

**PREDICTING STUDENTS' FUTURE LEARNING
BEHAVIOUR**

USING EDUCATIONAL DATA MINING

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

SALONI

11207797

Supervisor

AMRITPAL SINGH



School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May 2017

ABSTRACT

In this era of globalization and technological revolution, education is considered as a first step for every human activity. It plays a vital role in the development of human capital and is linked to an individual's well-being and opportunities for better living. It ensures the acquisition of knowledge and skills that enable individuals to increase their productivity and improve their quality of life. But there are number of factors which deviate student from the actual path resulting in degradation of academic performance. A series of variables is to be considered when to identify the affecting factors towards quality of academic success. Identifying the most contributing variables in the quality of academic performance is a very complex and challenging job. In this proposed work, a fuzzy rule based expert system has been designed to evaluate the performance of students and dropout rate on the basis of various parameters. Firstly, a discussion is held with highly experienced authorities of Lovely Professional University (LPU) with the reason to identify the critical factors that may be affecting students' performance and dropout rate.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled ‘PREDICTING STUDENTS’ FUTURE LEARNING BEHAVIOUR USING EDUCATIONAL DATAMINING’ in partial fulfillment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Amritpal Singh. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Saloni

11207797

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.TECH Dissertation entitled **“PREDICTING STUDENT FUTURE LEARNING BEHAVIOUR USING EDUCATIONAL DATA MINING”**, submitted by Saloni at **Lovely Professional University; Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Amritpal Singh)

Date: 29/04/2017

Counter Signed by:

1) Concerned HOD:

HoD's Signature: _____

HoD Name: _____

Date: _____

2) Neutral Examiners:

External Examiner

Signature: _____

Name: _____

Affiliation: _____

Date: _____

Internal Examiner

Signature: _____

Name: _____

Date: _____

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this dissertation. I am thankful for their aspiring guidance, invaluable constructive criticism and friendly advice during the thesis work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to this research.

I express my warm thanks to Assistant professor Amritpal Singh for her support and guidance.

I would also like to thank all the people who provided me with the facilities being required and conducive conceptions.

I would like to thank Lovely professional University for providing a grant for this research. I would also like to thank Amritpal Singh, COS Amandeep Nagpal and department of examination Navdeep Dhaliwal and Chavi Ralhan for giving a dataset and participating in my discussion as well as encouraging me for doing this research and helping me to identify the critical factors that may be affecting students' performance and dropout rate.

Thank you,

Saloni

TABLE OF CONTENTS

CONTENTS	PAGE NO.
Inner first page – Same as cover	i
PAC form	ii
Abstract	iii
Declaration by the Scholar	iv
Supervisor’s Certificate	v
Acknowledgement	vi
Table of Contents	vii
List of Figures	ix
List of Figures	x
CHAPTER 1: INTRODUCTION	1
1.1 DATA MINING	1
1.2 KNOWLEDGE DISCOVERY PROCESS	1
1.3 DATA MINING TECHNIQUES	3
1.3.1 CLASSIFICATION	3
1.3.2 CLUSTERING	5
1.3.3 ASSOCIATION RULES	
1.4 DATA MINING APPLICATIONS	5
1.5 INTRODUCTION TO FUZZY LOGIC	5
1.6 ORGANIZATION OF THESIS	7
CHAPTER 2: REVIEW OF LITERATURE	8
CHAPTER 3: PRESENT WORK	11

3.1 PROBLEM FORMULATION	11
3.3 OBJECTIVES OF THE STUDY	11
3.3 RESEARCH METHODOLOGY	12
CHAPTER4: RESULTS AND DISCUSSION	22
4.1 EXPERIMENTAL RESULT	22
4.2 IMPROVEMENT IN RESULT	27
CHAPTER5: CONCLUSION AND FUTURE SCOPE	37
5.1 CONCLUSION	37
5.2 FUTURE SCOPE	37
REFERENCES	38

LIST OF FIGURES

FIGURE NO.	FIGURE DESCRIPTION	PAGE NO
FIGURE 1	Knowledge discovery process	2
FIGURE 2	Decision tree	4
FIGURE 3	Example of crisp logic	6
FIGURE 4	Example of fuzzy logic	6
FIGURE 5	Main structure of a fuzzy system	12
FIGURE 6	Membership function for every input and the output	14
FIGURE 7	Mapping of input to output for student performance	19
FIGURE 8	Mapping of input to output for dropout	20
FIGURE 9	Flow chart of research methodology	20
FIGURE 10	Fuzzy input and output variable for student performance	22
FIGURE 11	Defining membership function	23
FIGURE 12	Designing the rule base	24
FIGURE 13	Predict and classify the students' performance	24
FIGURE 14	Fuzzy input and output variable for dropout	25
FIGURE 15	Defining Membership Function	25
FIGURE 16	Designing the Rule Base	26
FIGURE 17	Predict and classify the dropout	27
FIGURE 18	User interface of improve accuracy of student performance	27
FIGURE 19	Improve accuracy of	

	student performance	28
FIGURE 20	Accuracy,mse,rmse and correlation coefficient of student performance using centroid	29
FIGURE 21	Accuracy,mse,rmse and correlation coefficient of student performance using bisector	29
FIGURE 22	Accuracy,mse,rmse and correlation Coefficient of student performance using som	30
FIGURE 23	Accuracy,mse,rmse and correlation Coefficient of student performance using mom	30
FIGURE 24	Accuracy,mse,rmse and correlation Coefficient of student performance using lom	31
FIGURE 25	Accuracy comparison	32
FIGURE 26	MSE comparison	32
FIGURE 27	RMSE comparison	32
FIGURE 28	Correlation coefficient comparison	33
FIGURE 29	Accuracy,mse,rmse and correlation Coefficient of dropout using centroid	33
FIGURE 30	Accuracy,mse,rmse and correlation coefficient of dropout using bisector	34
FIGURE 31	Accuracy,mse,rmse and correlation Coefficient of dropout using som	34
FIGURE 32	Correlation coefficient comparison	35
FIGURE 33	RMSE comparison	35

FIGURE 34	MSE comparison	36
FIGURE 35	Accuracy comparison	36

LIST OF TABLES

TABLE NO.	TABLE DESCRIPTION	PAGE NO.
TABLE 1	Classification of student performance	15
TABLE 2	Classification of dropout rate	16
TABLE 3	Student performance rule base for FIS	16
TABLE 4	Dropout rate rule base for FIS	17
TABLE 5	Improve accuracy of student performance	28
TABLE 6	Comparative analysis of student performance by using different methods	31
TABLE 7	Comparative analysis of dropout by using different methods	35

CHAPTER 1

INTRODUCTION

1.1 Data Mining

It is the way of extraction of information and patterns from the very large number (or quantity) of data. In the present day, the use of data mining is growing day by day. It is used for decision making activities. In some cases data mining are called information or Knowledge discovery since it is the way toward examining the information from alternate points of view and compressing it into valuable information. Microsoft Academic Research provides ranking for data mining key words which have been grown recently (used in publications, organizations etc.). Few of them are listed below

- Association rules: These are if/then statements that are used to display the relationships between apparently independent data in a relational database or any other information store.
- Machine Learning: Both data mining and machine learning used same methods. But there is difference, machine learning focused on prediction, based on known properties, whereas data mining focuses on identification of unknown properties.
- Support Vector Machine: It is supervised learning algorithm which analyzes data used for classification.
- Cluster Algorithms: Clustering is one of the emerging research fields in data mining due to its numerous applications. Example : K-means
- Information Retrieval
- Search Engine
- Web Search
- Indexation
- Social Network

Data Mining is the process to find the hidden information as well as pattern from a bulk amount of data i.e. the data should be coming from different sources such as data ware house, Data mart etc.

1.2 Knowledge Discovery Process

- Knowledge Discovery Process it is also called KDD. It is a large-scale process of discovering the useful knowledge from the collection of facts and figures.
- Objective of KDD procedure is to extricate the information from data or facts in the surrounding of big databases.
- The application areas of KDD is marketing and manufacturing

The below figure shows the process of knowledge discovering:-

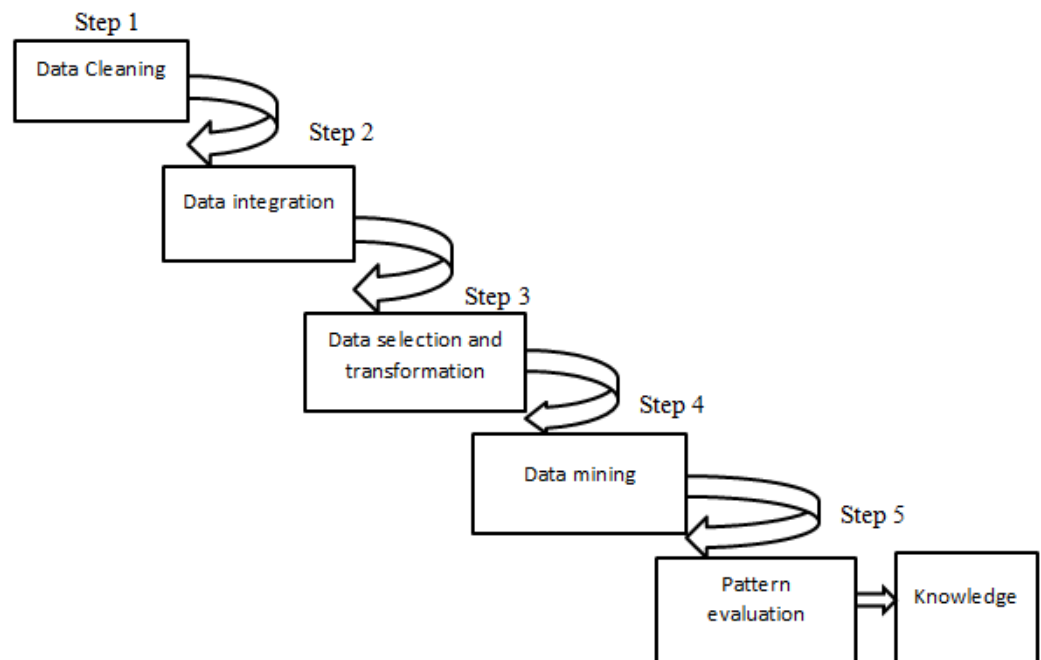


Figure 1: Knowledge discovery process

. There are several steps which are as follows:

- Cleaning of Data: It is a first step in this stage dirty data and inconsistent facts or data are eliminated or discard.
- Integration of Data: In this phase, different sources of data are integrated.
- Selection of Data: This is a phase in which selected the data that are used for analyzing.
- Transformation of Data: it is the stage in which chosen data is modify into many forms.

- **Data Mining:** This is the essential level in this step intelligent methods are used for searching the patterns.
- **Evaluation in Patterns:** To identify the factually interesting patterns representing knowledge based on some interestingness measured.
- **Knowledge representation:** It is the last step, in this Knowledge is presented.

1.3 Data Mining Techniques

1.3.1 Classification: The classification is done because of exactly guess the aimed class for all case in the data. One of the example of this model is it is help to predict the student performance.

- In inclusion, there are two stages in classification. The initial part is the learning process. In this part, the training data or facts are examined by classification algorithm and rules and design are created which are based on learned model or classifier.
- In the second part the model is used for classification and testing data are used for gaining the accuracy of classification design. Then, establish on the sufficient accuracy, the rules can be used for the classification of new or recently developed data or for unseen data.

1.3.1.1 Decision Tree: Decision trees are broadly used in the classification procedure. With the help of this, the model can be predicted and classified. Decision trees shows rules, which may followed by individual and used in knowledge structure like a database.

- **Example:** if the attendance is not matching the giving criteria then the chances of giving the exam is less.
- **Decision Tree can be represented as:**

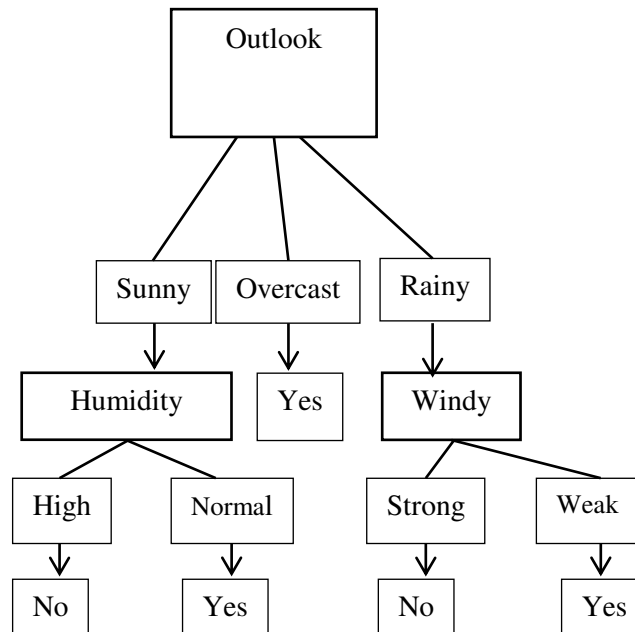


Figure 2: Decision tree

It's like a flowchart. In this rectangular shapes of boxes are called node internal nodes are those nodes that have a child and the leaf node are those who don't have children. The top nodes are root node. In the given figure outlook is a root node. Humidity and Windy is an internal node.

1.3.1.2 Naïve Bayes: It is a conditional probability approach. In which a mention problem case is to be classified, and it can be elected by a vector $x=(x_1, x_2, \dots, x_n)$ representing a few n features. Using Bayes theorem it is written as

Posterior=prior X likelihood/evidence.

1.3.1.3 Fuzzy logic: It is a method to determine the “degree of facts” instead of the general “true or false” (1 or 0).Data mining uses different methods (approaches) and assumption from a broad areas or fields for the knowledge extraction from huge amount of data. But uncertainty is a general phenomenon in data mining problems. Therefore, it is applied to manage with the uncertainty in actual world.

1.3.2 Clustering

Clustering is a procedure of dividing a gathering of information (or articles) into an arrangement of significant family, called cluster groups. Clustering can be utilized as stand-alone tool to get inside into information distribution or it can be utilized as pre-processing step for different calculations.

1.3.3 Association rules

Association rules are if/then articulations that are utilized to show the relationships between apparently independent data in a relational database or some other data store. Case of this would be "If a client purchases a dozen of eggs, he is 80% possibility also to buy the milk."

1.4 Data Mining Applications

Data mining is extremely use because of its many benefits. In this there are numerous low cost techniques to collect and manage the data or facts, but they are some approaches for extracting helpful knowledge from this data. Data mining has different applications in multiple fields.

- **Marketing and Retailing:** Marketers can make scheme to fulfill the each and every requirements and understand their buying behaviors with the help of data mining.
- **Banking:** Financial organizations can acquire the help of data mining in credit and loan details. A credit card issuer can detect fraud credit card transaction.
- **Research and Development:** Using data mining approaches researchers extract the knowledge by analyzing the data and precede their research work.
- **Education:** Data mining is very helpful in educational organizations or institutes because there is a large number of unused collected data and this data can be used in a proper way using data mining [1].

1.5 Introduction to Fuzzy Logic

As this thesis is organised around Fuzzy Logic, so below are discussed some basic concepts of fuzzy logic.

Fuzzy logic is a developed form of Boolean logic. According to Boolean logic (also known as crisp logic) everything can be indicated in binary terms (0 & 1). Consider

the Query “Temperature is 32’c”. The answer to this is a definite Yes/True or No/False as warranted by the situation. If Yes/True is assigned a value 1 and No/false is assigned a value 0, this statement results in a 0/1 type of situation. Such a logic which demands a binary type of handling is termed as crisp in the domain of Fuzzy set theory. Thus statements like “Is water colourless” are examples of crisp situations.

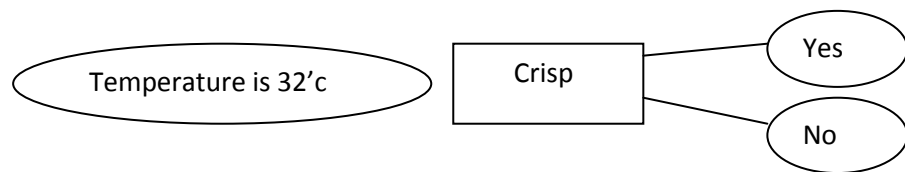


Figure 3: Example of crisp logic [13]

Then again, consider the statement “Is Saloni Honest”. The response to this question require not be a clear Yes or No. Considering how much one knows Saloni, an assortment of answers like “extremely honest”, “very honest”, “honest at times” could be created.

If for instance extremely honest is assigned a value 1, extremely dishonest a value of 0, honest at times and very honest can be assigned values of 0.4 & 0.85 respectively. Such a situation which can accept values between 0 & 1 is termed fuzzy.

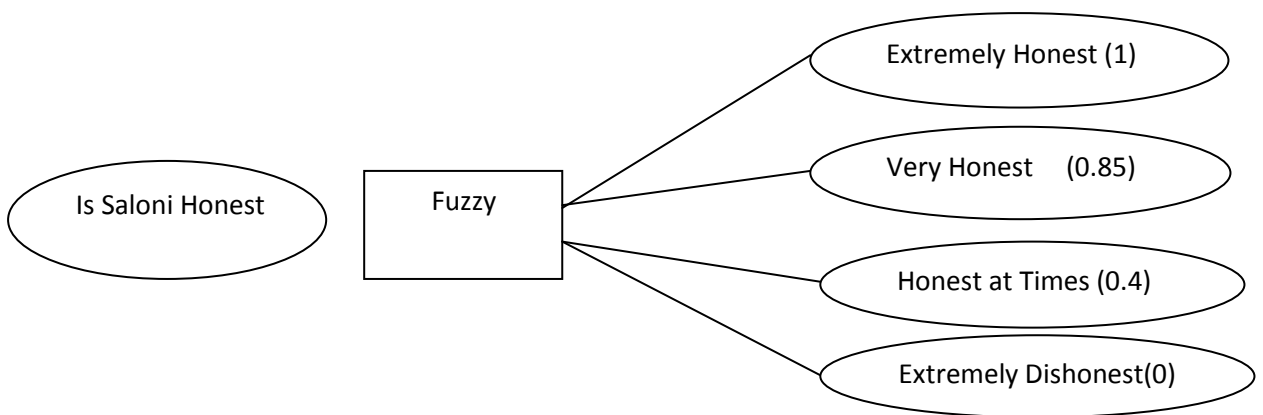


Figure 4: Example of Fuzzy logic[13]

1.5.1 Membership Function

While in crisp set theory, a component either has a place with or does not have a place with a set, in Fuzzy set theory many degrees of membership between 0 & 1 are allowed.

Thus a membership function $\mu_A(x)$ is related with a Fuzzy set A to such an extent that the capacity maps each component of the all-inclusive set X to the interim [0,1].

$$\mu_A(x): X \longrightarrow [0, 1]$$

It can be indicated in the form of a curve that characterizes how each point in the input space is mapped to a membership value or a level of truth in the vicinity of 0 and 1. The most widely recognized state of a membership function is triangular, although trapezoidal and bell curves are additionally utilized.

A Fuzzy set is characterized as:

In the event that X is an all-inclusive set and x is a specific component of X, then Fuzzy set A characterized on X might be composed as an accumulation of requested sets

$$A = \{(x, \mu_A(x)), x \in X\}$$

Where each match $(x, \mu_A(x))$ is known as a singleton.

Example: Let $X = \{a,b,c,d,e\}$ be the reference set of students. Then the Fuzzy set A based on linguistic term “smart” can be represented as

$$A = \{(a, 0.4), (b, 0.5), (c, 1), (d, 0.9), (e, 0.8)\}$$

1.6 Organization of Thesis

This thesis is partitioned into five sections. Chapter 1 contains a summary of introduction to the area of this thesis work. It also includes some basic fundamentals of fuzzy logic as this thesis is based on a fuzzy approach. Chapter 2 gives an overview of literature review. It discusses the work done till now in this area of research. Chapter 3 presents the problem statement; and discusses the proposed methodology in this thesis work. Further it includes information about the simulation methodology used in carrying out the work.

Chapter 4 outlines the work which has been carried out. It contains the results and discussion of the observations. Chapter 5 contains conclusion of the thesis and directions in which work can be extended in future.

CHAPTER 2

LITERATURE REVIEW

This chapter include the literature overview about the student performance or dropout. Numerous journals and conference papers, articles identified with our point are considered and some of those papers are given beneath. This chapter presents the basic concept and proposed theory of the papers which are studied.

MuslihahWook et al.(2012) [2]This aim of the research is to use data mining methods such as Artificial Neural Network (ANN) and the mixture of clustering and decision tree approaches to predict and classify academic achievement of the students in NDUM.

.G.Kesavaraj et al.(2013) [3] The objective of the paper is to give a complete description of various classification methods in data mining. Paper represents the some kinds of classification methods along with decision tree induction, k-nearest neighbour classifier Bayesian networks. The objective of classification algorithms is to produce more certain, precise and accurate outcomes.

V.K.Deepa et al.(2013) [4] Data Mining and its application is not a modern word, but in the present days its growth is increasing. Now, it has scattered in almost all the areas. This paper describes data mining application in exactly three areas such as Heart Disease Prediction, Intrusion Detection and Business process development. It is cleared that Data mining instrument or tools supports in extracting meaningful knowledgeable attributes from the impossible huge data.

Yohannes Kurniawan et al.(2013)[5] The paper discuss when the students data are modified such as assignment, attendance and score the prediction can also change. These predictions may not be suitable or proper if the data is not available (completed). This paper showed that how helpful data warehouse and data mining can be used in the schools to estimate the final progress of the students. This paper concentrates on estimating the student progress by using the benefit of data mining techniques.

K V Krishna Kishore et al.(2014)[6] This paper uses the students Past educational History, Regularity, Degree of Intelligence, Grade ,No. of Backlogs, and Working Nature to predict the GPA of the students. This paper uses these techniques J48, Multilayer Perceptron Naïve Bayes, CART to predict the GPA based on these factors or attributes. The result concluded that from these other classification methods, the Multilayer Perceptron has attained 97.37% accuracy on test data set. After predicting the student's data those student who have probability of failing provide an extra classes and material to make better the academic progress of those students.

M. Mayilvaganan et al.(2014) [7] The aim of this research is to improve the Prediction techniques that are used to analyse the skill expertise based on the academic achievement this paper is used to by the range of knowledge. With the help of weka tool the experimentation has been done. To analyse the performance of students this paper also compares the algorithms such as C4.5, Naïve Bayes and Multi-labelled K-Nearest Neighbour to find the accuracy of classification and decision tree algorithms. The result is decided Multi-labelled K- Nearest Neighbour has high (great) accuracy of time taken in classification as contrast to alternative techniques by the importance of examination outcome.

Nguyen Thai Nghe et al.(2014) [8] This research paper compares the accuracy of Bayesian Network and Decision tree algorithm for estimating the studious performance of students. Based on it they will get a scholarship.

Sanjay Kumar et al.(2014) [9] In this paper, firstly predict the 5 factors or attributes that may influence the students' performance. The factors include teaching methodology, university system, university environment, personal reasons and family issues. A system has been designed on basis of these different parameters which will decide the performance of students. This work is done with the help of Matlab software. In this software, fuzzy logic is used. The information is collected from the engineering students of LPU. Based on this data, a system is designed which will decide the achievement will be increased or decreased or remain consistent.

Subaira.A.S et al.(2015) [10] This paper describes the various approaches such as Neural network, K- Nearest Neighbour, Bayesian Classifier, Fuzzy Logic and decision tree classification Algorithms for implementation of intrusion Detection system. With the help of this paper, it is clear that the data mining methods are used to

perform the intrusion detection system But this paper don't describe which technique is best for all of these (means which is used for better result).

J.Naren et al.(2016) [11]Feedback is an important action for each and every student. The judgement makes help for students and teachers so that teacher knows the students capability and implement new methods for teaching.it is done with the help of weka.

Pooja Asopa et al. (2016) [12] This paper uses the students exam performance and class record to predict the performance of the students. This paper uses fuzzy-logic to predict the performance based on these factors or attributes. Based on these it has attained 72% accuracy on data set.

CHAPTER 3

PRESENT WORK

3.1 Problem Formulation

In our thesis we have designed a fuzzy inference system that compute a student performance of the first semester and dropout rate as most of the students drop the college in the first semester. There is a need to identify various factors that lead to dropout and degradation in student performance. Firstly, a discussion is held with highly experienced authorities of Lovely Professional University (LPU) with the reason to identify the critical factors that may be affecting students' performance and dropout rate. After that, a one to one formal interaction with engineering students of LPU, is done to identify which kind of factors are affecting particular student. Based upon the collected information, a system is developed. This study will help the institutions and universities to cope up with the problem of student dropping out of the studies. This study will be useful for making certain changes in academic structure etc. This study provides the insight of the issues and reasons of dropping the study in between which may further be useful for the universities and education institutes to enhance their efforts in this particular area.

3.2 Objective of Study

Main objectives of this research work are:

1. To predict student's academic performance based on their past academic performance
2. To predict the factors which can affect the grade of student that leads to dropout.

Fuzzy Logic is used to predict students' performance of the first semester and Dropout rate. It is hoped that the prediction could be done in the early semesters and intervention given to the students in the form of extra tutorials and so on to help them improve their academic performance and thus achieve

a better grade than that as predicted. This study will help the institutions and universities to cope up with the problem of student dropping out of the studies. This study will be useful for making certain changes in academic structure etc. This study provides the insight of the issues and reasons of dropping the study in between which may further be useful for the universities and education institutes to enhance their efforts in this particular area.

3.3 Research Methodology

The proposed method uses a Fuzzy system for the prediction of the Students Final Achievement (SFA) of the first semester and Dropout rate. Figure 5 demonstrates the fundamental structure of a Fuzzy Framework. In this structure a Fuzzy framework comprises of four segments: Fuzzifier, Rule Base, Inference engine, Defuzzifier.

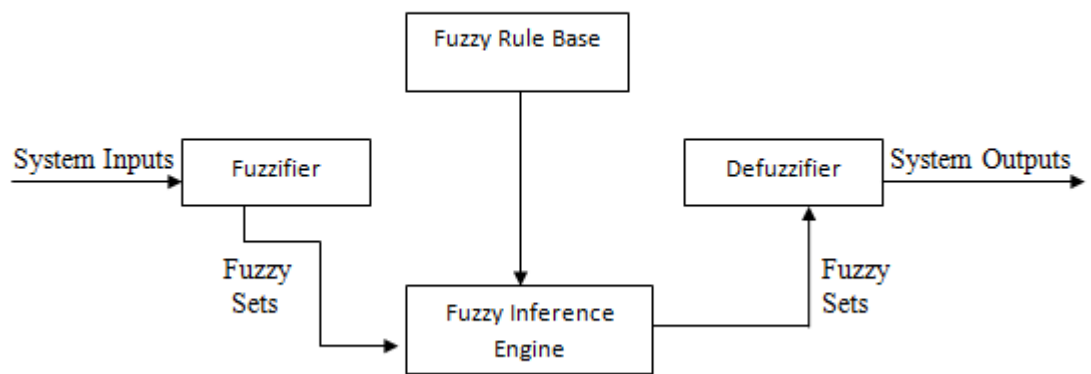


Figure 5: Main structure of a fuzzy system [13]

A fuzzy system is an information-based rule system. The core of this system is a database which is configured with if-then rules. A fuzzy inference system (FIS) tries to conclude answers from a knowledgebase by utilizing a fuzzy inference engine. The inference engine which is examined to be the brain of the expert systems gives the methodologies to reasoning around the data in the knowledgebase and explains the results. Fuzzy Inference Systems are extremely very basic.

A FIS comprises of an information stage, a preparing stage, and output stage. The information arranges maps the inputs to the appropriate membership functions and truth values. The preparing stage invokes each appropriate rule and generates a

result for each. It then joins the results of the rules. At last, the output stage converts the combined result back into a specific output value.

As examined before, the preparing stage, which is known as the inference engine, is depends upon a group of logic rules in the form of IF-THEN explanations, where the IF part is known as the "predecessor" and the THEN part is known as the "subsequent". Ordinary fuzzy inference subsystems have many rules. These rules are put away in a knowledgebase.

A case of fuzzy IF-THEN standards are:

“IF Continuous Assessment of a student is Low and mid-term is Low and Attendance is medium and health is Critical, and Friend circle is Bad THEN Student Final Achievement (SFA) will be Low” in which Continuous Assessment of a student, mid-term, Attendance, health, Friend circle are linguistics variables and low, critical and Bad are linguistics terms.

Another case for dropout is

“IF financial issue of student is moderate and confused vision is moderate and less self-confidence is Moderate and Student Final Achievement (SFA) is Low THEN Chances of Dropout will be moderate” in which financial issue of student, confused vision, less self-confidence , Student Final Achievement (SFA) are linguistics variables and moderate are linguistics terms.

3.3.1 Overview of the Proposed Algorithm

3.3.1.1 Introduction

This algorithm is used for predicting the student performance and dropout. It uses a triangular and trapezoidal membership function.

If the inputs or the Continuous Assessment will be in the scope of 0 to 20, then three membership functions for each input and the output would be chosen as appeared in figure 7

Type – Triangular, Range 0 – 20, Low [0, 5, 10], Medium [5, 10, 15], High [10, 15, 20]

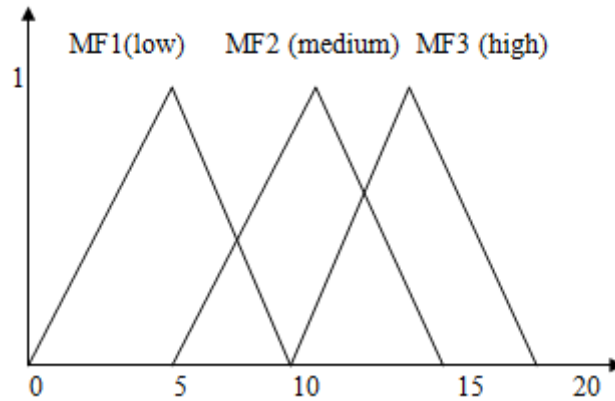


Figure 6: Membership function pattern for every input and the output [14]

In case we have more than one input variable (in fact, the case we have), the degree of membership for the output value will be the minimum value of the degree of membership for the various inputs. Fuzzy Linguistic Variables and their membership value for both student performance and dropout are shown as

Fuzzy Linguistic Variables and their membership value for Student Performance Achievement are shown as

Membership Functions:

For Continuous Assessment

Range: 0-100, low-[0-40], medium-[30-70], High [60-100]

For Mid term

Range: 0-100, low-[0-40], medium-[30-70], High [60-100]

For Attendance

Range: 0-100, low-[0-75], medium-[70-80], High [80-100]

For Health issue

Range: 0-5, Critical-[0-2], mild-[1-4], Fit [3-6]

For friend Circle

Range: 0-5, Bad-[0-2], Average-[1-4], Good [3-6]

Student Performance Achievement can be classified as

Range: 0-100

No.	Grade value	Classification label
1	80-100	High
2	65-85	Above average
3	50-70	Average
4	35-55	Below average
5	0-40	Low
6	0-1	Detained

Table 1: Classification of student performance

Similarly, for Dropout are shown as

For financial problem

Range: 0-5, low-[0-2], medium-[1-4], High [3-5]

For Confused vision

Range: 0-5, low-[0-2], medium-[1-4], High [3-5]

For less self-confidence

Range: 0-5, low-[0-2], medium-[1-4], High [3-5]

For Student Performance Achievement

Range: 0-100, Detained-[0-1], Low-[0-40], below average [35-55], average [50-70], above average [65-85], High [80-100]

Chances of dropout rate can be classified as

Range: 0-10

No.	Dropout value	Classification label
1	7-10	High
2	3-8	Medium
3	0-4	Low

Table 2: Classification of dropout rate

3.3.1.2 Rule Base for FIS

Table 3 shows the rule base used for the proposed FIS (Fuzzy Inference system). It is a group of logic rules in the form of IF-THEN explanations, where the IF part is known as the "predecessor" and the THEN part is known as the "subsequent". Ordinary fuzzy inference subsystems have many rules. Some of them are as follow

S.No	Attendance	Mid-term	CA	Friend circle	Health issue	Z (output)
1	Low	Low	Low	Bad	Critical	Low
2	Low	Low	Low	Bad	Mild	Low
3	Medium	Low	Low	High	Fit	Below avg.
4	High	High	High	Medium	Critical	Above avg
5	High	High	High	Medium	Mild	High
6	High	High	High	Medium	Fit	High
7	High	High	High	High	Critical	Above avg.
8	High	High	High	High	Mild	High
9	High	High	High	High	Fit	High

Table 3: Student performance rule base for FIS

S.No	Financial problem	Confused vision	Less self-confidence	Student Final Achievement	Z(Output)
1	Yes	Yes	Yes	Detained	High
2	Yes	Yes	Yes	Low	High
3	Yes	Yes	Yes	Below avg.	High
4	Yes	Yes	Yes	Average	High
5	Yes	Yes	Yes	Above avg.	High
6	Yes	Yes	Yes	High	High
7	Medium	Yes	Yes	Detained	Medium
8	Medium	Yes	Yes	Low	Medium
9	No	Medium	Medium	Below avg.	Low

Table 4: Dropout rate rule base for FIS

3.3.1.3 Fuzzy Inference Engine

Then the degree of membership for the output value will be defuzzified by using Mamdani's defuzzification method. Mamdani has several methods i.e Centroid, Bisector, Mom, Lom and Som. Centroid method is the most commonly used defuzzification method. The formula for centroid method is given below.

$$X^* = \frac{\sum A \bar{x}}{\sum A} \dots\dots\dots(v)$$

Where,

- i. X* is the desired crisp value (Next Predicted CPU burst)
- ii. A is the area of the segment of aggregated fuzzy set.
- iii. \bar{x}_i is the corresponding centroid

Research Methodology involves following steps:

- 1. Data Collection:** The process of data gathering that required collecting all information about students. The group of attributes or factors should be find that can affect student's performance and collected from available data source.
- 2. Data Pre-processing:** It is very important task. It removes the noisy and inconsistent data from the collected data. Pre- processing techniques such as data separating and variable transformation and other techniques for attribute selection must be used.
- 3. Data Preparation:** This step is performed after data Pre-processing once the noisy data is removed then the preparation of data is easy. In addition to this stage choose the factors that are used to predict the student performance.
- 4. Define input and output variable:** The input variables in a fuzzy system are mapped by group of membership functions alike to this known as "fuzzy sets". The procedure of converting a crisp input value to a fuzzy value is known as "fuzzification".

An Example of fuzzy input variable and output variable is:-

IF Continuous Assessment of a student is Low and mid-term is Low and Attendance is medium and health is Critical, and Friend circle is bad THEN Student Final Achievement (SFA) will be Low".

In this example, the input variables are "CA" , "mid-term" ,"Attendance" ,"Health issue" and "friend circle" that have values defined as fuzzy sets. The output variable "Student Final Achievement" is also defined by a fuzzy set.

- 5. Selecting membership function (MF):** In fuzzy logic there are many membership functions like triangular, gaussian, Trapezoidal. This is the one of the important part of the fuzzy logic. It is a curve that describes how each and every point in the input space is mapped to a membership value between 0 and 1.

The basic membership functions are formed by using the straight lines. The simplest is the triangular function instead of others, and it has the function name trimf. It is a collection of three points forming a triangle. The trapezoidal membership function, trapmf, has a flat top and it is a truncated triangle curve. These straight line membership functions have the benefit of simplicity.

6. Fuzzy rules: There are large numbers of fuzzy rules are made. Example of the fuzzy rule is:

If a is P then b is Q

In a Fuzzy sets, a and b are linguistic values on the ranges P and Q respectively. Here 'a is P' define antecedent and 'b is Q' define consequent or conclusion.

7. Apply fuzzy operators: A fuzzy set operation is an operation on fuzzy sets. These operations are generalisation on crisp set operation. The most broadly used operations are known as standard fuzzy set operations. There are 3 operations which are as:

a). Fuzzy complement: it is also called NOT. In this complement of its value is taken.

b). Fuzzy intersection:-In this minimum values are considered.

c). Fuzzy union:-It is also called OR .In this maximum values are considered.

8. Fuzzy Inference System (FIS): Fuzzy inference is the procedure of expressing input/output mapping using by applying the fuzzy logic. The Fuzzy Logic Toolbox software gives command-line functions .These functions are

a) Design Fuzzy system

b) Choose a membership function

c) Define fuzzy rules

d) Evaluate and See fuzzy structure.

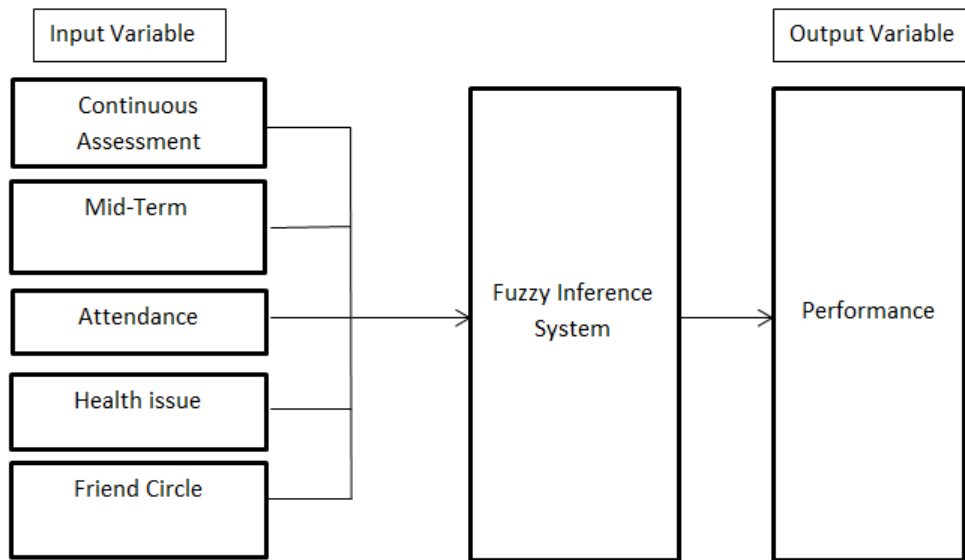


Figure 7: Mapping of input to output for student performance

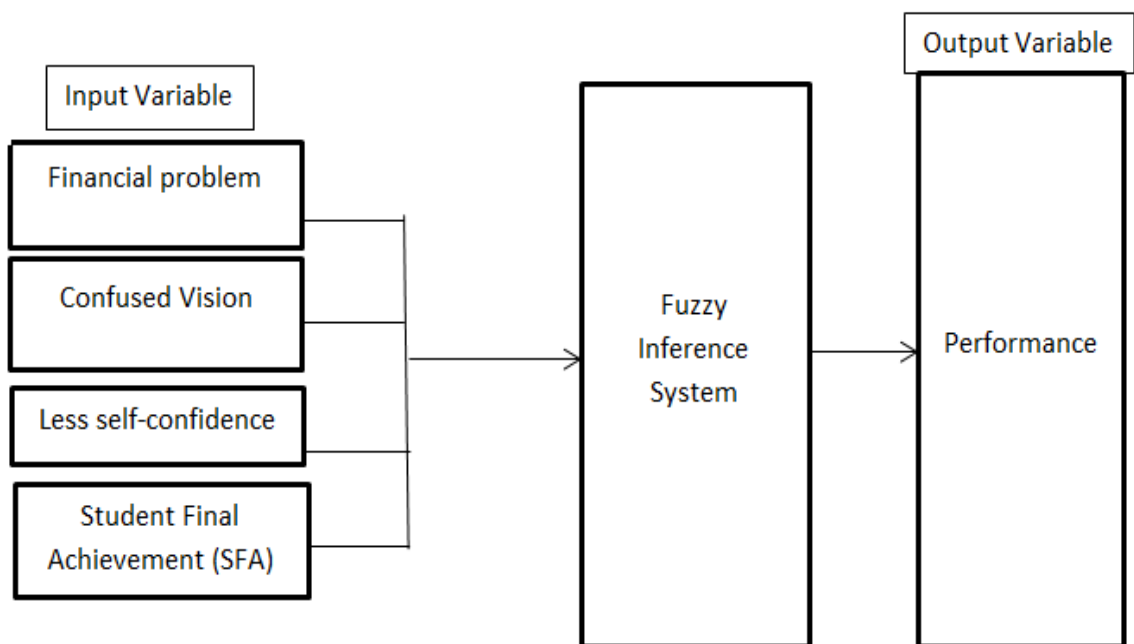


Figure 8: Mapping of input to output for dropout (Design for FIS)

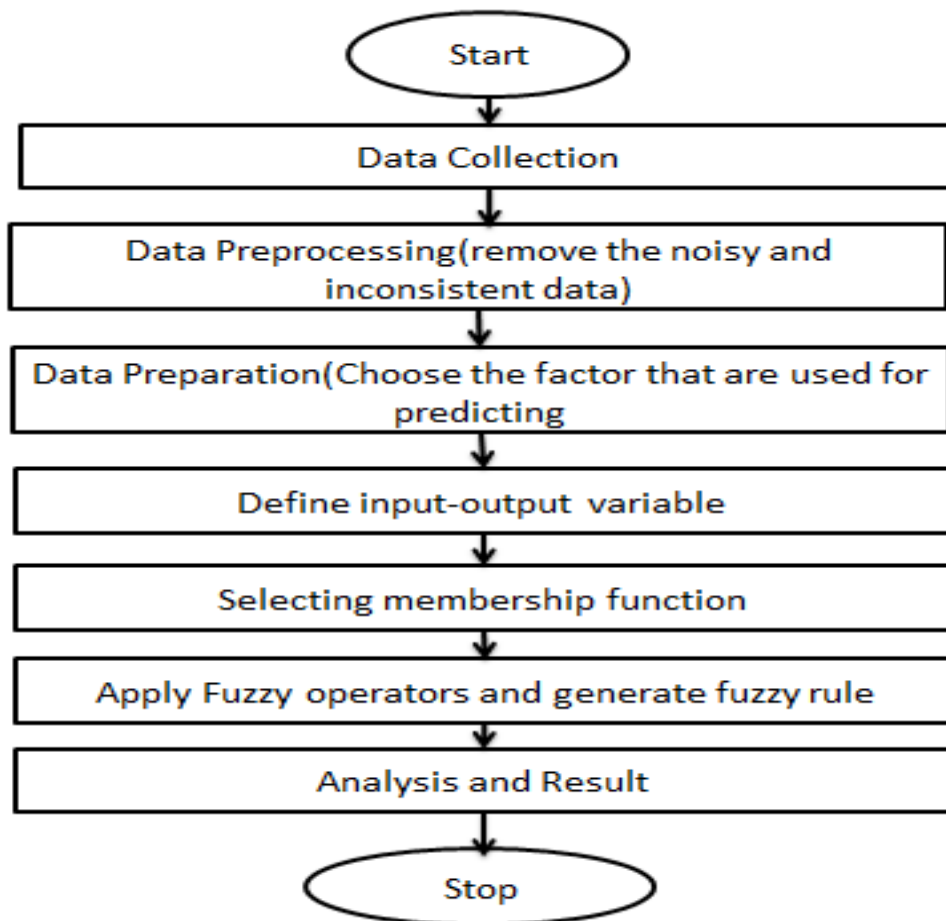


Figure 9: Research Methodology

3.3.2 Simulation Methodology

MATLAB is used to simulate the proposed Fuzzy based algorithm. Matlab is a numerical figuring condition and fourth-period programming language. It permit plotting of functions and information, matrix manipulations, executing of calculations, making of UIs, and interfacing with programs written in various languages, including C, Fortran and C++.

Fuzzy-based algorithm is simulated using Fuzzy logic toolbox in Matlab. This tool kit lets you demonstrate complex framework behavior utilizing basic logic rules and then executes these rules in a fuzzy inference System (FIS). Fuzzy Logic Toolbox is a gathering of functions based on the MATLAB numeric figuring condition. It gives instruments for you to make and alter fuzzy inference systems inside the structure of MATLAB.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 EXPERIMENTAL RESULT

Fuzzy logic toolbox in MATLAB was used to implement the proposed Fuzzy based algorithm. Fuzzy rule based expert system has been designed to assess the performance of students and dropout rate on the basis of different parameters. Firstly, a discussion is held with highly experienced authorities of Lovely Professional University (LPU) with the reason to identify the critical factors that may be affecting students' performance and dropout rate. After that, a one to one formal interaction with engineering students of LPU, is done to identify which kind of factors are affecting particular student. Based upon the collected information, a system is developed. It was slightly found that the student's performance is not totally dependent on their academic efforts, in spite; there are many other factors that have equal to greater influences as well. This is shown in the screenshot below.

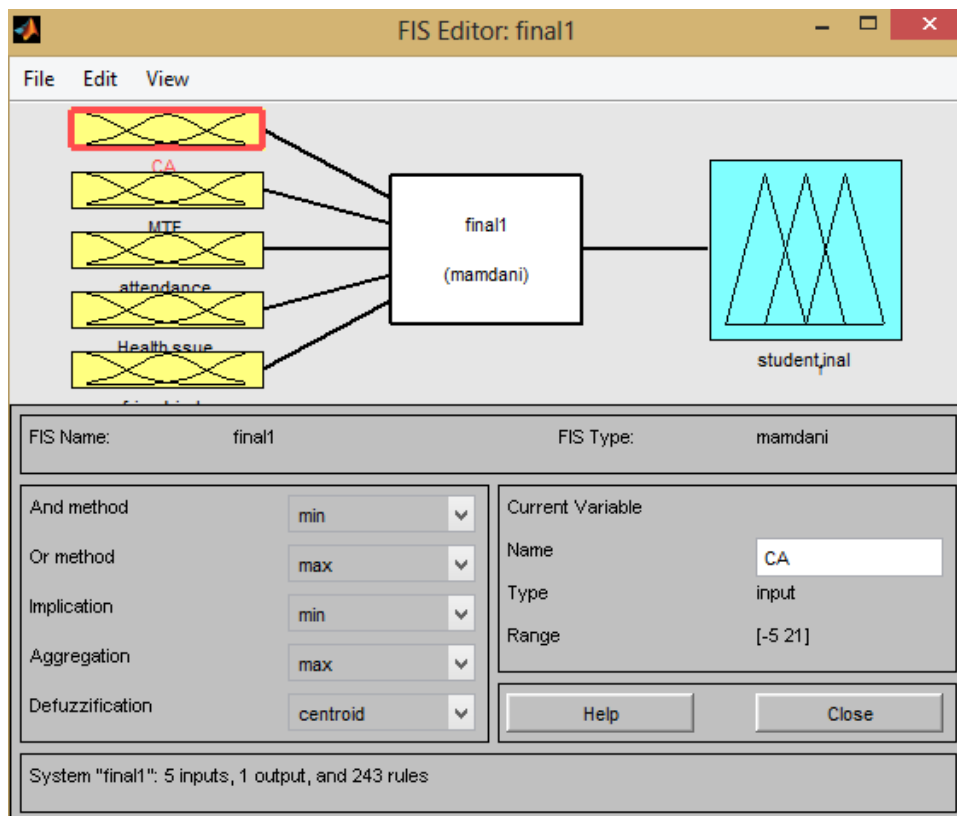


Figure 10: Fuzzy input and output variable for dropout

Then the membership function for each Input and Output was defined. Fuzzy logic tool box has a list of prominent membership functions, you just have to choose the one you want to use. We have used triangular and trapezoidal membership function in our work. This is shown in Figure 11

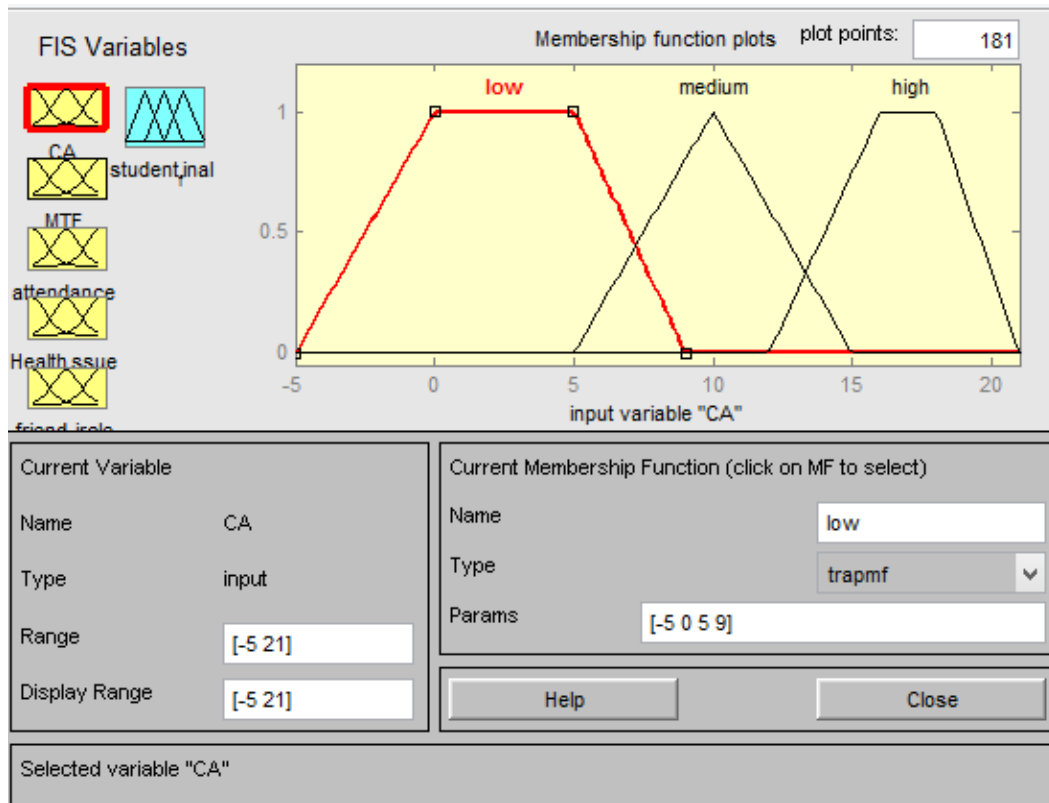


Fig 11: Defining membership function

After defining the membership functions, the rule base to be used was designed or added to the fuzzy system. Fuzzy logic toolbox allows you to simply add rules to the designed system. Figure 12 shows the rule base added. It has a total of 243 rules defined in it. Some of the rules are shown:

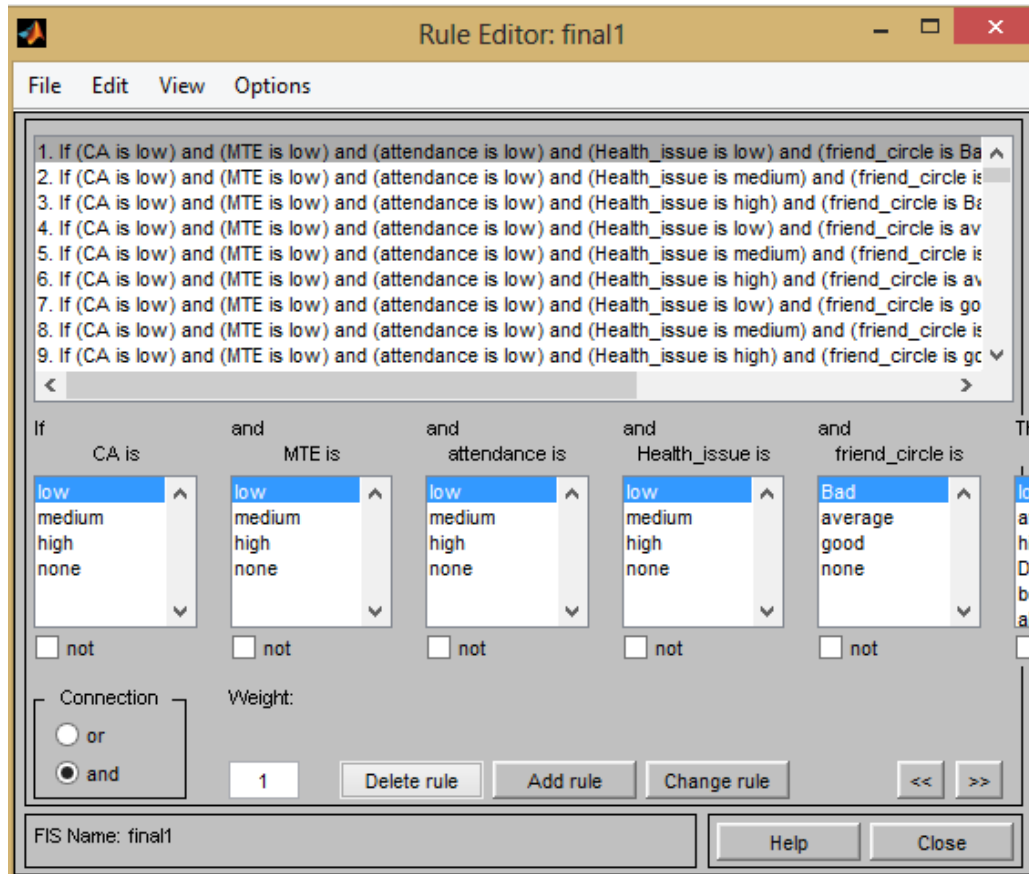


Figure 12: Designing the rule base

After that, we predict and classify the students' academic performance and it is divided into several categories.

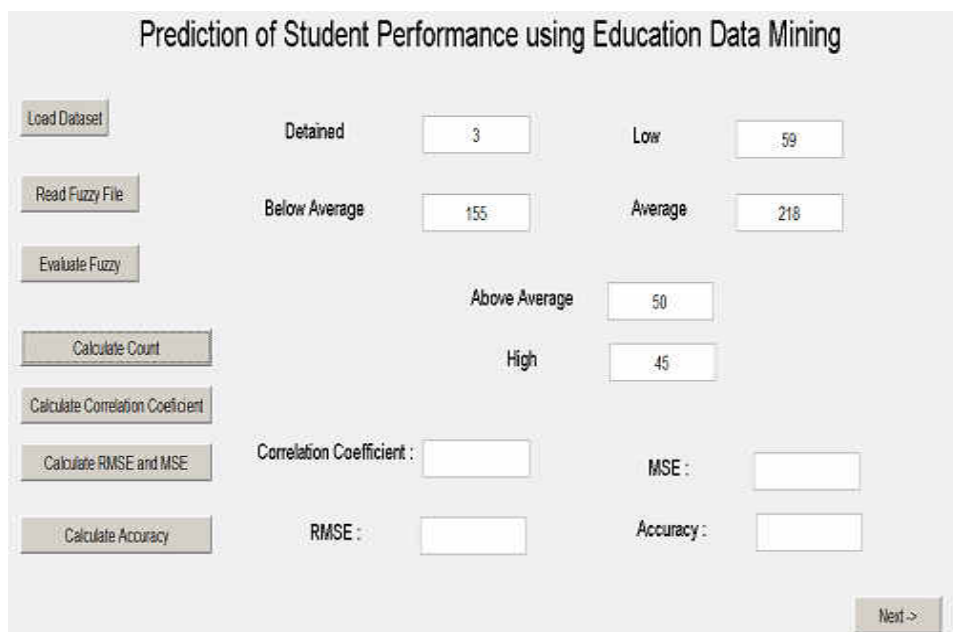


Figure 13: Predict and classify the students' performance

After this, we design the fuzzy system for Dropout.

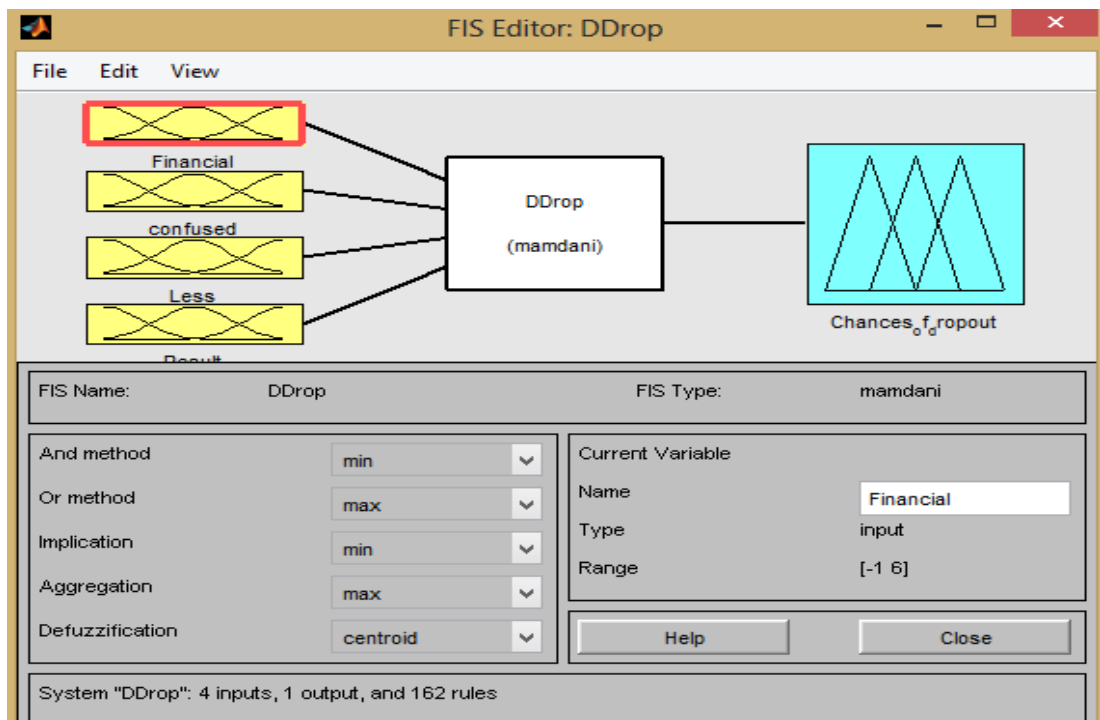


Figure 14: Fuzzy input and output variable for dropout

Then the membership function for each Input and Output was defined. Fuzzy logic tool box has a list of prominent membership functions, you just have to choose the one you want to use. We have used triangular and trapezoidal membership function in our work. This is shown in Figure 15

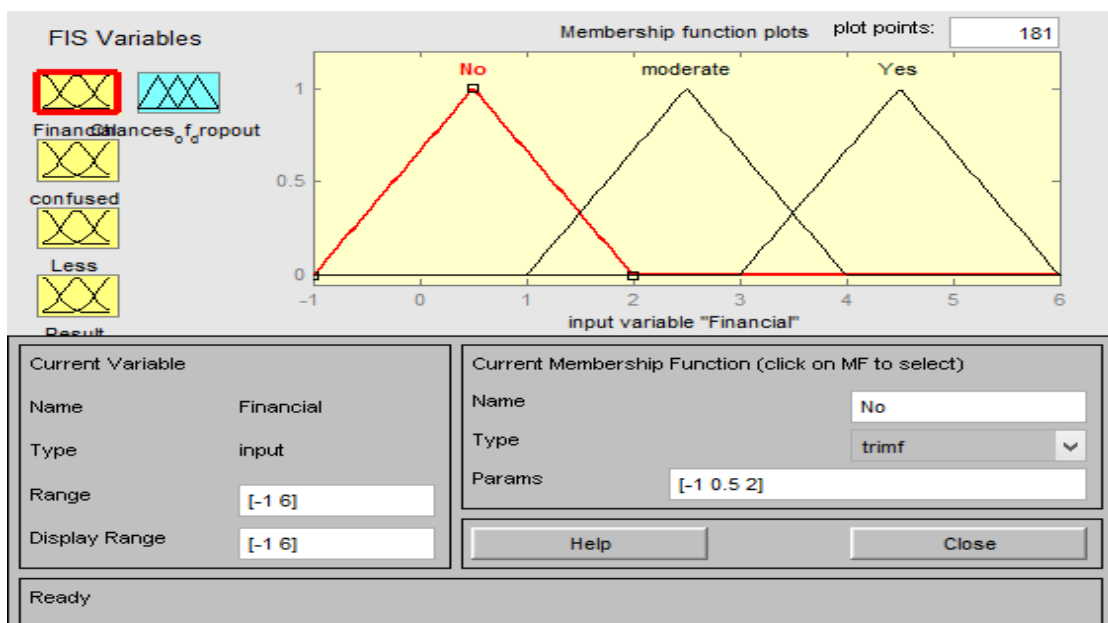


Fig 15: Defining membership function

After defining the membership functions, the rule base to be used was designed or added to the fuzzy system. Fuzzy logic toolbox allows you to simply add rules to the designed system. Figure16 shows the rule base added. It has a total of 162 rules defined in it. Some of the rules are shown:

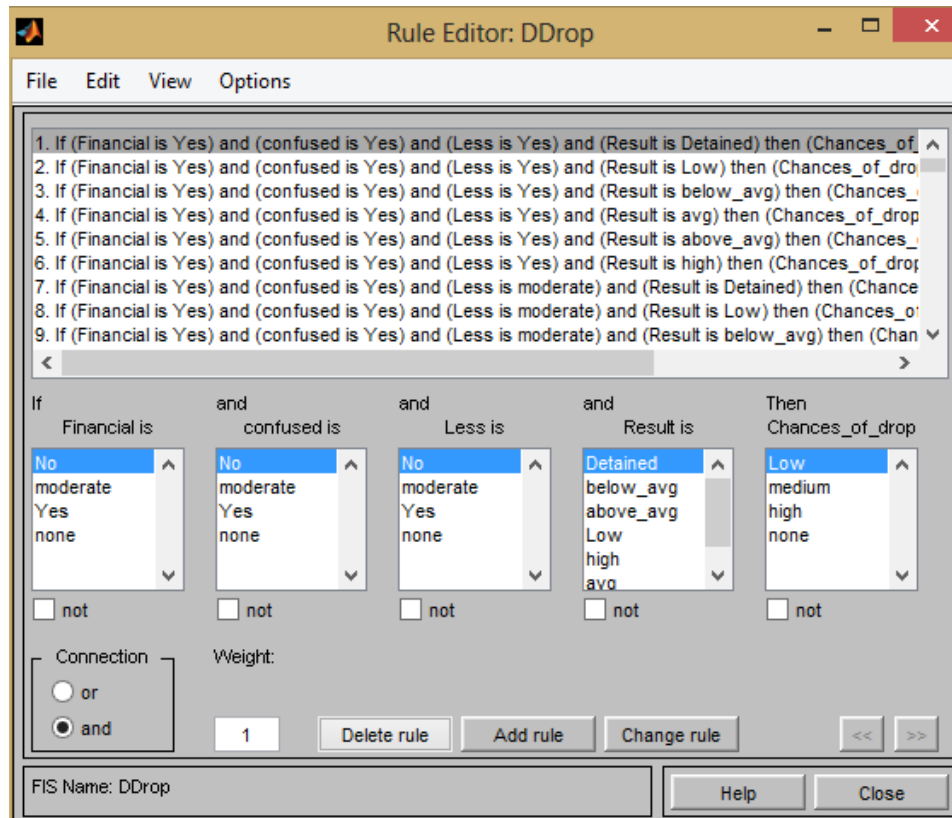


Figure 16: Designing the rule base

After that, we predict and classify the students' academic performance and it is divided into several categories. These are

1. Low chances of dropout
2. Medium chances of dropout
3. High. chances of dropout

Prediction of DropOuts using Education Data Mining

<input type="button" value="Load Dataset"/>	Less chance of Dropout	<input type="text" value="6"/>	Medium Chance of dropout	<input type="text" value="386"/>
<input type="button" value="Read Fuzzy File"/>	High Chance of dropout	<input type="text" value="138"/>		
<input type="button" value="Evaluate Fuzzy"/>				
<input type="button" value="Calculate Count"/>				
<input type="button" value="Calculate Correlation Coefficient"/>	Correlation Coefficient :	<input type="text"/>	MSE :	<input type="text"/>
<input type="button" value="Calculate RMSE and MSE"/>	RMSE :	<input type="text"/>	Accuracy :	<input type="text"/>
<input type="button" value="Calculate Accuracy"/>				

Figure 17: Predict and classify the dropout

4.2 IMPROVEMENT IN RESULT

The Improvement is shown in the result by comparing this accuracy level. This accuracy is find by using the centroid method.

Prediction of Student Performance using Education Data Mining

<input type="button" value="Load Dataset"/>	Detained	<input type="text"/>	Low	<input type="text"/>
<input type="button" value="Read Fuzzy File"/>	Below Average	<input type="text"/>	Average	<input type="text"/>
<input type="button" value="Evaluate Fuzzy"/>				
<input type="button" value="Calculate Count"/>			Above Average	<input type="text"/>
<input type="button" value="Calculate Correlation Coefficient"/>			High	<input type="text"/>
<input type="button" value="Calculate RMSE and MSE"/>	Correlation Coefficient :	<input type="text" value="0.79522"/>	MSE :	<input type="text" value="3.434"/>
<input type="button" value="Calculate Accuracy"/>	RMSE :	<input type="text" value="0.76725"/>	Accuracy :	<input type="text" value="98.586"/>
				<input type="button" value="Next ->"/>

Figure 18: User interface of improve accuracy of student performance [12]

Calculated accuracy [12]	Accuracy resulted from thesis work
72	96.56

Table 5: Improve accuracy of student performance [12]

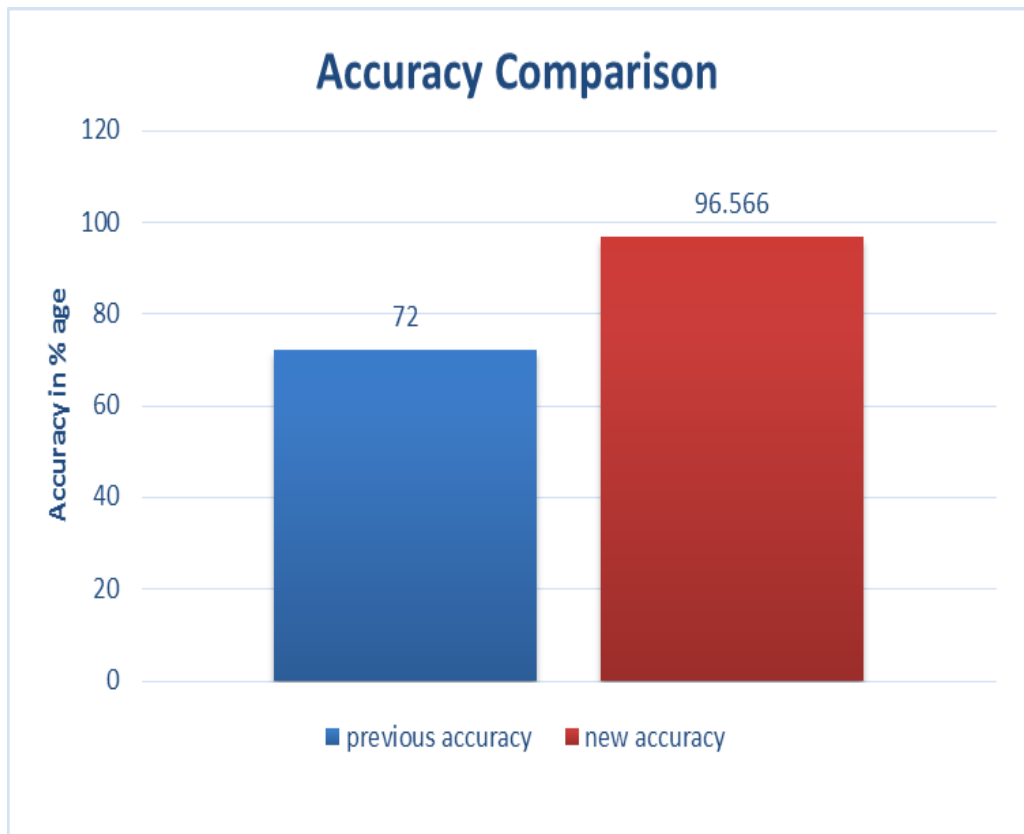


Figure 19: Improve accuracy of student performance [12]

We have also found the Accuracy, MSE, RMSE and Correlation coefficient by using Centroid, Bisector, Mom, Lom and Som method. It can be shown below. We took a sample of those students and fed their Inputs and calculated the student performance and dropout rate.

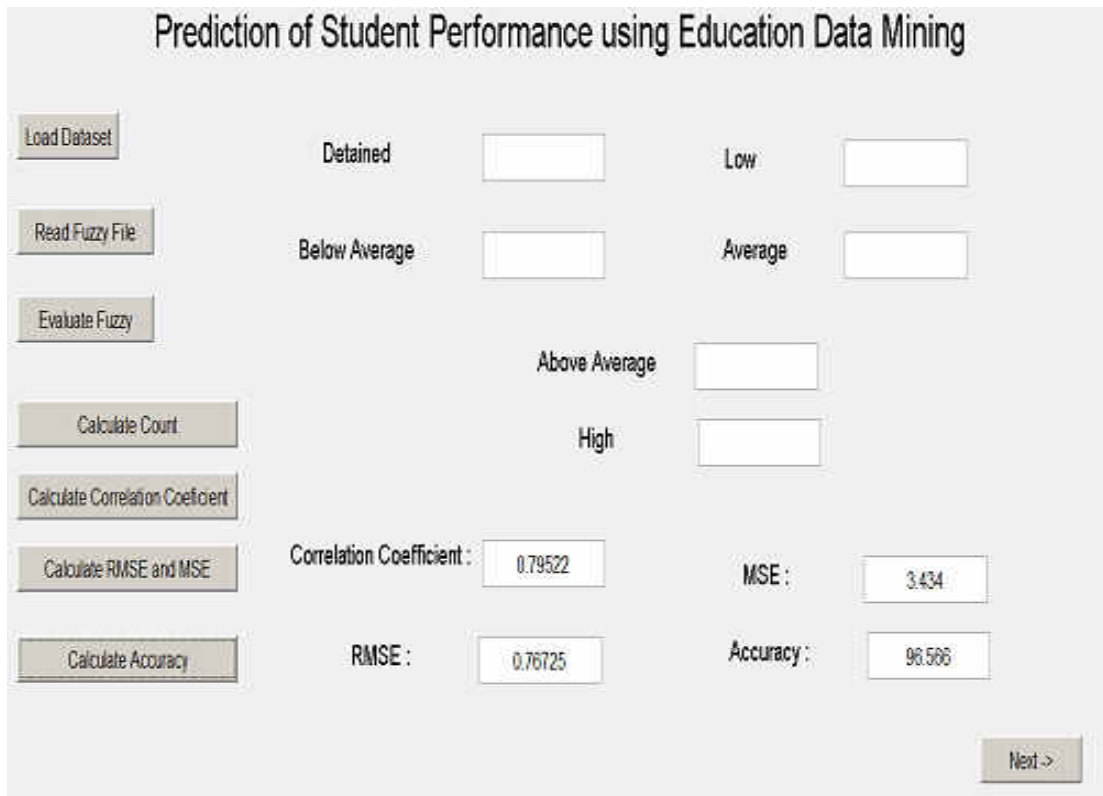


Figure 20: Accuracy,mse,rmse and correlation coefficient of student performance using centroid method

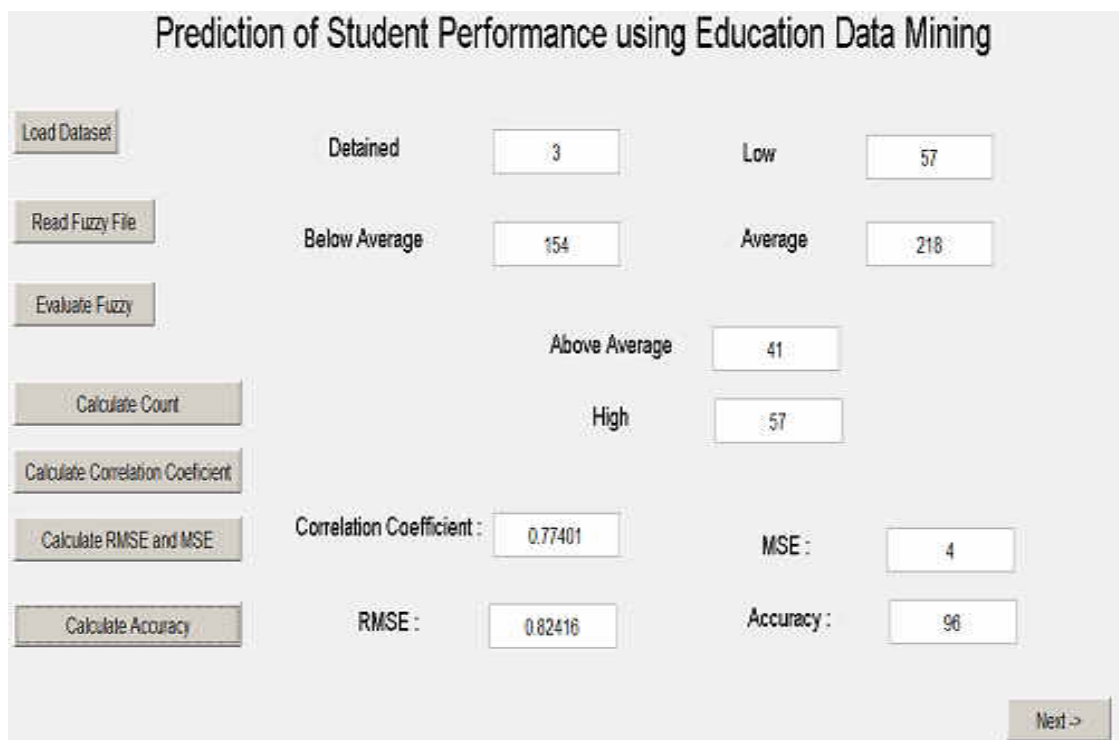


Figure 21: Accuracy,mse,rmse and correlation coefficient of performance using bisector method

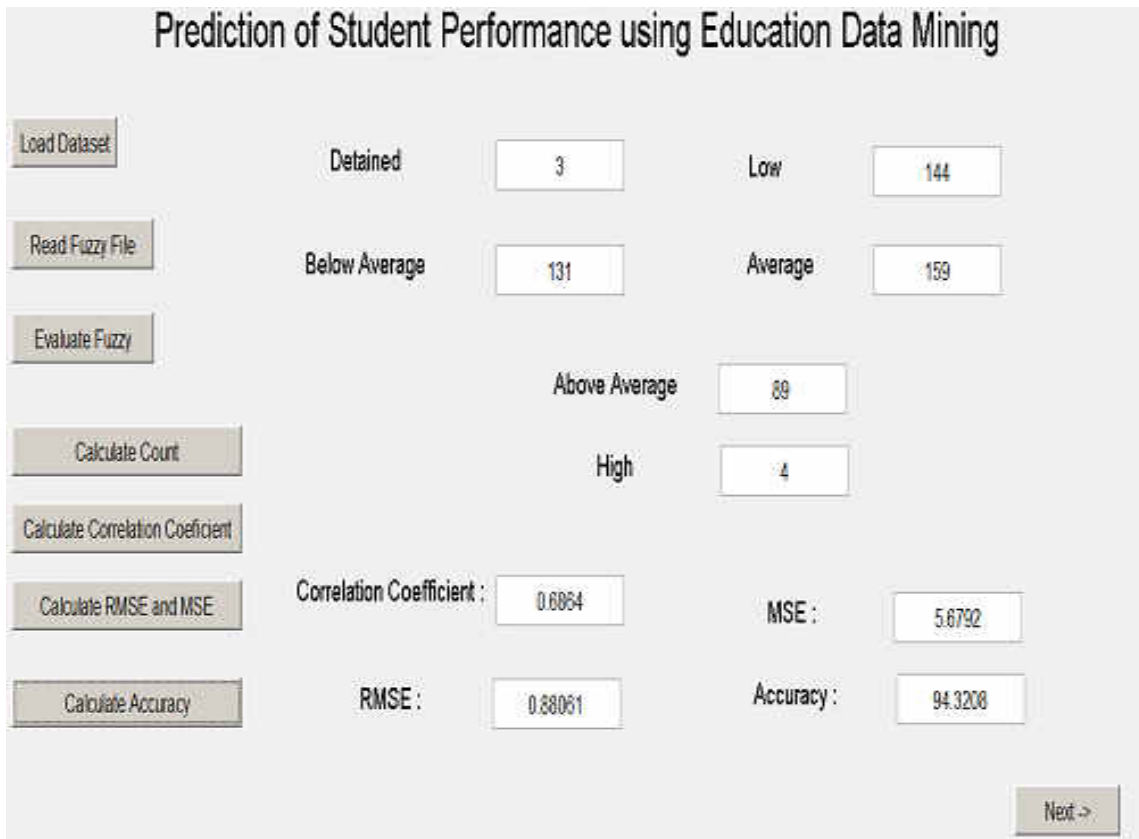


Figure 22: Accuracy,mse,rmse and correlation coefficient of performance using som method

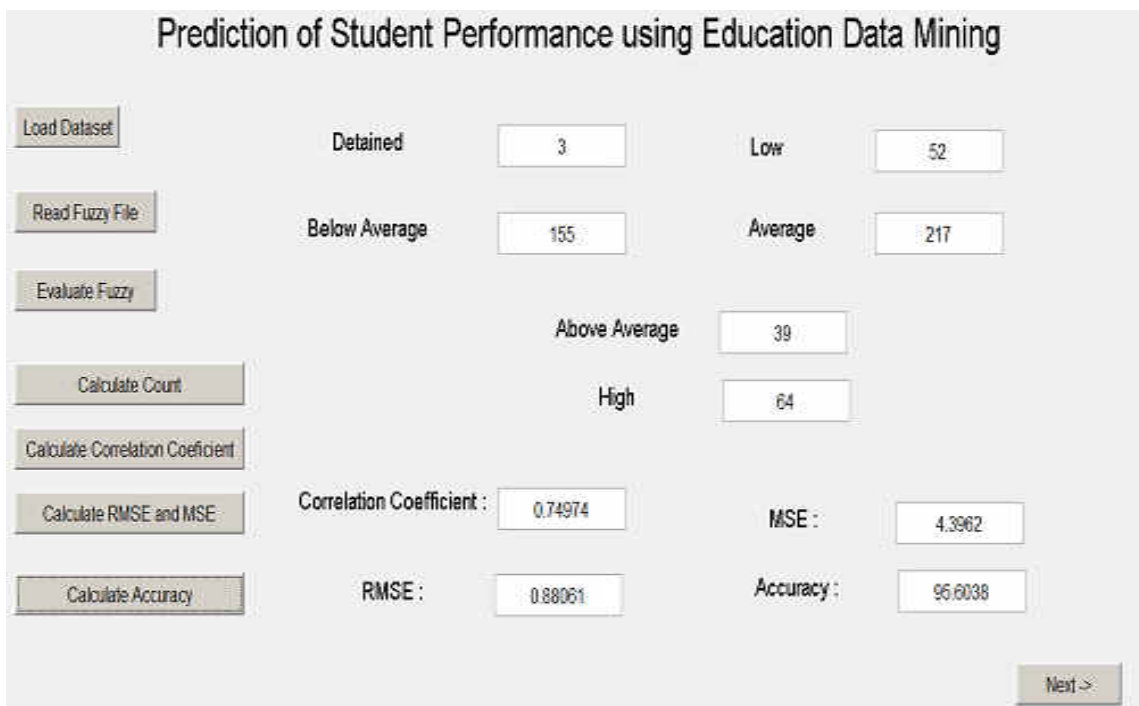


Figure 23 Accuracy,mse,rmse and correlation coefficient of performance using mom method

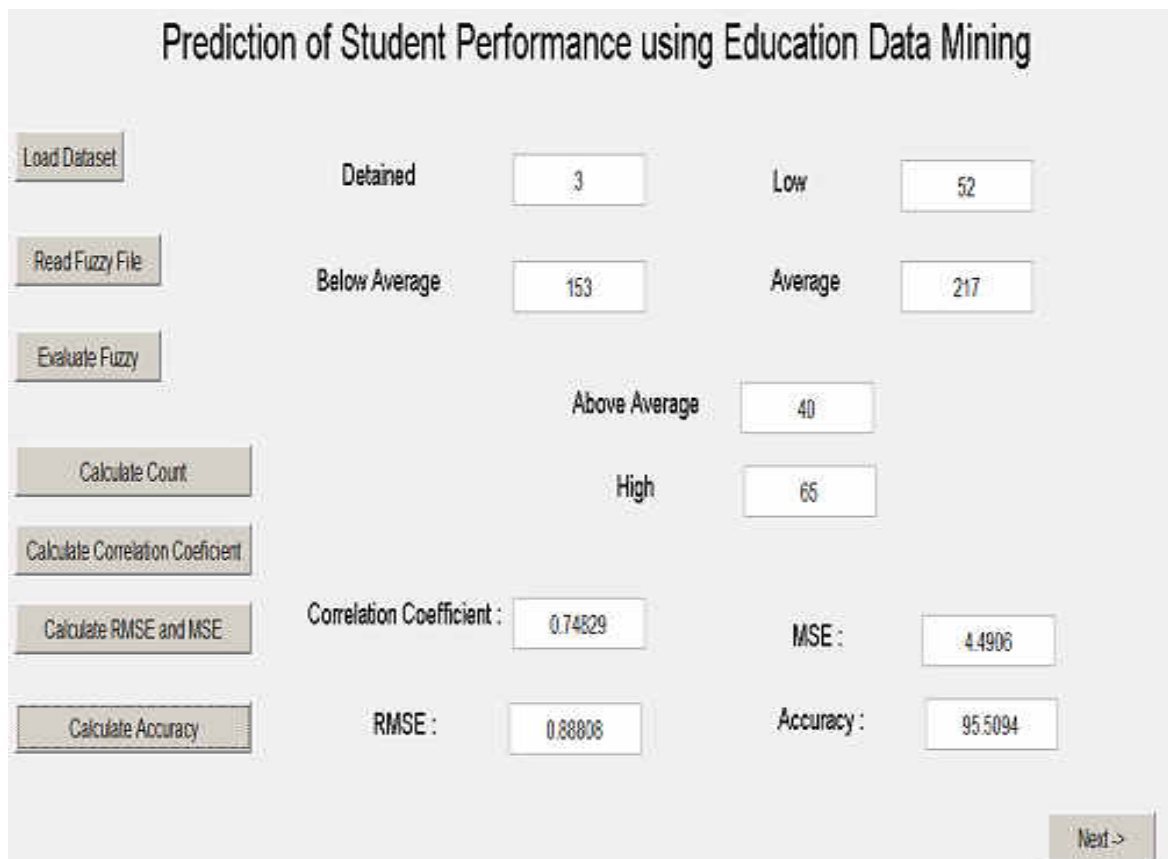


Figure 24: Accuracy,mse,rmse and correlation coefficient of performance using lom method

Method	Accuracy	MSE	RMSE	Correlation Coefficients
Bisector	96	4	0.8242	0.7740
Centroid	96.5660	3.4340	0.7673	0.7952
Mom	95.6038	4.3962	0.8806	0.7497
Lom	95.5094	4.4906	0.8881	0.7483
Som	94.3208	5.6792	0.8806	0.6864

Table 6: Comparative analysis of student performance by using different methods

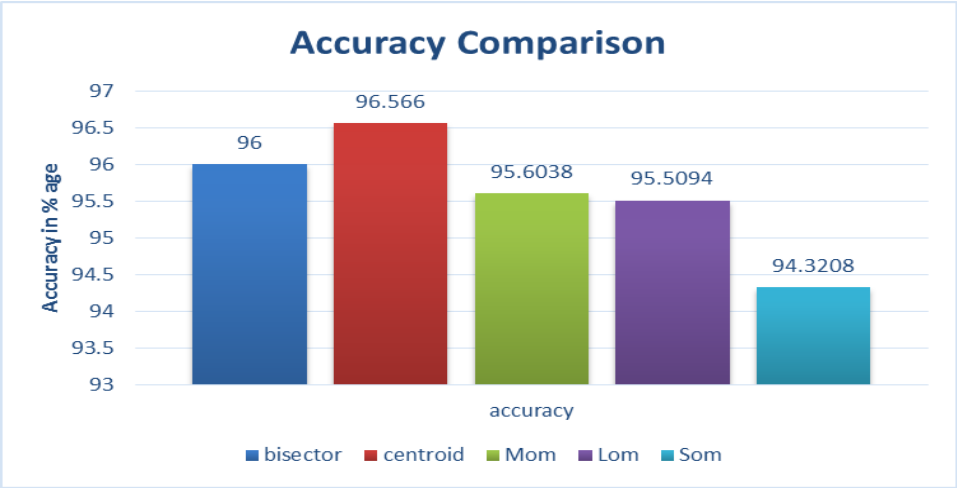


Figure 25:Accuracy comparison

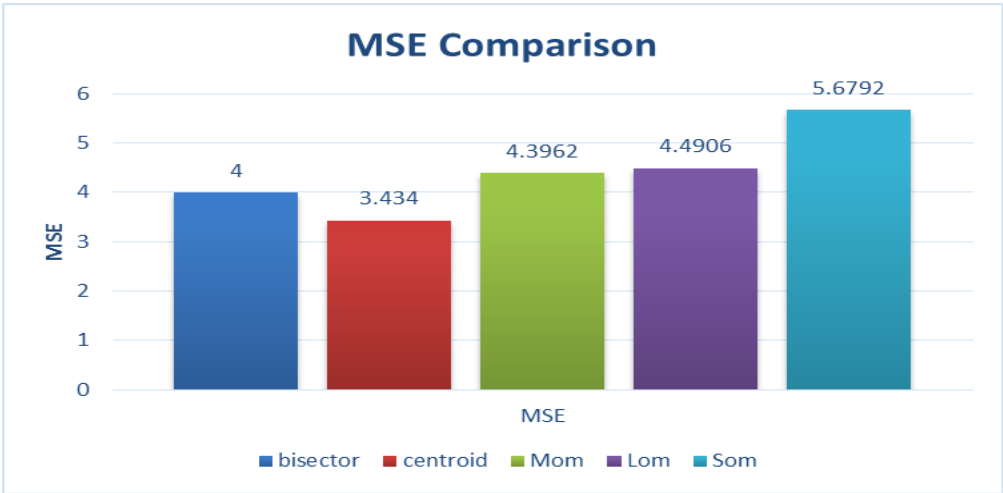


Figure 26:MSE comparison

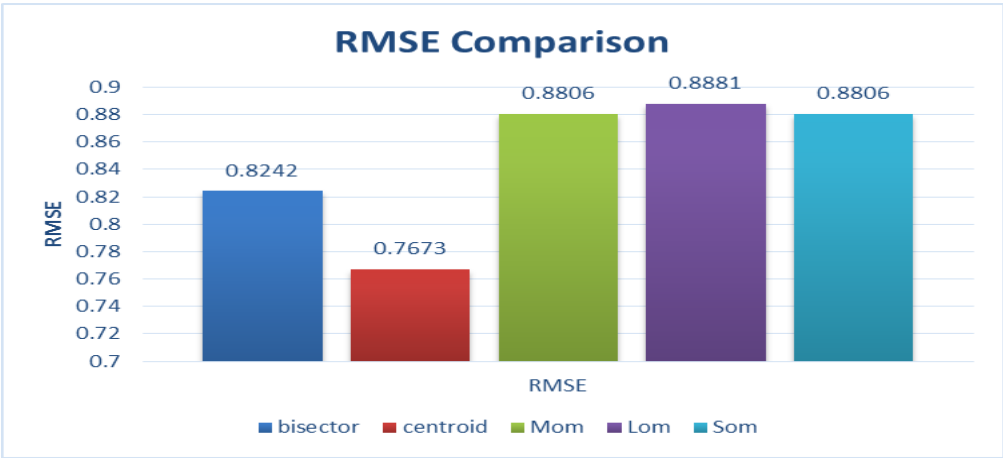


Figure 27: RMSE comparison

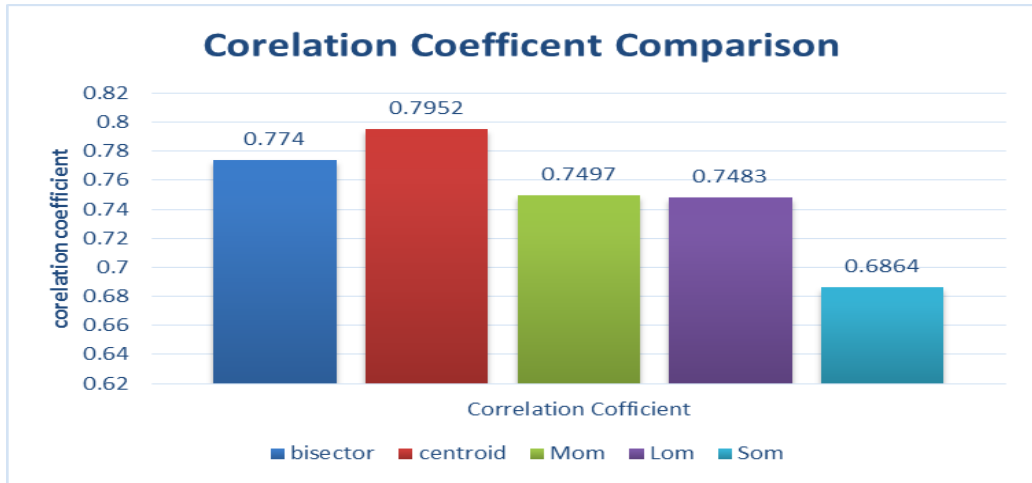


Figure 28: Correlation coficient comparison

Similarly, for dropout

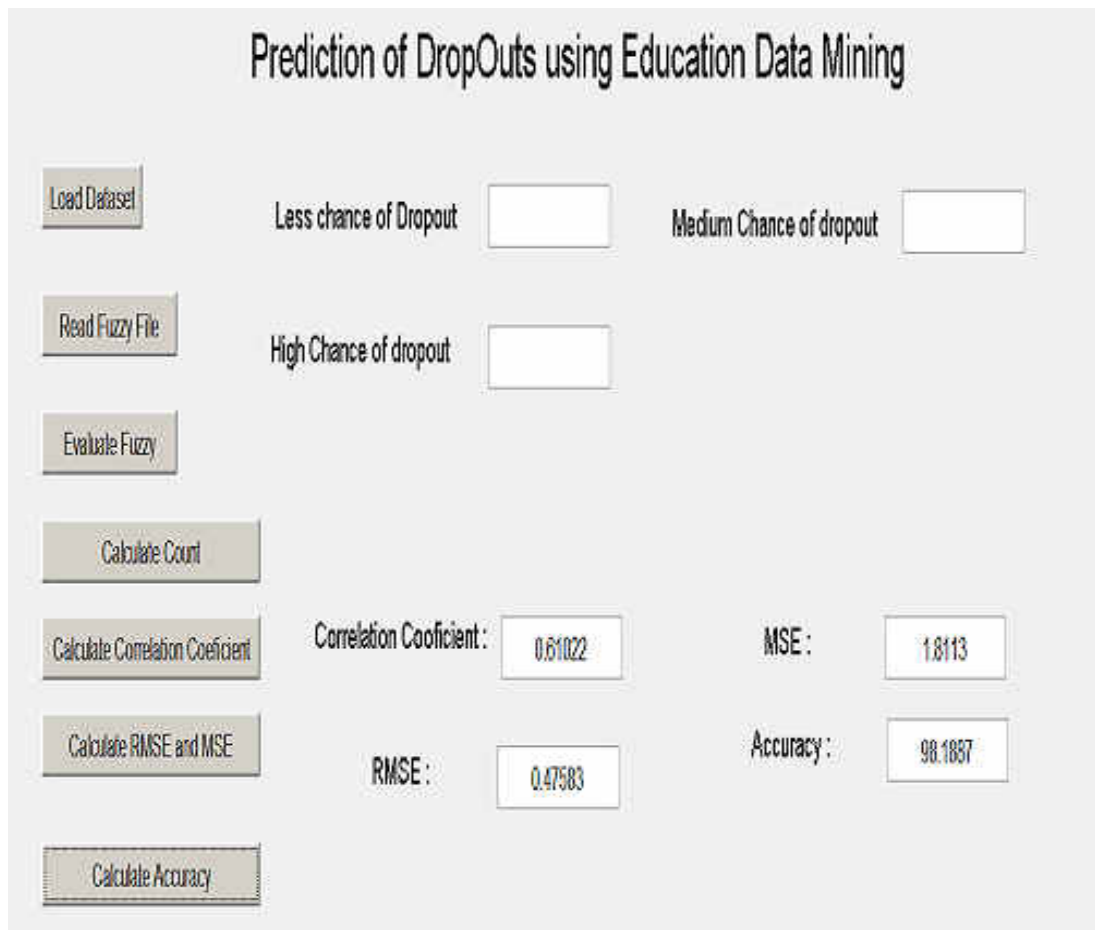


Figure 29: Accuracy,mse,rmse and correlation coefficient of dropout using centroid method

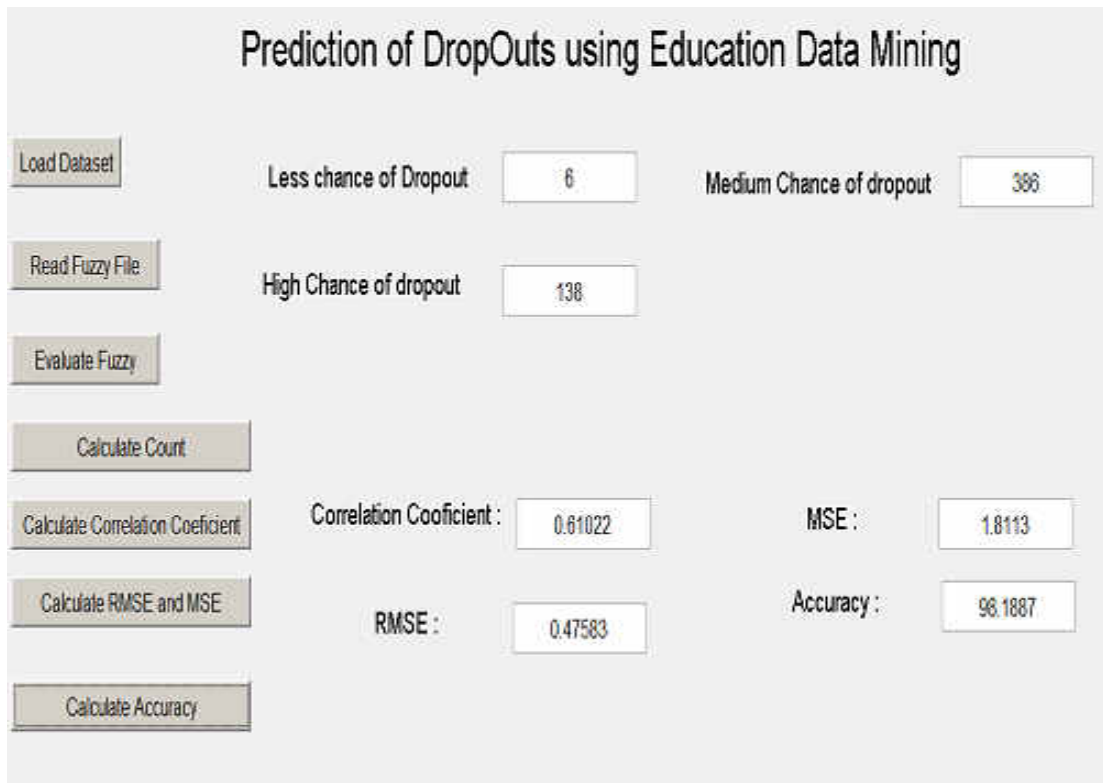


Figure 30: Accuracy,mse,rmse and correlation coefficient of dropout using bisection method

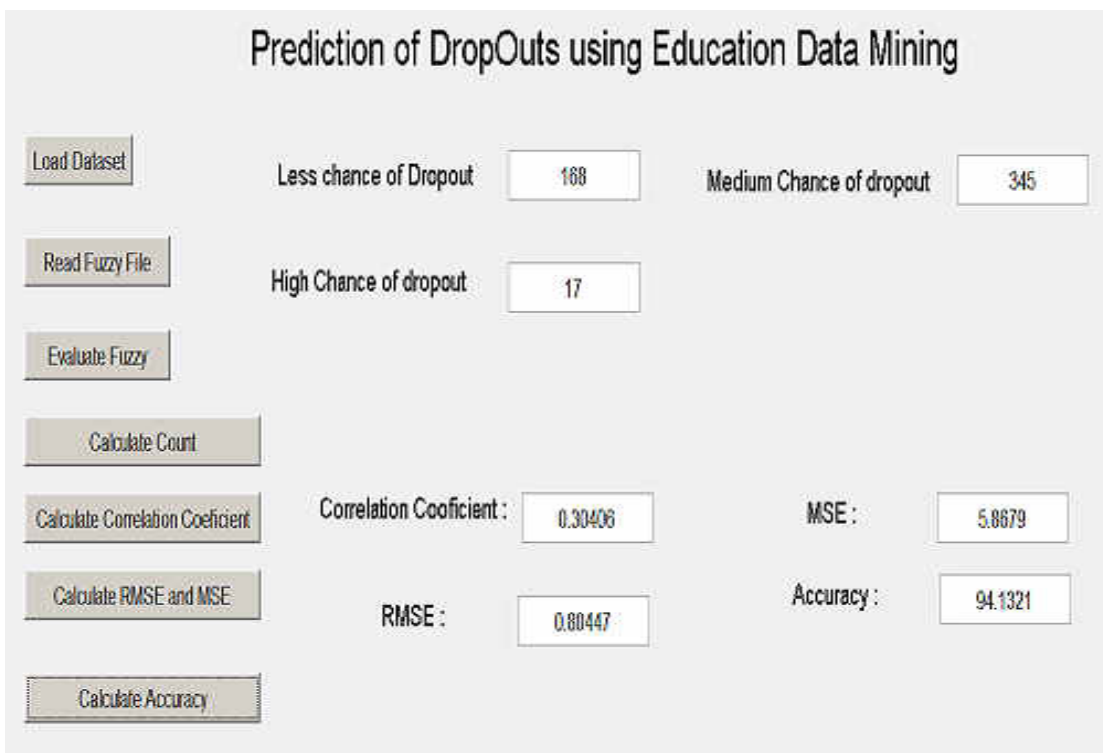


Figure 31: Accuracy,mse,rmse and correlation coefficient of dropout using som method

Method	Accuracy	MSE	RMSE	Correlation Coefficient
Bisector	98.1887	1.8113	0.4758	0.6102
Centroid	98.1887	1.8113	0.4758	0.6102
Mom	98.1887	1.8113	0.4758	0.6102
Lom	98.1887	1.8113	0.4758	0.6102
Som	94.1321	5.8679	0.8045	0.3041

Table 7: Comparative analysis of dropout by using different methods

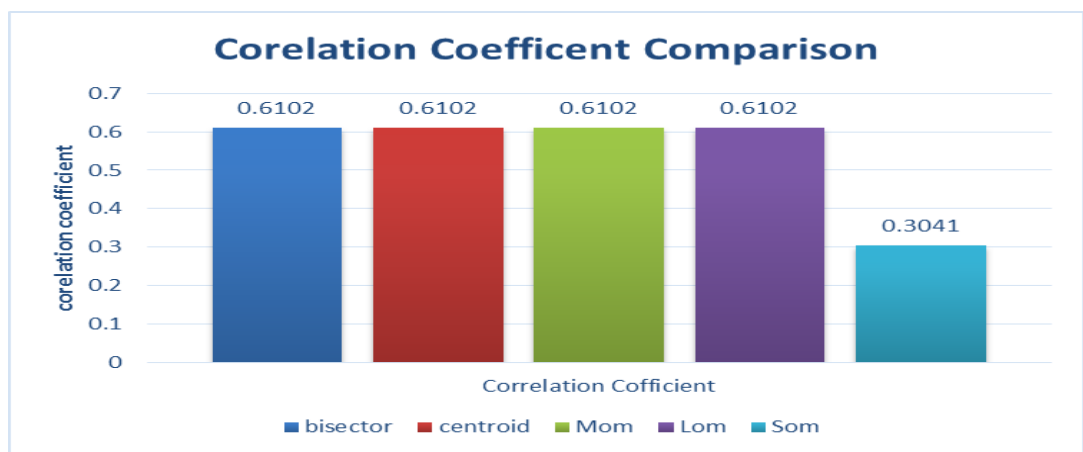


Figure 32: Correlation coefficient comparison

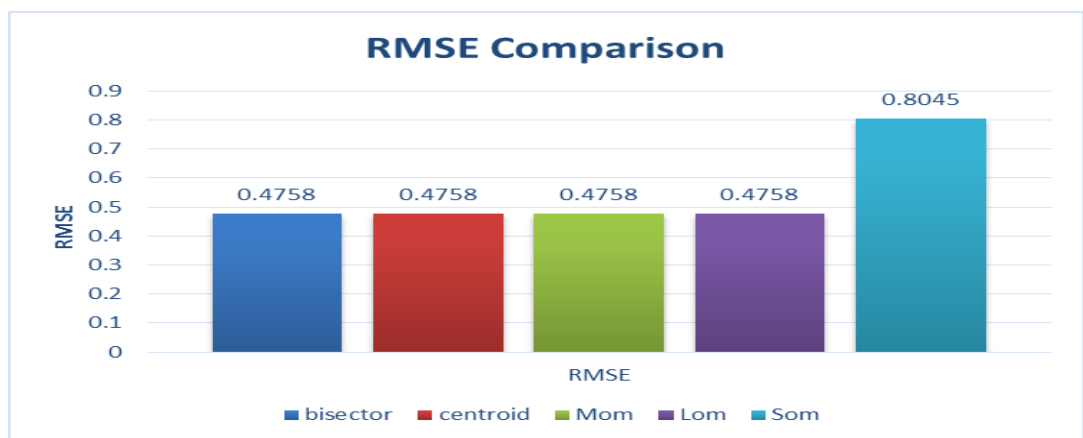


Figure 33: RMSE comparison

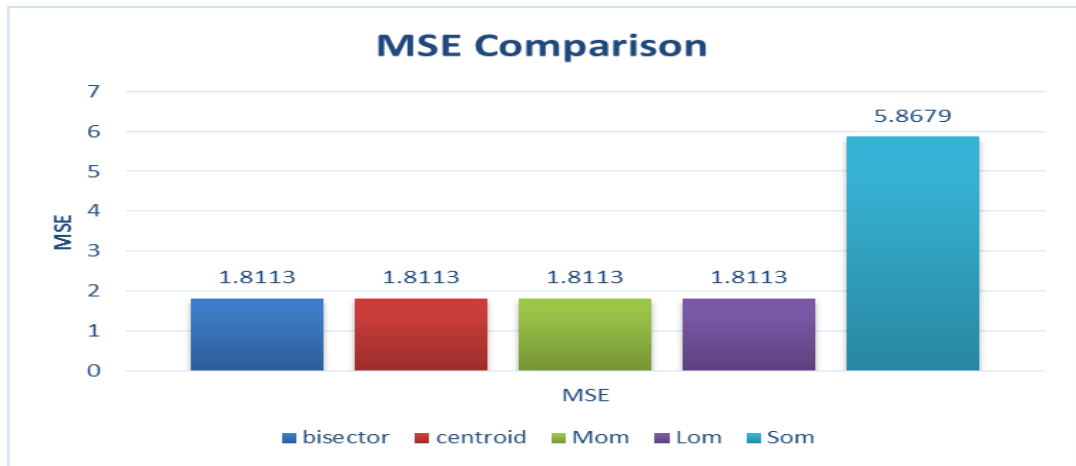


Figure 34: MSE comparison

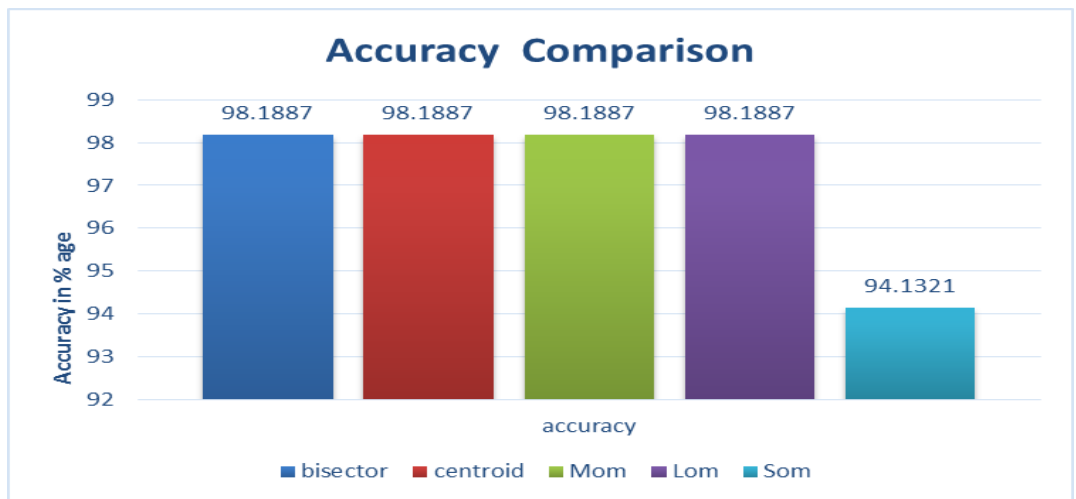


Figure 35: Accuracy comparison

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

For building the classification models from an input data set the most applicable approach is classification. Selecting the right data mining method or approach is another main fact to solve the problems. Fuzzy Logic is used for predicting the student performance and dropout rate. In this work, the fuzzy logic is applied corresponding to the student's future prediction and dropout rate and to improve the students' achievement and dropout rate and success more effectively in an efficient way using this technique. It could bring the benefits and impacts to students, educators and academic institutions. Determining the factors on which the performance of certain student depends, the performance of that student can be made better by working on the factors or attributes which are responsible for the downgrading of their achievement. In the current study, it was found that the student's performance is not totally dependent on their academic efforts, in spite; there are many other factors that have equal to greater influence as well. In conclusion, this study can motivate and help universities to perform data mining tasks on their students' data regularly to find out interesting results and patterns which can help both the university as well as the students in many ways. This will help universities not only to focus more on bright students but also to initially identify students with low academic achievement and find ways to support them.

5.2 FUTURE SCOPE

The Future scope of this study is as follows

- This system can also be extended to include more factors that may affect the performance of the students.
- We have worked on fuzzy logic to predict students' performance and dropout. In near future we will be testing our work using neural network as neural network proved to be good option as optimizer which can improve the results.

References

- [1] Book: Han Jiawei and Micheline Kamber. *Data Mining: Concepts and techniques*. Elsevier, 2011.
- [2] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. S. H. Y. Seong, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques," *2009 Second Int. Conf. Comput. Electr. Eng.*, vol. 2, pp. 357–361, 2009.
- [3] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," *2013 Fourth Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–7, 2013.
- [4] V. K. Deepa, "Rapid Development of Applications in Data Mining," pp. 1–4, 2013.
- [5] Y. Kurniawan and E. Halim, "Use Data Warehouse and Data Mining to Predict Student Academic Performance in Schools: A Case Study (Perspective Application and Benefits)," no. August, pp. 98–103, 2013.
- [6] K. V. Krishna Kishore, S. Venkatramaphanikumar, and S. Alekhya, "Prediction of student academic progression: A case study on Vignan University," *2014 Int. Conf. Comput. Commun. Informatics Ushering Technol. Tomorrow, Today, ICCCI 2014*, no. 2, pp. 3–8, 2014.
- [7] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," *Commun. Netw. Technol. (ICCNT), 2014 Int. Conf. Comput. Intell. Comput. Res.*, pp. 113–118, 2014.
- [8] N. T. Nghe, P. Janecek, and P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance," pp. 7–12, 2007.
- [9] P. Kaur, S. K. Singh, P. Agrawal, and L. Jain, "Fuzzy Rule Based Students' Performance Analysis Expert System."

- [10] A. S. Subaira and P. G. Scholar, "Efficient Classification Mechanism for Network Intrusion Detection System Based on Data Mining Techniques : a Survey," pp. 274–280, 2014.
- [11] K. Ashokkumar and V. Raghavendran, "A semantic Feedback on student performance using Data Mining Techniques" pp. 1605–1609, 2016.
- [12] Pooja Asopa "Evaluating Student Performance Using Fuzzy Inference System in Fuzzy ITS", Elsevier Science Pub. Co., 1985.
- [13] Upadhyaya, M.S., "Fuzzy Logic Based Evaluation of Performance of Students in Colleges," Journal of Computer Applications (JCA), ISSN: 0974-1925, Vol. 5, Issue 1, 2012.
- [14] Semerci, C., "The Influence Of Fuzzy Logic Theory On Students' Achievement," The Turkish Online Journal of Educational Technology- TOJET, ISSN: 1303-6521, Vol. 3, Issue 2, Article 9, April 2004.
- [15] Pavani, S.; Gangadhar, P.V.S.S.; Gulhare, K.K., " Evaluation Of Teacher's Performance Using Fuzzylogic Techniques," International Journal of Computer Trends and Technology, ISSN: 2231-2803, Vol. 3, Issue 2, 2012.
- [16] Khan. A. R.; Amin, H. U.; Rehman, Z. U., "Application of Expert System with Fuzzy Logic in Teachers' Performance Evaluation," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 2, February 2011.
- [17] Gobert, J. D., Y. J. Kim, et al. (2015). "Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld." *Thinking Skills and Creativity*.
- [18] Huang, S. and N. Fang (2013). "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models." *Computers & Education* 61: 133-145.
- [19] Huebner, R. A. (2013). "A survey of educational data mining research." *Research in Higher Education Journal*.

[20] Khan. A. R.; Amin, H. U.; Rehman, Z. U., "Application of Expert System with Fuzzy Logic in Teachers' Performance Evaluation," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 2, February 2011.

[21] Rawat, S. S., S. S. Reddy, et al. (2015). "Ascertaining the Factors Influencing Students' Performance for Engineering Pedagogy." Journal of Engineering Education Transformations: 30-33.