Excellent
19	6.68	5.9	3	10	13	25.5	Average
20	6.36	5.5	3	11.83	5	23.5	BelowAverage
21	4.14	6.7	4	10.17	11.5	27.5	Average
22	6.68	6.7	4	11.17	9	31.5	Good
23	5.09	5.4	4	11.67	15.5	27.5	Good
24	6.68	6.8	4.5	13.83	15	22.5	Good
25	5.73	5.9	3	7.17	3.5	29	BelowAverage

Figure 3.4: Target dataset after pre-processing phrase

Now, after the pre-processing phrase data is transformed for applying data mining algorithm on the data. Data mining and other phrase are described in the future section of this chapter. From table 5.2 we can see the description of the attributes that we select for the study.

Table 3.2: Provides the attribute information of the student

Attribute Name	**Description of the attribute**
Attendance	Students attendance mark
Presentation	Students presentation mark
Quiz	Average quiz mark of the student
Mid	The mark that student get in the mid-term exam
Final	The mark that student obtain from the final exam
Assignment	The mark the student get after submitting the assignment
Grade	This is the class attribute. We want to predict this attribute depending on the above attributes

3.3 Use Weka for the Classification process

Data mining (Classification)

In KDD process, the third step is data mining. Various kinds of data mining algorithm used in this phrase to discover the hidden pattern or the information from the data set.In this phrase, we analyze the data using the data mining algorithm.

The data mining process occurred in two ways. One of them is Descriptive Analysis and another in Predictive Analysis.

Descriptive analysis used when we want to descriptive information from the summarized information and the information must be understood by the human.There are three basic algorithms for descriptive analysis.The algorithms are

- Clustering
- Association rule mining
- Sequential rule mining

Predictive analysis is used when we want to predict one variable depending on the other variables. In this analysis, we try to figure out what would possibly happen. The predictive analysis gives the probability of the future outcome. There are some algorithms for predictive analysis. The algorithms are

- Classification
- Regression
- Deviation detection

For our prediction, we use classification Algorithms.

Using Weka tools

Weka is a collection of machine learning algorithms and data processing tools.Weka is a software and it is written in java. It contains many tools for data pre-processing, classification, regression, clustering and association rules. For complete the data mining process in KDD, we need to select and use many algorithms. We complete this process using Weka.

Classification Algorithm

In data mining classification is a technique and it basically based on machine learning.It is used to classify each of the items in the data set. And in the dataset, there is always a data set that is called the class. It makes use of some of the mathematical technique like a decision tree, neural network, linear programming etc. There are some algorithms for classification and the algorithms are decision trees, Neural Networks, Genetic algorithms, Bayesian networks, Rule-based induction. There are some steps for applying the classification rule. First of the all, from the entire dataset, about 60% to 80% data is used for creating the training dataset. It takes the dataset for well-known output values and uses the dataset to create our model. After creating the training dataset,we take the rest of the dataset and divided them two or more sets to creating test set. Then we compare the two datasets to test the accuracy of our model.

In our research, we want to predict the student result using classification. For this purpose, we need to find out the models using a classification algorithm. So, we apply two algorithms of classification. They are Naïve Bayes and Decision tree. Decision tree called j48 in weka.

First of all, we load our datasets in weka. After loading all the data it will be looked like figure 3.5 in weka.

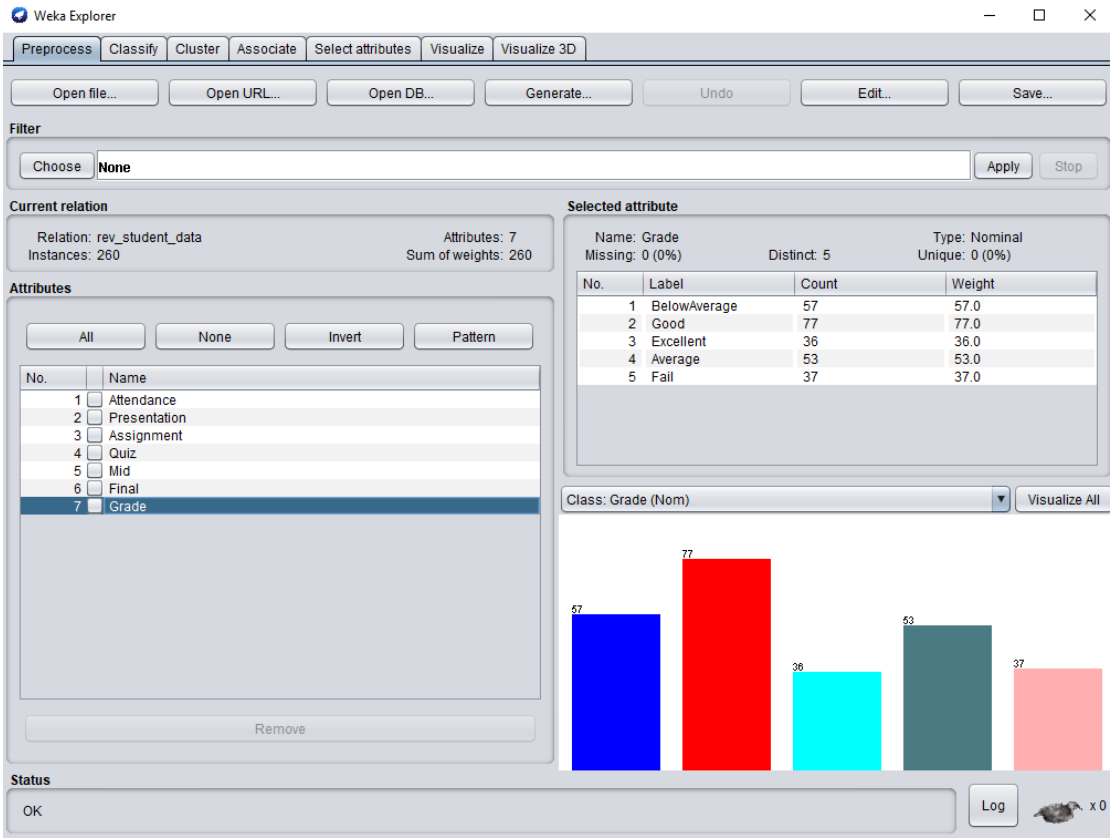

Figure 3.5: Student result prediction classification in weka

From figure 3.5, we can see that there are 260 instances and 6 classes. Among all the attributes the class attribute needs to be nominal. The classification process mainly depends on that nominal attribute. In our case, this class attribute is grade and the attribute is nominal. If the class attribute is not nominal then virtually it is not possible to apply the classification algorithm. After the loading, all data in weka if we want to apply some data pre-processing technique before classify we can do that. The weka also shows some vital information for all the attribute in figure 3.5. Such as, for grade attribute, it shows the type of the attribute, distinct values of that attribute, count value of each distinct value and their weight. It also indicates if there is any missing value in the attribute.

For classify the data click on the classify icon that shown beside the pre-processor icon in weka. Then we get a screen like a figure 3.6

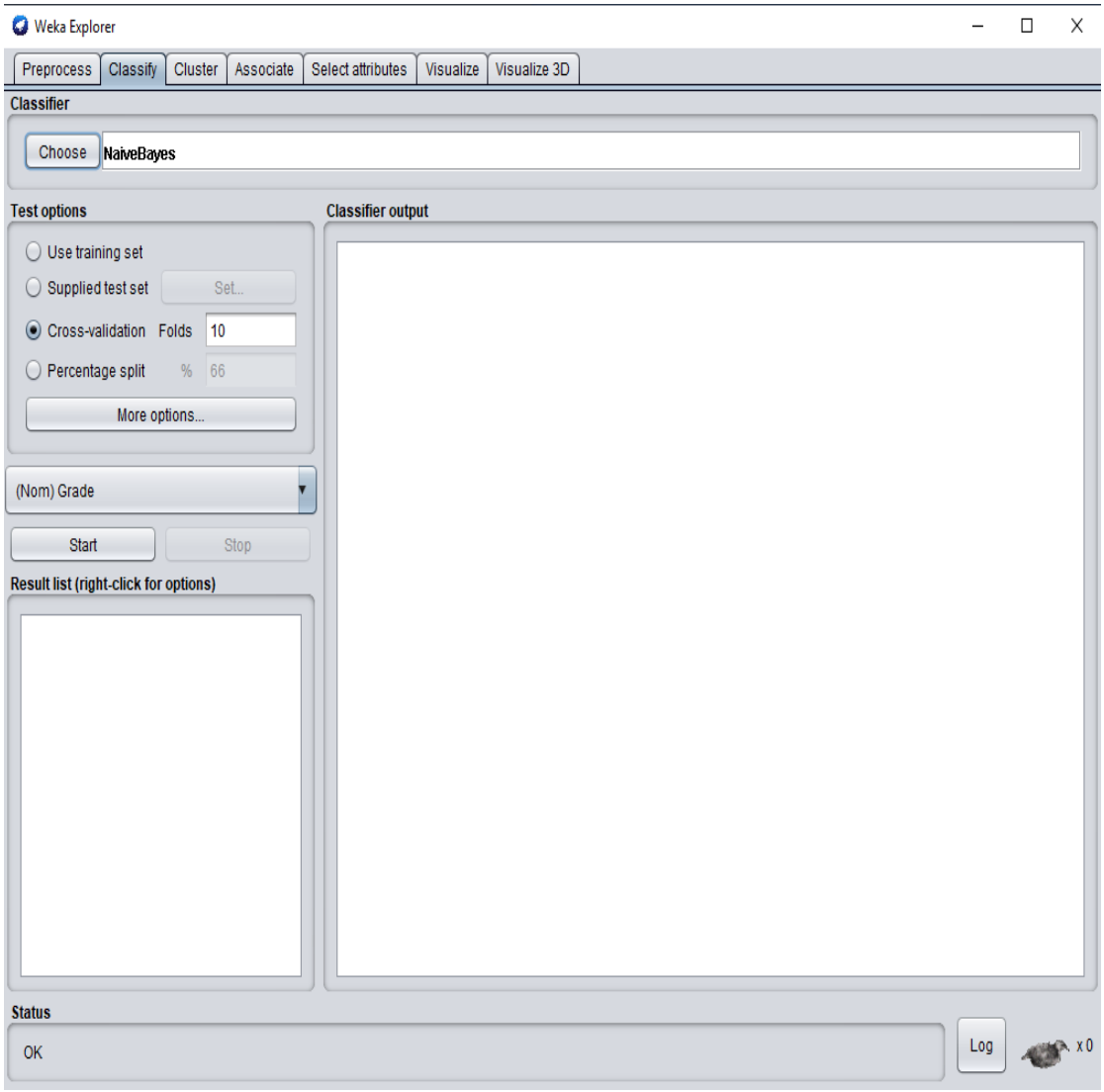

Figure 3.6 Step before the classification process in weka

For classification, we use two algorithms like Naïve Bayes and Decision tree (j48 in weka).We use Naïve Bayes in figure 3.5 and use cross-validation as test options. In cross-validation, all the dataset folds into some number of data. In our case, we fold 10 data are in each fold. After that, a set of the fold acts like a test set and it compares with the other sets and thus we get the accuracy for our model. We check the data set for two algorithms to find out which suits best for our dataset. First, we use Naïve Bayes classifier.For that, we choose the classifier from the classifier menu. After that, we

select the test option and then we select the class attribute. Generally, the class attribute is the last attribute in weka. We can see in figure 3.4. The class attribute is in the last in the attribute column.We also select that from classify step. The option for that is under the test options in weka figure 3.5. Then we click the start button to start the classification process. The result of applying the classifier is illustrated in figure 3.7.

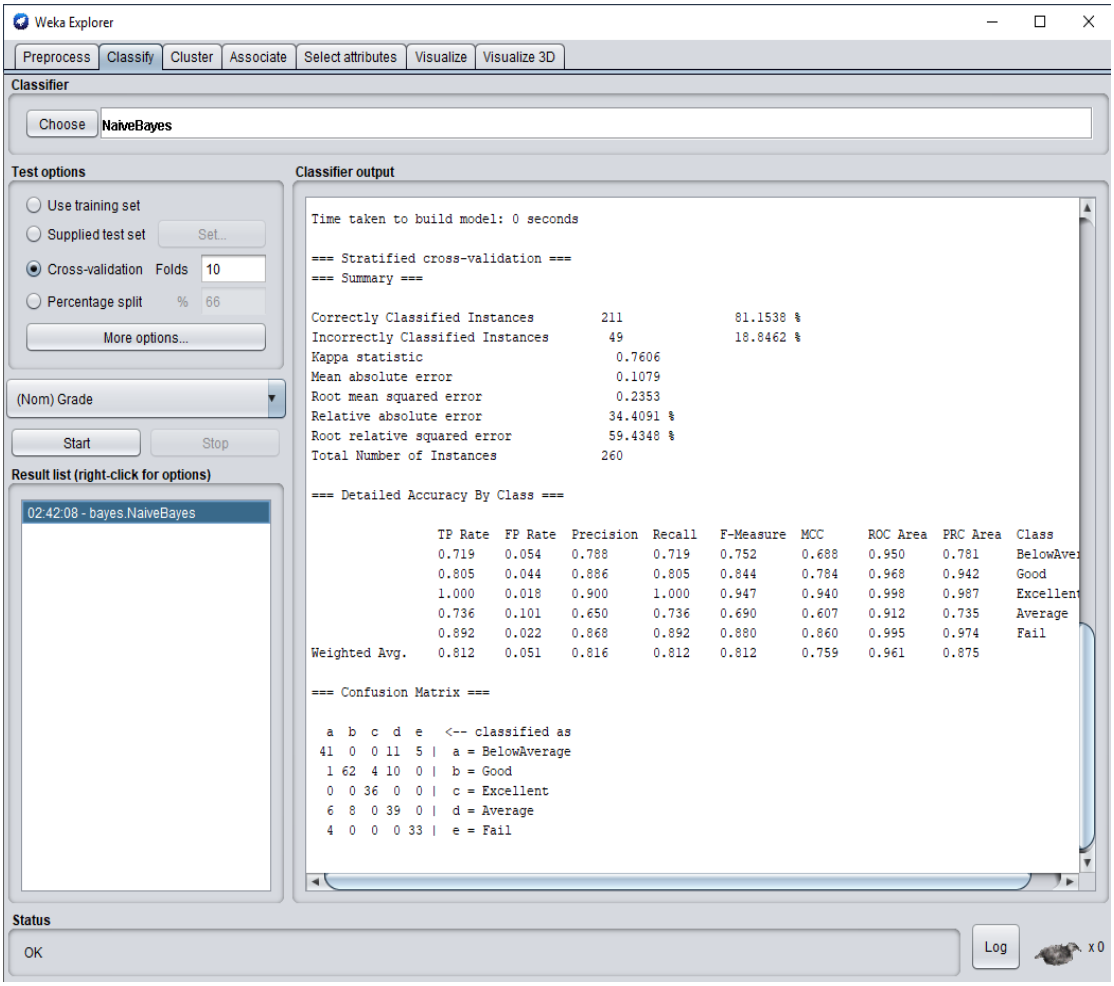

Figure 3.7: The classifier output model using a Naïve Bayes Algorithm in weka

From figure 3.7, we can see that result of using Naïve Bayes classifier. We can see the two vital information from figure 3.7. One of them is correctly classified instance and the other is an incorrectly classified instance. Here in this model, the accuracy is about 81.15% . So, we can say that the model is a smart model for our dataset. Now, some questions arise on our mind that what the other things mean in figure 3.7 like accuracy, confusion matrix, TP Rate, FP Rate, Precision, Recall.

The confusion matrix is a matrix that used to describe or justify classification model whether it is true or not. TP means true positive FP means false positive.FN means false negative and TN means true negative. From TP,FP,FN,TN we get the value of accuracy, Precision and Recall value. For Accuracy first, we add TP and TN and then divided with the total number of the instances. For Precision FP divided by the total number of instances. For recall FN divided with a total number of instances.

We also use Decision tree (called J48) in weka. By using this algorithm, we can generate a visualize tree from the model. First, we want to check the dataset with this algorithm and that showed in figure 3.8.

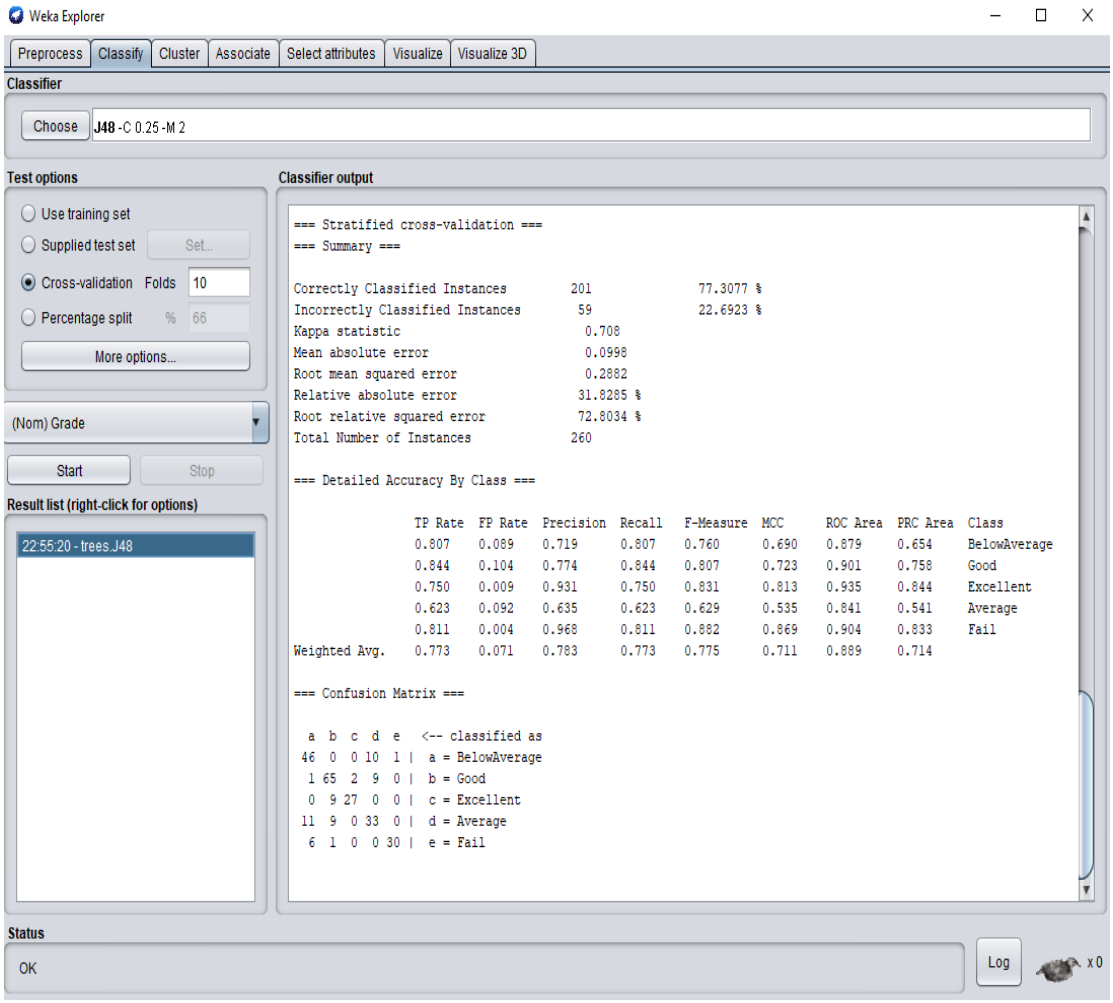

Figure 3.8: The classifier output model using a Decision Tree Algorithm in weka

From figure 3.8 we see that the accuracy of this model is about 77.30 percent. So, we call that this model is also a smart model for our test set. Decision tree accuracy is less

than the Naïve Bayes classification. But the classification we did here we set test option as cross-validation. But after analyzing with training and test set we can say that which one is better. We can see that in the future section in this chapter. Again the difference rate of the accuracy between two model is not so high. We can also visualize a tree view model with a Decision tree. For this, we have to right click in the result list then we have to click on the visualize tree. Then we got the tree view model of the dataset and that is shown in figure 3.9.

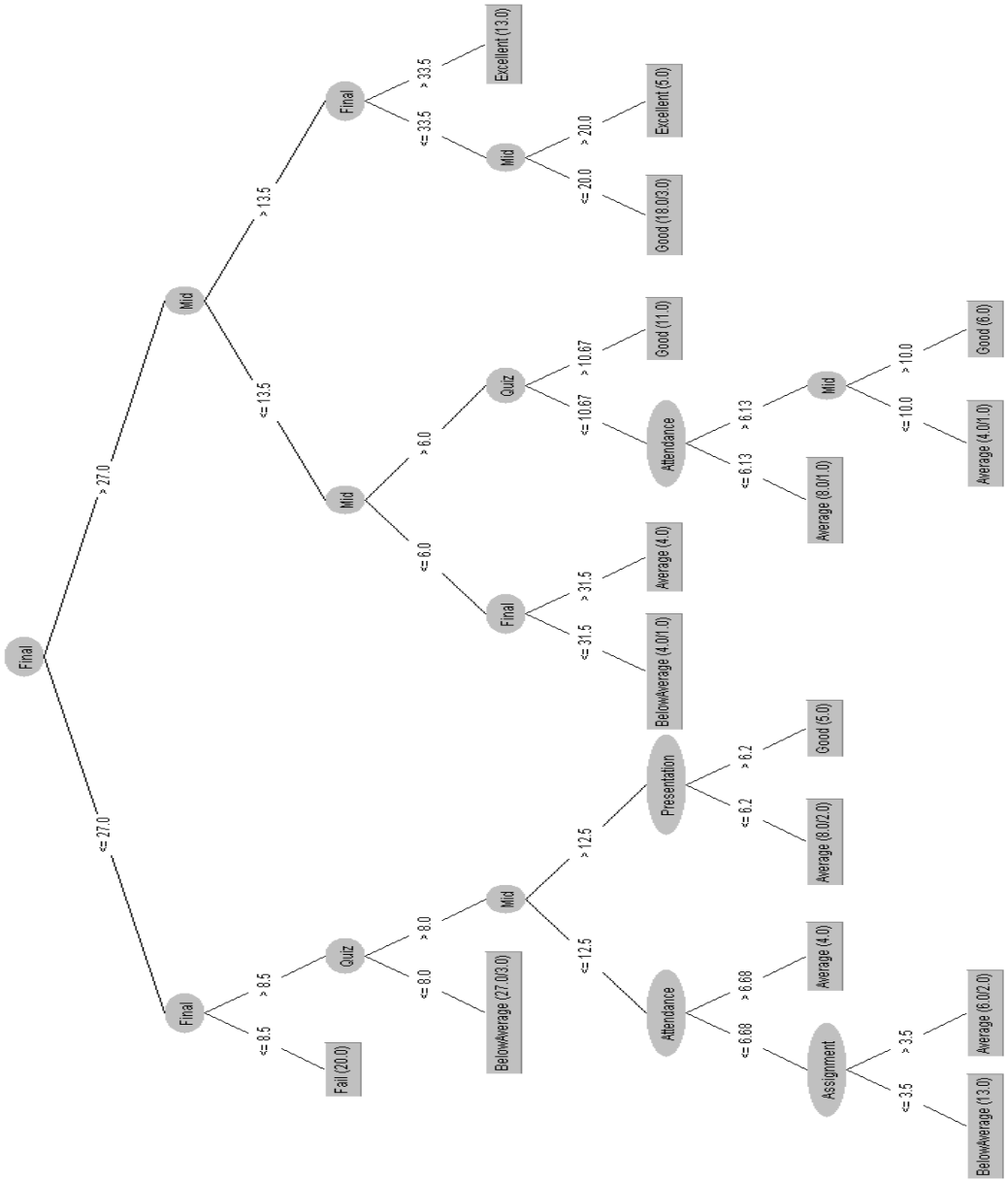

Figure 3.9: Visualize the tree view model in Decision tree in weka

3.4 Experiments using Weka Tool

Experiment area is necessary for done half of the analysis work. Experiment step is tough for every researcher for their research work. In data mining, the result of the research work depends on the what method or style we use for analysis [11].

Experiment

For the experiment our data set in weka. First of all, we divided the dataset into two sets. We select about 60% of our instances for the training set and the rest of the data set we use as a test set. Then we divided the test datasets into two sets as test set-1 and test set-2. The information on the number of instances of each set is shown in table 3.3.

Table 3.3: Provides information about the instances of the dataset

Instances	Number of instances
Total instances	260
Training instances	156
Test Dataset-1 instances	52
Test Dataset-2 instances	52

Figure 3.10: Information about the number of instances

For our experiment, first of all, we open our training dataset by clicking the open file in the weka software from the pre-process. Then select classify option to apply the algorithm.

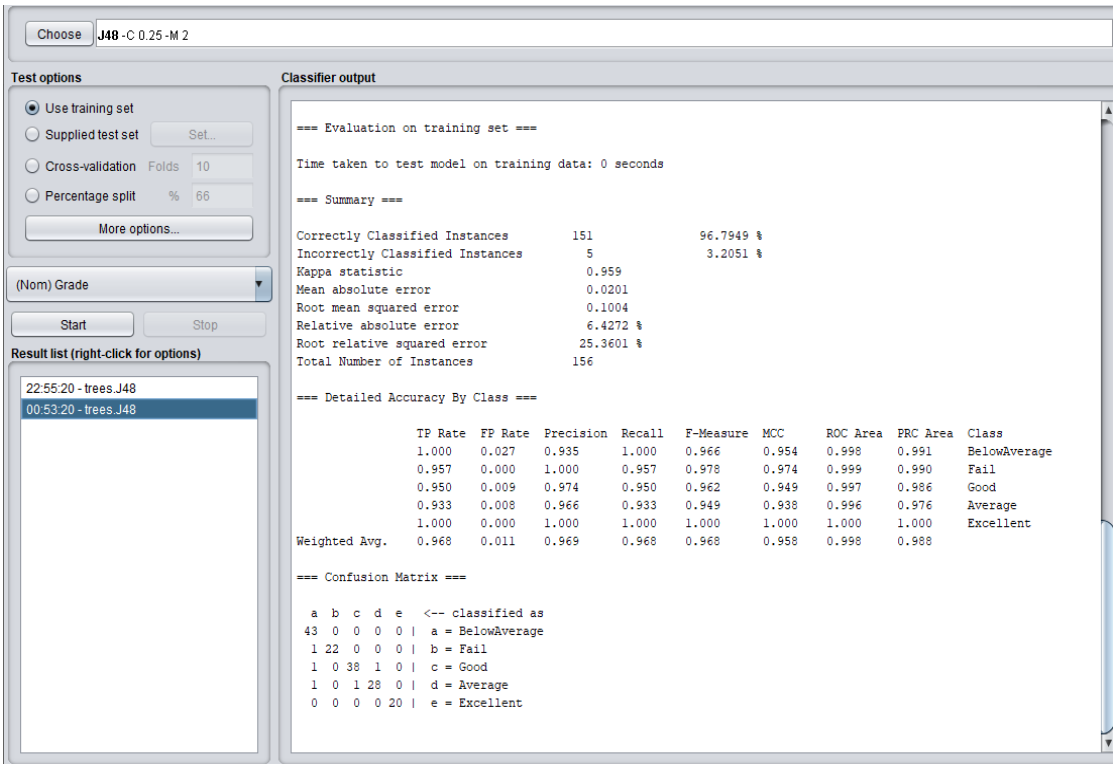

Figure 3.11: Run Results on Training dataset using Decision Tree (j48)

Here,we first use Decision tree named j48 in weka. Then select the text options as use training set. After that, we click the start button to run the process.The results we get are shown above in figure 3.11.

From the test, we have to get the result like-

- Correctly Classified instances 151 96.7949%
- Incorrectly Classified instances 5 3.2051%

We get the result from the training data sets now we run our Test dataset-1 with the training dataset to see how our experiment goes. So, we have click on the test options named use supplied test then click start to run the Test Data-set 1.

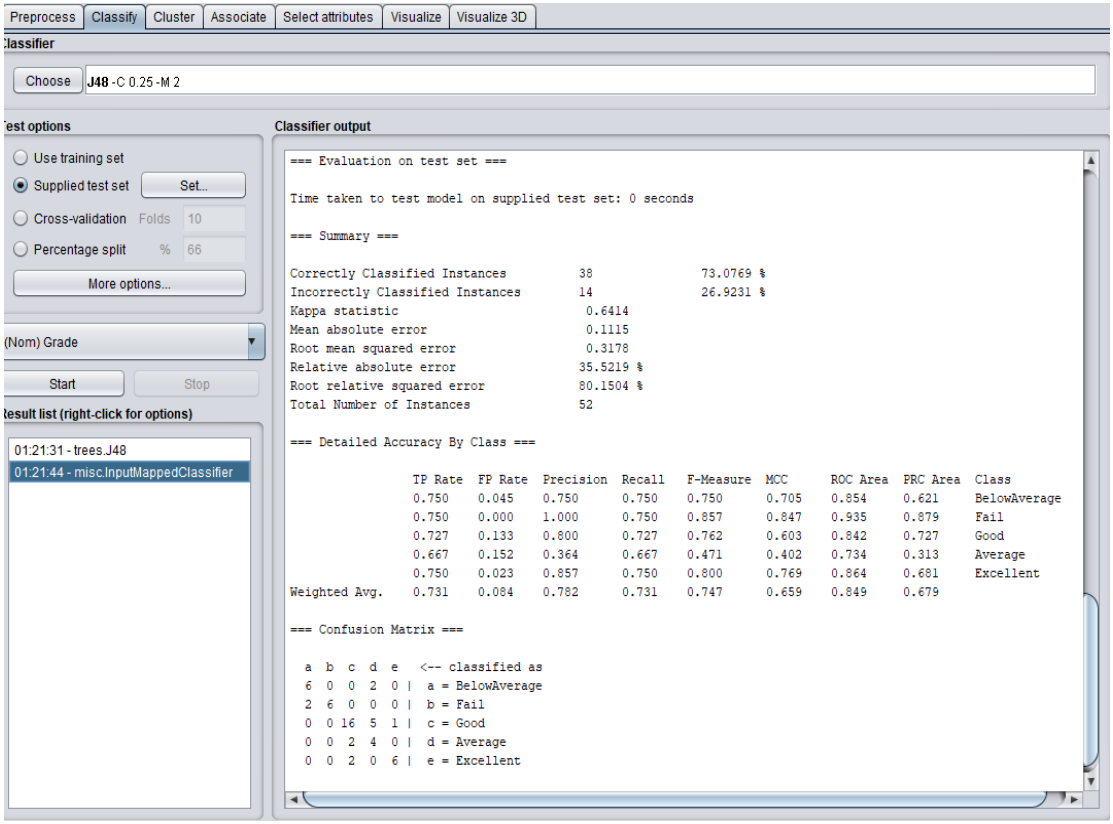

Figure 3.12: Run Results on Test dataset-1 using Decision Tree (j48)

From the run result on Test dataset-1, we get the result like-

- Correctly Classified instances 38 73.0769%
- Incorrectly Classified instances 14 26.9231%

Now we can run the Test dataset-2 as the supplied test with the training datasets and check the run result on the Test dataset-2. The run result is shown below in figure 3.13.

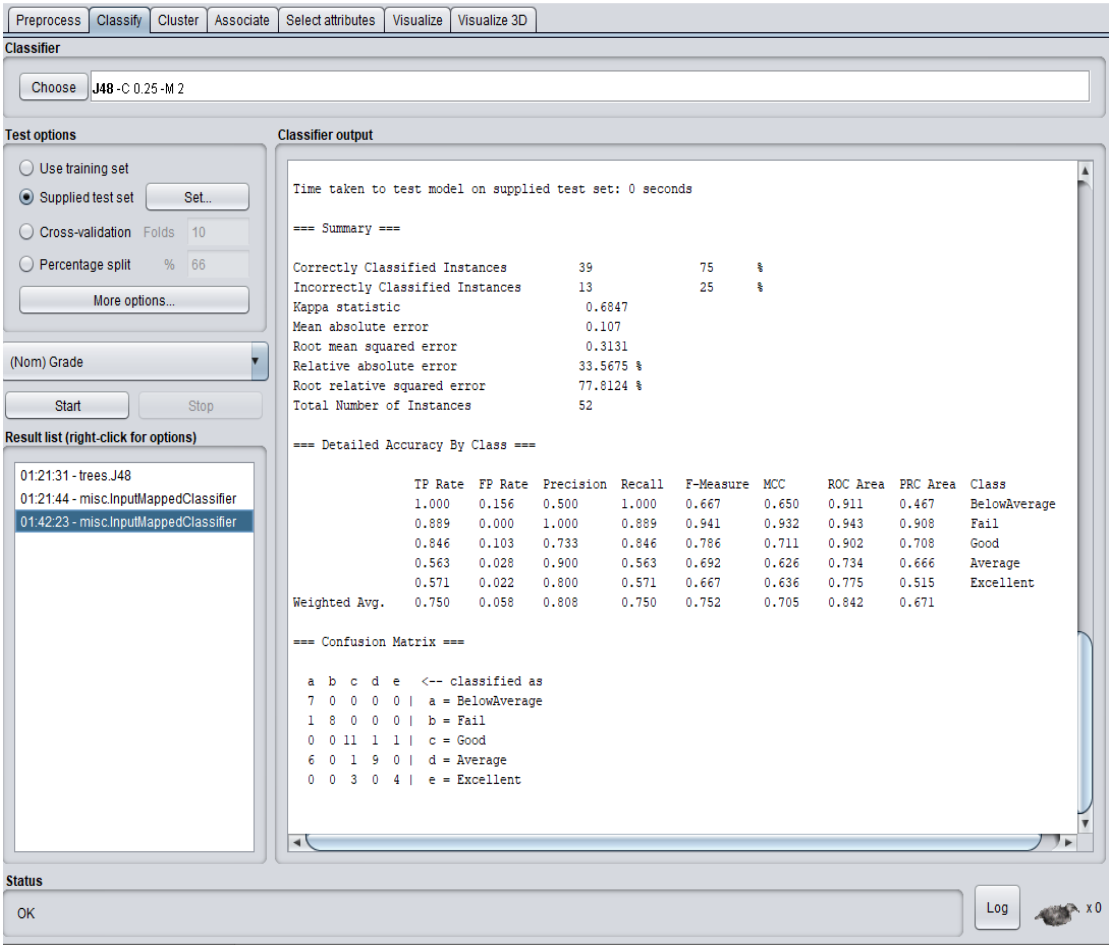

Figure 3.13: Run Results on Test dataset-2 using Decision Tree (j48)

From the run result on Test dataset-2, we get the result like-

- Correctly Classified instances 39 75%
- Incorrectly Classified instances 13 25%

From table 3.4, we can see the information that we get from the above experiments using Decision Tree (j48)

Table 3.4: Provides information about the run result using the training and testing datasets using Decision tree (j48)

Types of the dataset	Number of correctly classified instances	Number of incorrectly classified instances	Percentage of correctly classified instances	Percentage of incorrectly classified instances
Training set	151	5	96.7949%	3.2051%
Test set-1	38	14	73.0769%	26.9231%
Test set-2	39	13	75%	25%

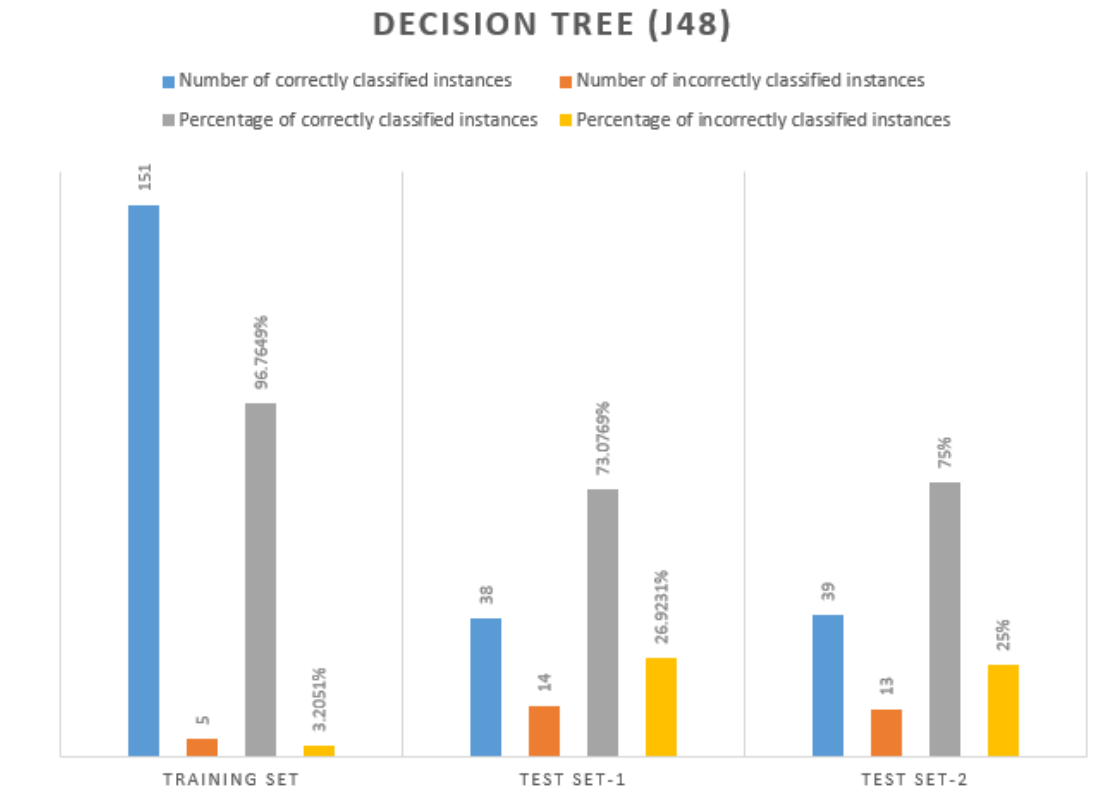

Figure 3.14: Information about the run results on different test sets using a Decision tree

Now we use the Naïve Bayes classification algorithm. First, we apply the classifier on training data set in weka. Then use the test dataset-1 and test dataset-2 to check the accuracy of the dataset with the training dataset.

At first, we choose the classifier and we choose Naïve Bayes for our experiment. After running the process we get results that are illustrated in figure 3.15

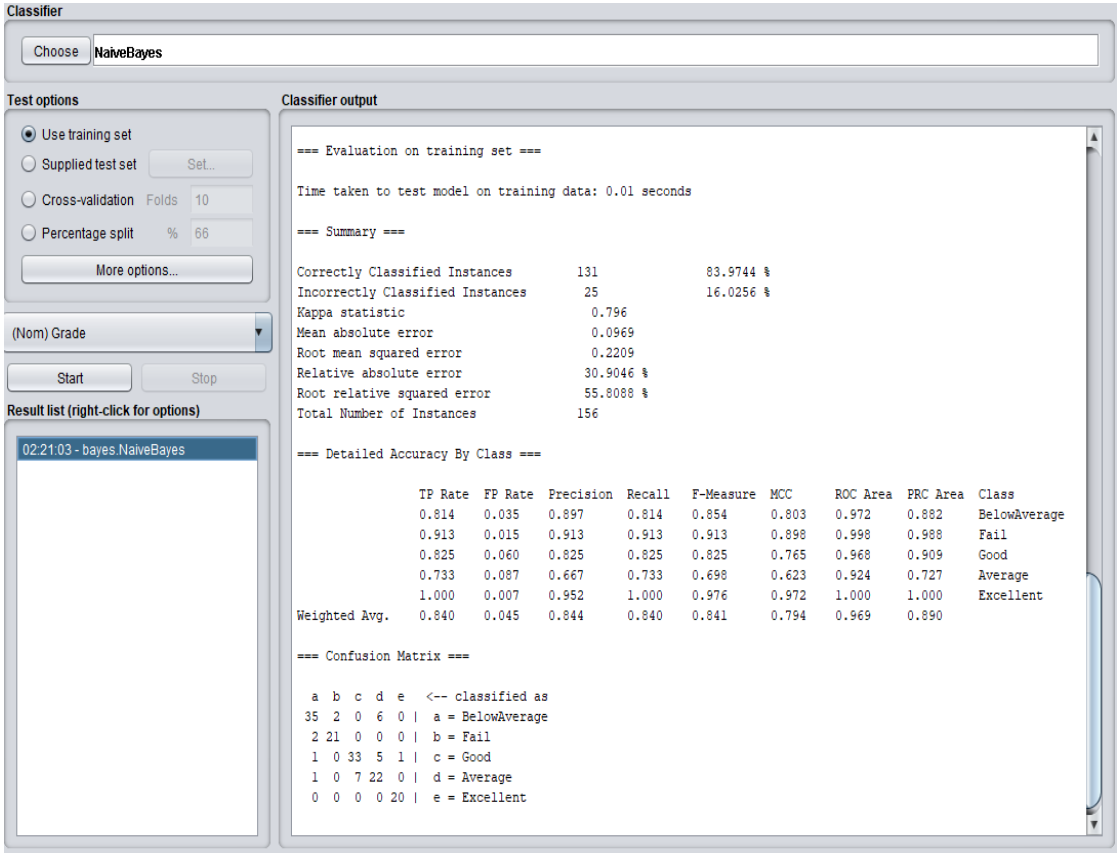

Figure 3.15: Run Results on Training dataset using Naïve Bayes

From the test, we get the result like this-

- Correctly Classified instances 131 83.9744%
- Incorrectly Classified instances 25 16.0256%

Now we load the test dataset-1 in weka software as the supplied test. After running the process we get the results that are shown below in figure 3.16.

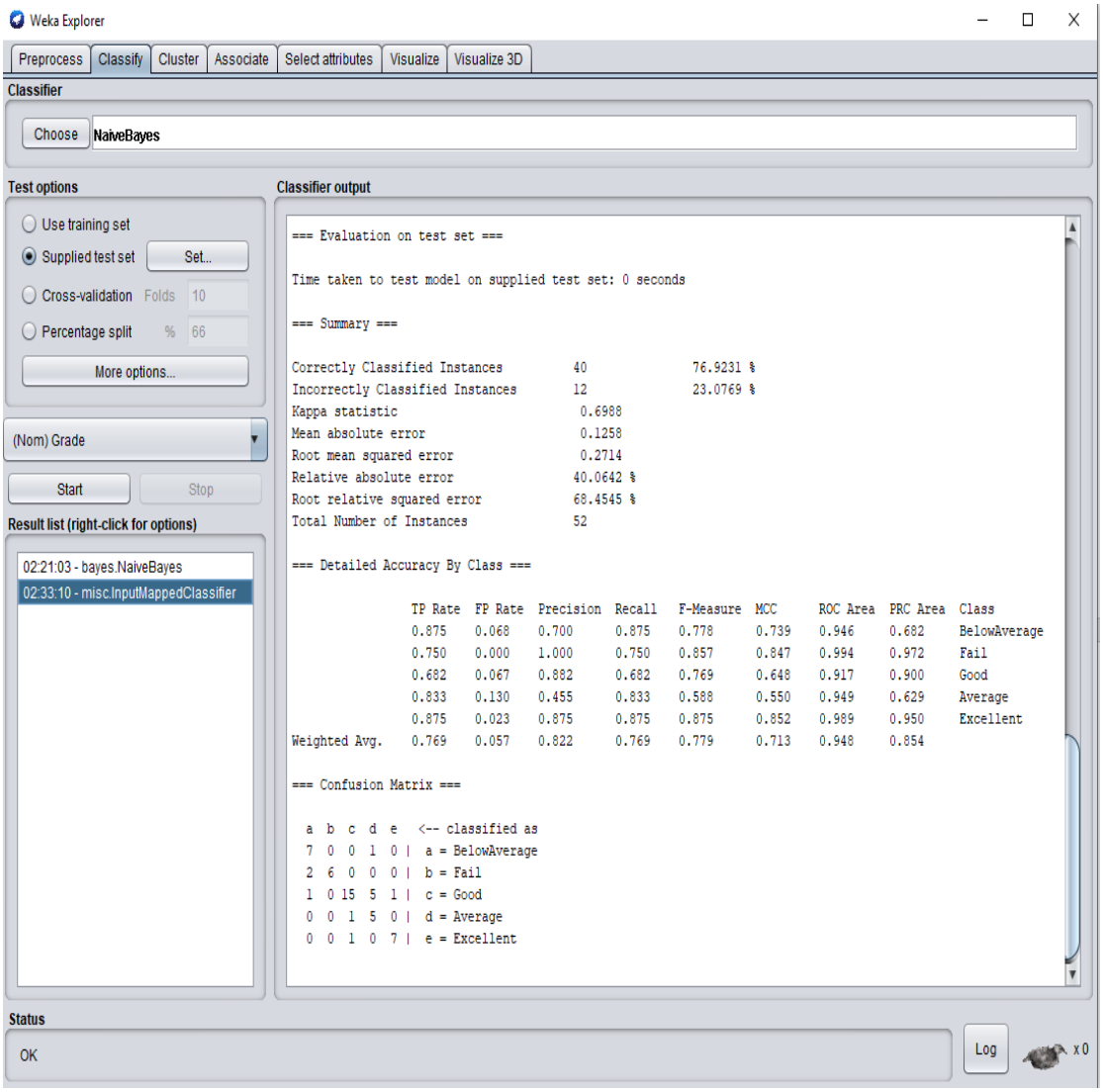

Figure 3.16: Run Results on the Test dataset-1 using Naïve Bayes

From the test, we get the result like this-

- Correctly Classified instances 40 76.9231%
- Incorrectly Classified instances 12 23.0769%

Now we load the test dataset-2 in weka software as the supplied test. After running the process we get the results that are shown below in figure 3.17.

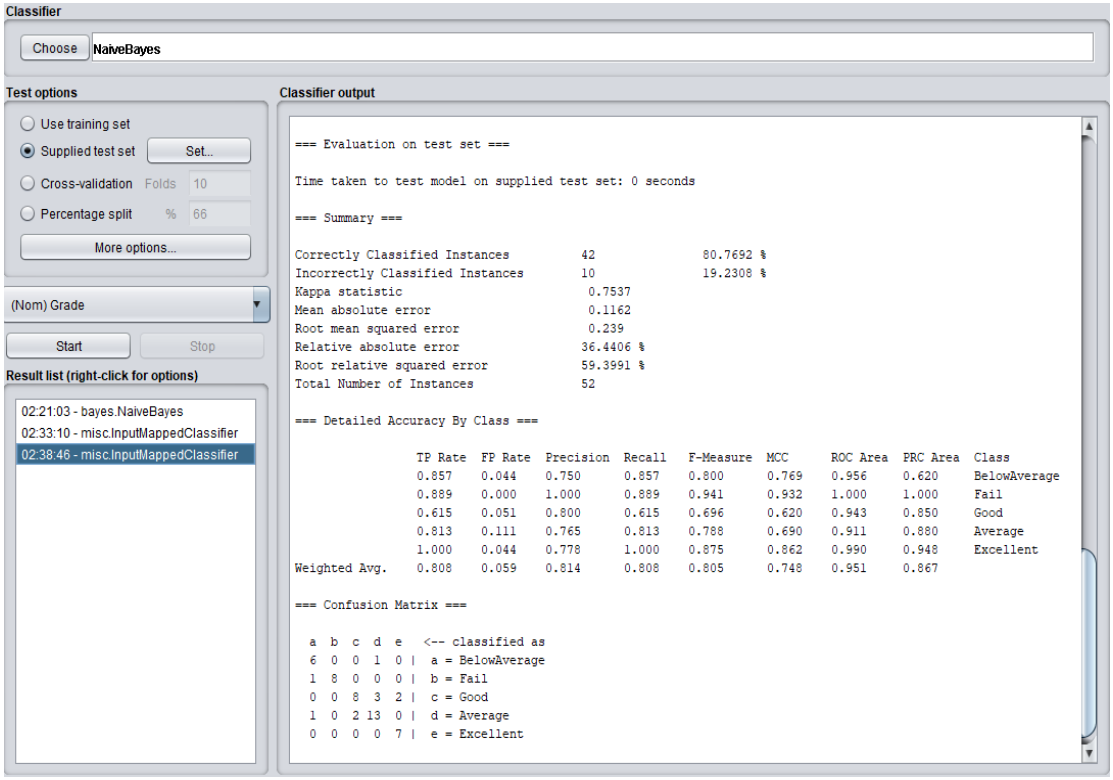

Figure 3.17: Run Results on the Test dataset-2 using Naïve Bayes

From the test, we get the result like this-

- Correctly Classified instances 42 80.7692%
- Incorrectly Classified instances 10 19.2308%

From table 3.5, we can see the information that we get from the above experiments using Naïve Bayes

Table 3.5: Provides information about the run result using the training and testing datasets using Naïve Bayes

Types of the dataset	Number of correctly classified instances	Number of incorrectly classified instances	Percentage of correctly classified instances	Percentage of incorrectly classified instances
Training set	131	25	83.9744%	16.0256%
Test set-1	40	12	76.9231%	23.0769%
Test set-2	42	10	80.7692%	19.2308%

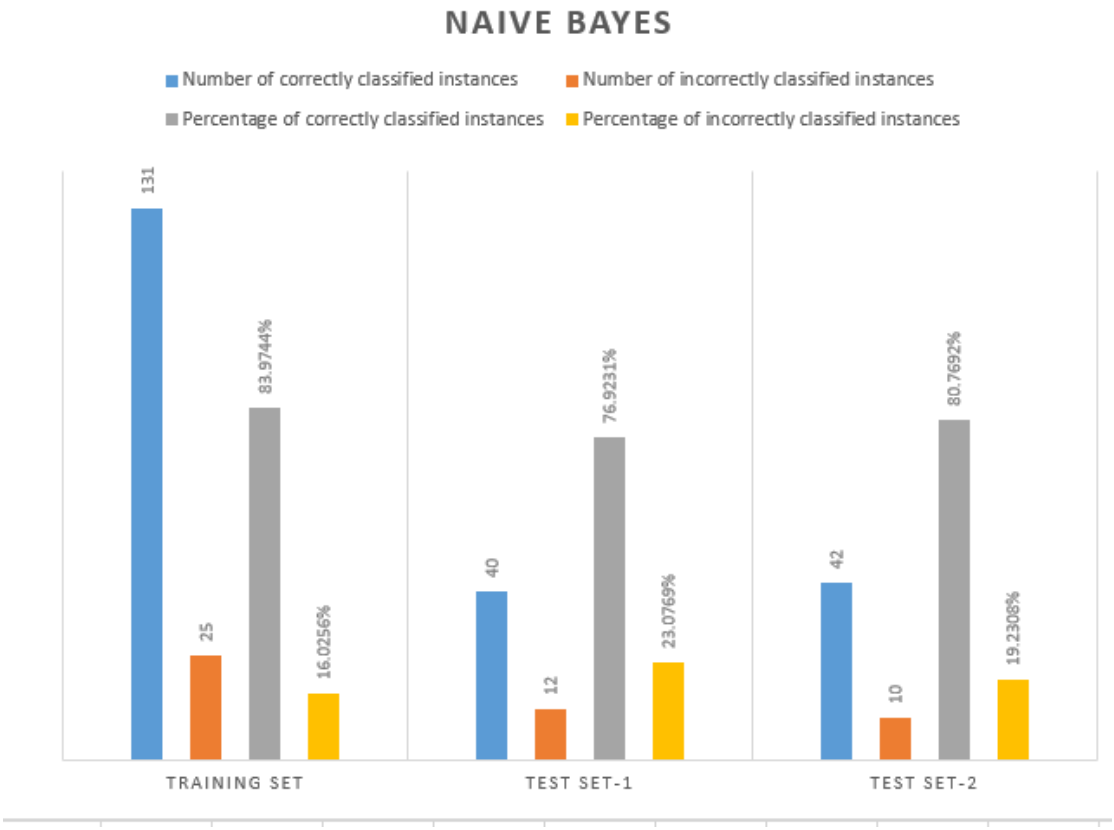

Figure 3.18: Information about the run results on different test sets using Naïve Bayes

Thus we complete training, testing process for our data sets. And the two algorithms build the smart models for our datasets.

3.5 Challenges

We face huge challenges in this section such as data collection is a vital part of our research work. We not only face problem in data collection but also face problem in data fitting, data filtering. Now we mention the challenges that we face in the section are given below-

- We have to verify the kinds of data that we want to use.
- We have to find out the attribute that we need for research.
- We have to rearrange the data within the tabular format.
- We have to design a proper structure for our analysis.
- We have to sort the datasets for our analysis.
- We have to learn about different options to work in weka.
- We have to learn a different classifier that uses in weka.
- We have to learn about Decision Tree and Naïve Bayes algorithms.
- We have to find out how to analyze the result that we find out in weka.

CHAPTER 4

Analysis of Results and Discussion

4.1 Result Analysis

Total number of the instance: 260

Number of Attributes: 7 classes

For each attribute: (numeric and nominal valued)

1. Attendance (numeric value)
2. Presentation (numeric value)
3. Assignment (numeric value)
4. Quiz (numeric value)
5. Mid (numeric value)
6. Final (numeric value)
7. Grade (nominal value)

Class Attribute: Grade

Missing Attribute values: None

Table: 4.1: Class value with the number of instances

Class Value	Number of instances
Below Average	57
Good	77
Excellent	36
Average	53
Fail	37

From table 4.1, we can see the class values and their number of instances of total datasets that we use for this research work.

Table 4.2: Provide information about the statistical analysis of all numeric attributes

Attribute Name	Minimum value	Maximum value	Mean value	Standard Deviation	Distinct value	Unique value
Attendance	0	7	5.362	1.787	66	29
Presentation	0	7.4	5.221	2.308	24	1
Assignment	0	5	3.327	1.615	17	5
Quiz	0	14.67	8.325	3.787	65	16
Mid	0	24.5	11.596	5.293	43	2
Final	0	40	24.333	10.476	54	13

From table 4.2, we can see the different types of value for each attribute. Such as presentation attribute. In presentation attribute, the minimum number a student get is 0, maximum number a student get is 7.4, average number a student get is 5.22, the standard deviation of the number is 2.308, in presentation, we 24 distinct value and 1 unique value. From distinct, it indicates that there are 24 values in the presentation attribute.

4.2 Comparative Study Analysis using Different Technique

We try to find out different types of comparison in this section

Comparison Between Before and after training of the dataset

We can see various kinds of accuracy throughout the experiment in chapter 3. Now, in this section, we try to compare the accuracy of each experiment and analysis the result. First of all, we chose the test option as the cross-validation and we run the whole datasets with that. We use two algorithms one of them is Decision Tree and another is Naïve Bayes. We also divided the dataset into some parts and run our algorithm. In this section, we can compare with different results.

We try to find out the difference of accuracy using cross-validation and using different sets as training and test set with the classifier as Decision Tree and Naïve Bayes.

From figure 4.1, we see the accuracy using different data sets with cross-validation

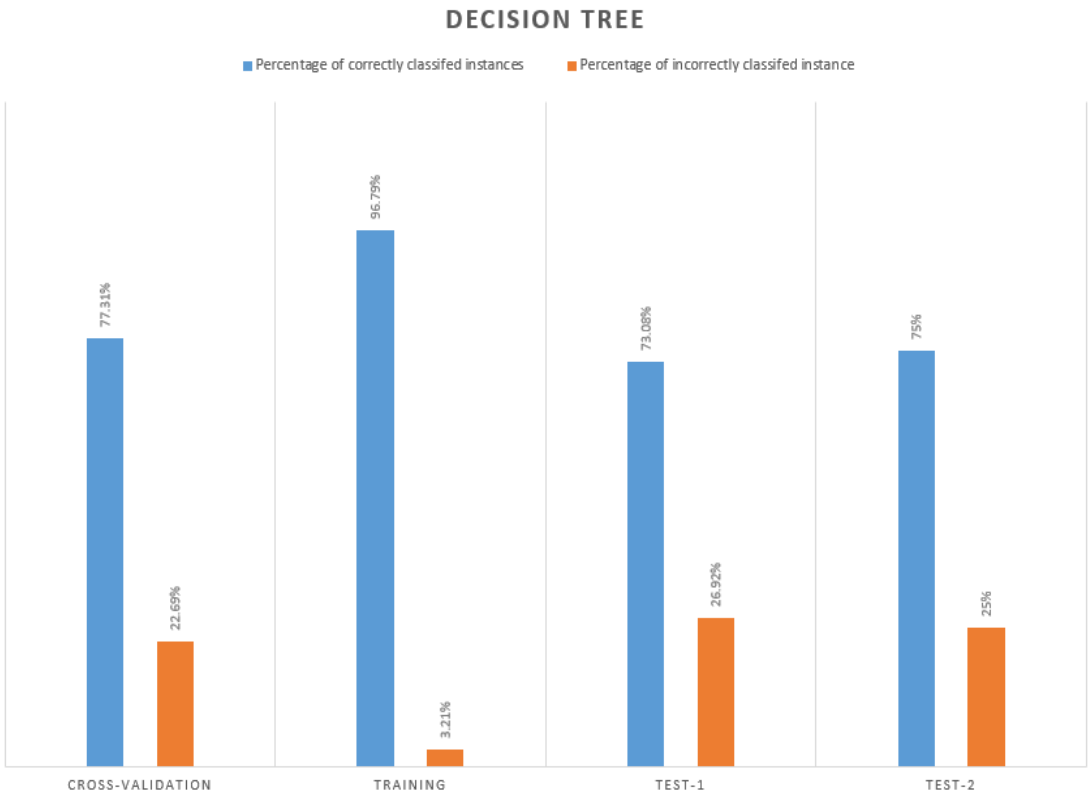

Figure 4.1 Comparison between different data sets with Decision Tree

We can see the difference relative to the accuracy of the dataset. Here we get better accuracy in the training dataset. Before training dataset in cross-validation the accuracy of the correctly classified instance was 77.3% and after the training set, the accuracy is 96.79%. We also created two datasets and check the accuracy of the dataset with the training dataset.

From figure 4.2, use see the accuracy using different data sets with cross-validation

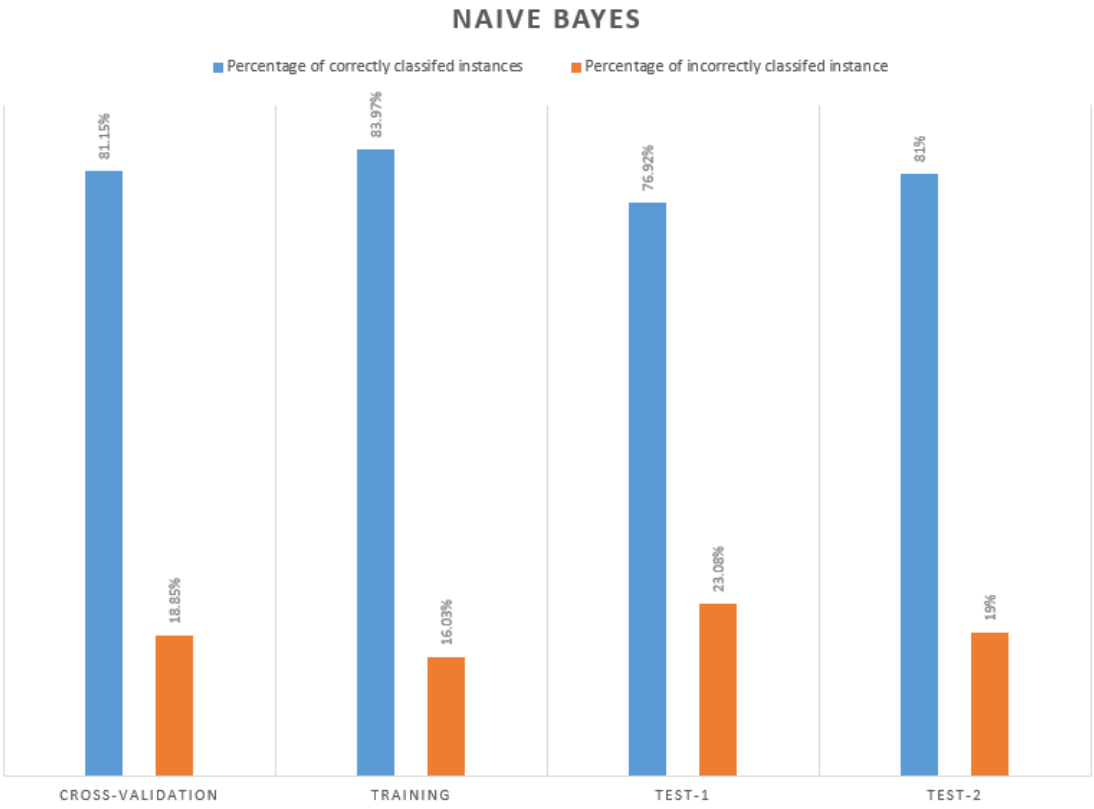

Figure 4.2: Comparison between different data sets with Naïve Bayes

We can see the difference relative to the accuracy of the dataset. Here we get better accuracy in the training dataset. Before training dataset in cross-validation the accuracy of the correctly classified instance was 81.15% and after the training set, the accuracy is 83.97%. We also created two datasets and check the accuracy of the dataset with the training dataset.

Comparison of the different datasets accuracy between two classifier

Here we try to compare the difference between the two classifiers. We use cross-validation as the test options.

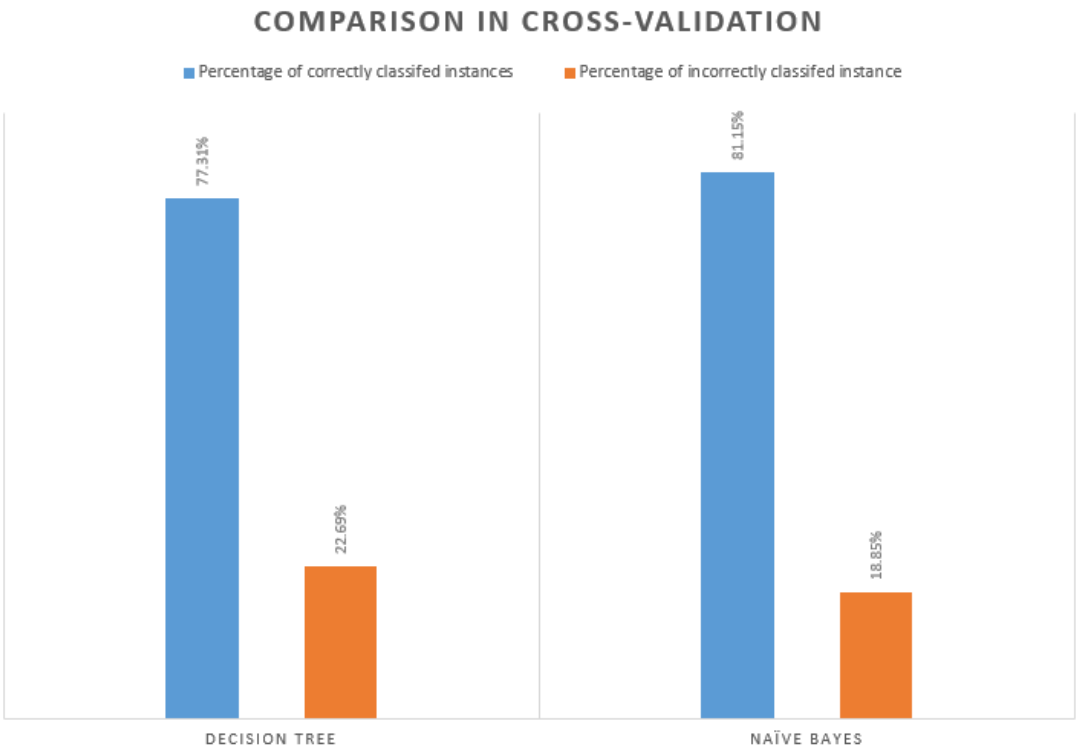

Figure 4.3: Comparison in cross-validation

From figure 4.3, we can see that the accuracy of the correctly classified instances of Naïve Bayes is higher than the Decision Tree. From this comparison, we can say that the Naïve Bayes model is slightly better than Decision Tree in this situation.

Now we try to compare the two classifiers on the basis of their training set

Figure 4.4: Comparison of the training set

From figure 4.4, we can see that the accuracy of the correctly classified instances of Decision Tree is higher than the Naïve Bayes. From this comparison, we can say that the Decision Tree model is absolutely better than Naïve Bayes in this situation.

Now we try to compare the two classifiers on the basis of their test set-1

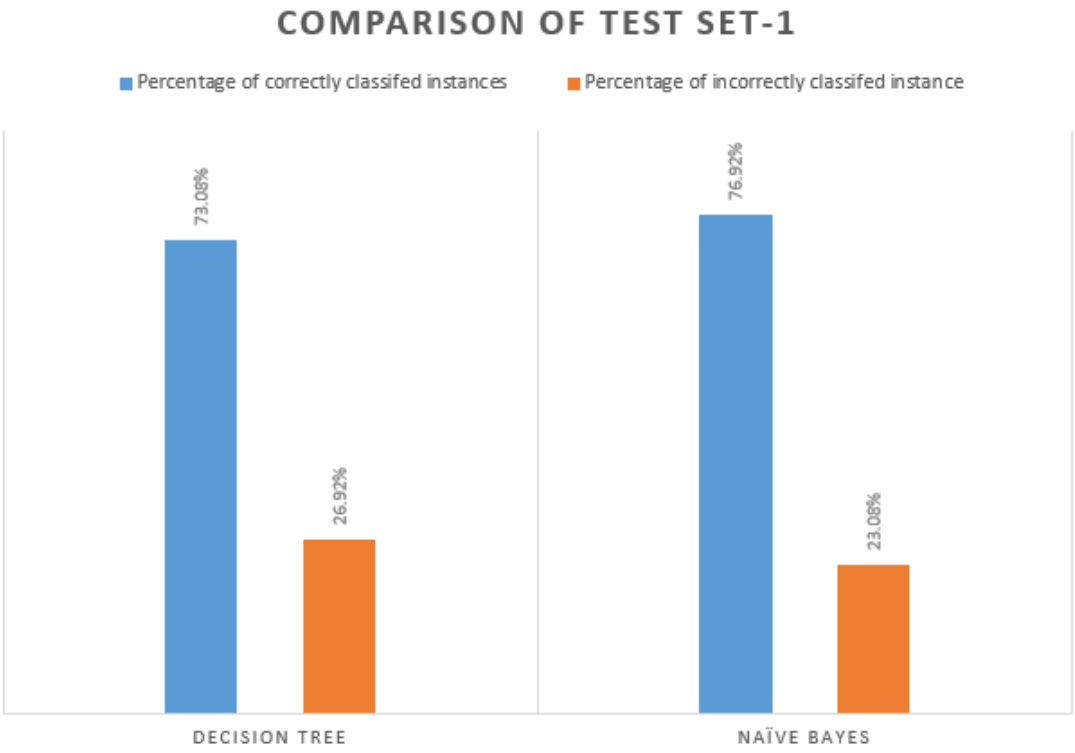

Figure 4.5: Comparison of test set-1

From figure 4.5, we can see that the accuracy of the correctly classified instances of Naïve Bayes is higher than the Decision Tree. From this comparison, we can say that the Naïve Bayes model is slightly better than Decision Tree in this situation.

Now we try to compare the two classifiers on the basis of their test set-2

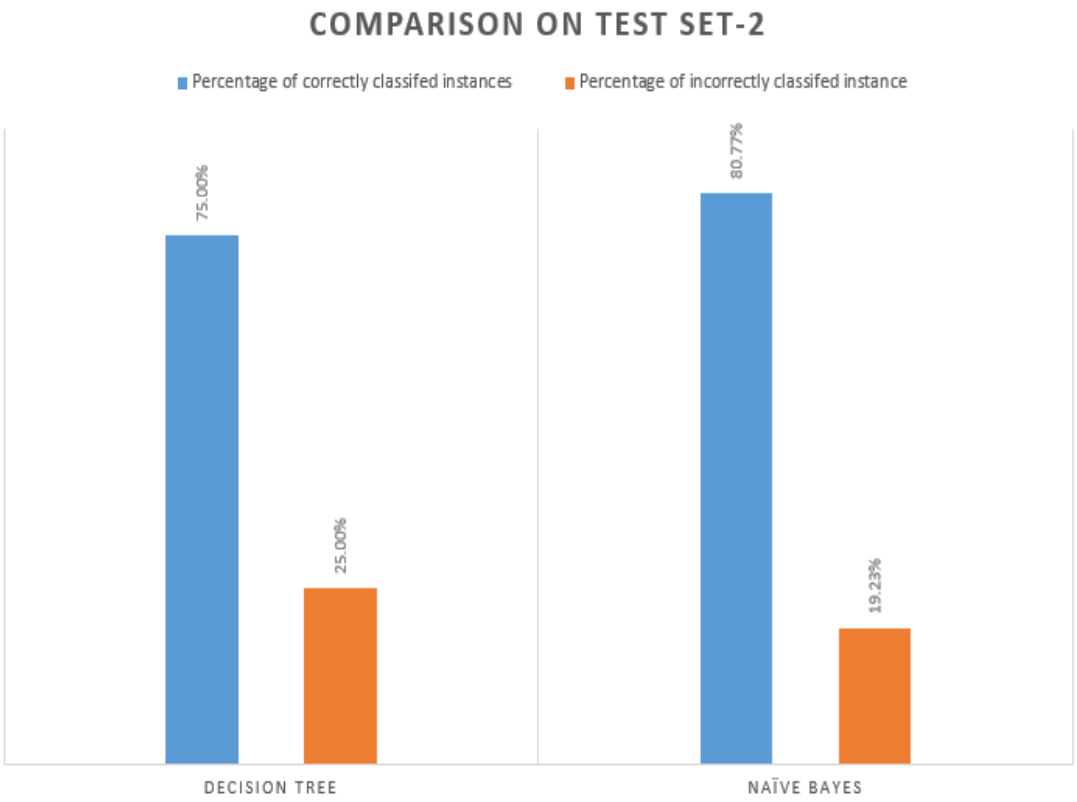

Figure 4.6: Comparison of test set-2

From figure 4.6, we can see that the accuracy of the correctly classified instances of Naïve Bayes is higher than the Decision Tree. From this comparison, we can say that the Naïve Bayes model is slightly better than Decision Tree in this situation.

We can visualize a tree using Decision tree and for our research, we need a tree view model. So, we created our system using the tree view model. It's also can be called the final result of our research. Because we try to find out a knowledge pattern from our datasets.

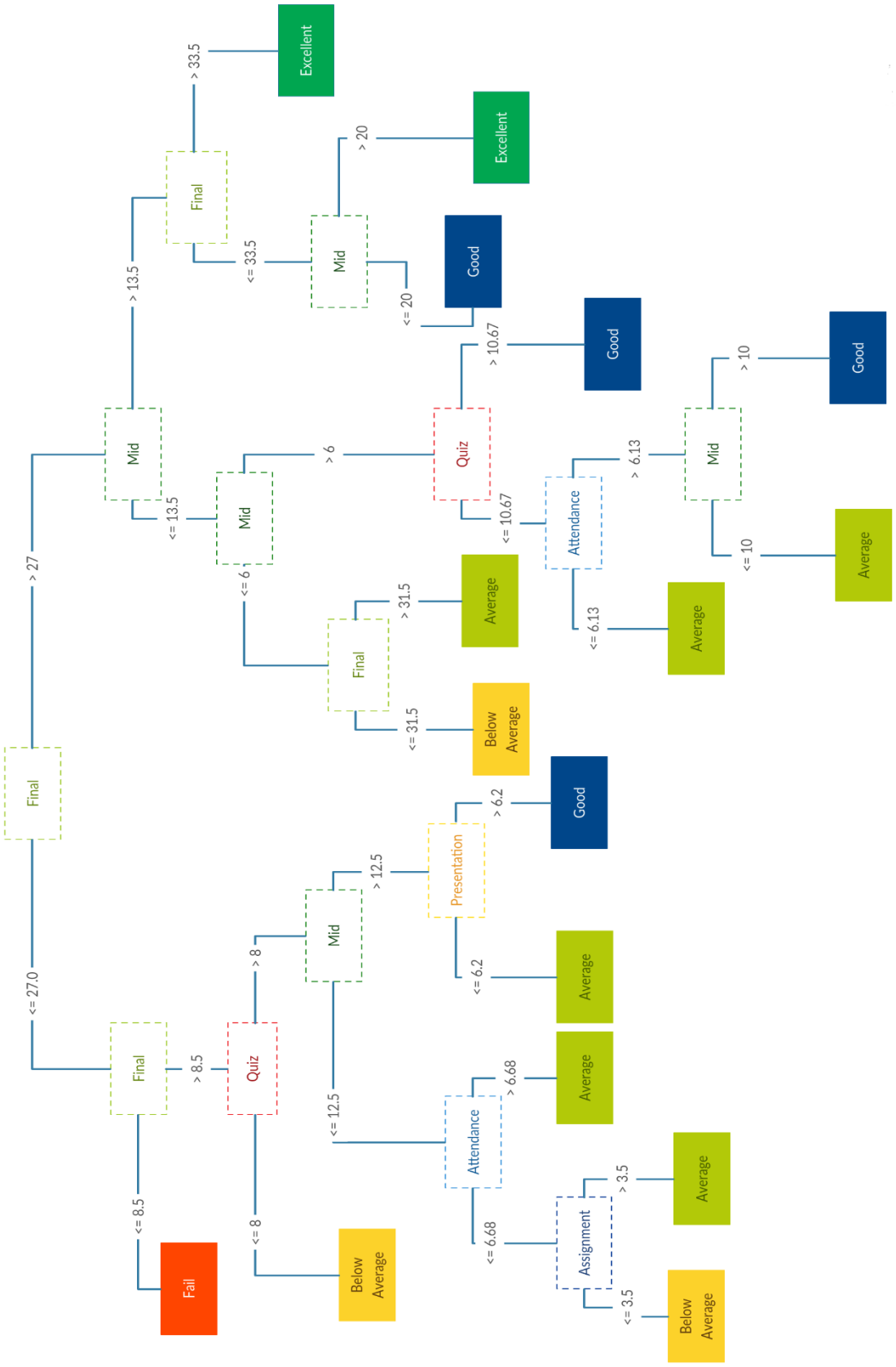

Figure 4.7: Visualize the tree view model using Decision Tree

From figure 4.7, we see the useable knowledge from our data sets. Now we can use this knowledge and implement the knowledge to Student Result Prediction System.

4.3 Discussion on Findings

Finally, we completed our analysis section. In this section, we try to compare the different types of experiment that we did in the previous chapter. Now, we note down the findings that we find throughout the chapter are given below

- Class value of each attribute.
- Statistical information about the numeric attribute such as maximum, minimum, mean value etc.
- Observe comparison of Decision Tree and Naïve Bayes on cross-validation test option
- Observe comparison of Decision Tree and Naïve Bayes with the training dataset and get the higher value of accuracy for Decision Tree
- Visualize the tree view model using Decision Tree
- Find out a ruleset that can be implemented in any education system
- Find out the way of implementation of this research knowledge with the tree view model

4.4 Challenges

The analysis part is the most critical part of our research-based project. We face many challenges during this chapter. Let's discuss the challenges that we face is this particular chapter.

- Understand the difference between the numeric and the nominal value
- Select the classification algorithm that suits for the analysis
- Used to with the general operation of the weka
- Understand the comparison manner to describe the accuracy of the classifier
- Decide to use what type of classifier we need for implementing the system
- Evaluating the result of each experiment

CHAPTER 5
Conclusion and the Future Scope

5.1 Conclusion

Student result is important for any student. Our intention is to provide a system that can help student for their educational purpose. Now, with the help of our system, a student can predict his result before the final exam on a particular. So, this system can be helpful for the student if they are worried about their result. All the result of the student before final is saved in the student database by the course teacher. In our system, we can predict the student result on a particular subject without a final examination. If a student can have the idea what number, he can get in the final exam then he can check his predicted result. So, our research is dedicated to improving the student result.

5.2 Future Scope

This research can be implemented and further developed with the collaboration of any educational institution. Because the student data is confidential. This research can be further extending because with the lots of data. The prediction result can be more accurate and realistic if the model is trained with a huge amount of data. This research can be used to predict the subject wise result of any institution.

REFERENCES

[1] Romero, C. & Ventura, S. (2007),Educational Data Mining: a Survey from 1995 to 2005, Expert Systems with Applications, Elsevier, pp. 135-146.

[2] Romero, C. , Ventura, S. and Garcia, E. (2008) 'Data mining in course management systems: Moodle case study and tutorial', Computers & Education, vol. 51, no. 1, pp. 368-384.

[3] F. Grivokostopoulou, I. Perikos, and I. Hatzilygeroudis, "Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance," *Proc. IEEE Int. Conf. Teaching, Assess. Learn. Eng. Learn. Futur. Now, TALE 2014*, no. December, pp. 488–494, 2015.

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

[5] A. Ktona, D. Xhaja, and I. Ninka, "Extracting Relationships between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques," *2014 Sixth Int. Conf. Comput. Intell. Commun. Syst. Networks*, pp. 6–11, 2014.

[6] T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," *Int. Conf. Adv. Comput. Commun. Technol. ACCT*, pp. 255–262, 2014.

[7] J. Jacob, K. Jha, P. Kotak, and S. Puthran, "Educational Data Mining techniques and their applications," *2015 Int. Conf. Green Comput. Internet Things, ICGCIoT 2015*, pp. 1344–1348, 2016

[8] Varun Kumar, Anupama Chadha, Mining Association Rules in Student's Assessment Data, 2012.

[9] << https://pubs.rsc.org/en/content/articlehtml/2016/rp/c5rp00144g >> 06 Oct 2018, accessed at 10:15 am.

[10] << https://www.quora.com/What-is-classification-in-data-mining >> 14 Oct 2018 accessed at 4.45 pm.

[11] Markov, Z., & Russell, I. (2006). An introduction to the WEKA data mining system. ACM SIGCSE Bulletin, 38(3), 367–368.

.