**International Journal of Swarm Intelligence Research**
Volume 11 • Issue 1 • January-March 2020

# Intelligent Techniques for Prediction of Engineering Colleges After XII

Mukta Goyal, Jaypee Institute of Information Technology, Uttar Pradesh, India
Rajalakshmi Krishnamurthi, Jaypee Institute of Information Technology, Uttar Pradesh, India
Gokul Gupta, Jaypee Institute of Information Technology, Uttar Pradesh, India
Abhishek Sharma, Jaypee Institute of Information Technology, Uttar Pradesh, India

## ABSTRACT

Today, students are very confused while selecting colleges based on their ranking after XII standard exam. If students are willing to go for engineering, then they are interested to know the name of colleges on the basis of their merit. The particular college depends on several factors. More and more colleges are interested in mapping students' other features such as extra-curricular activities and financial background, so that they can provide better platforms to sharpen their skills. Thus, this paper proposes an intelligent technique to provide students a platform that will help them to match the colleges based on their academics and extra-curricular qualifications. A fuzzy inference and weighted fuzzy decision tree are used to calculate the score of each student based on the multiple factors of the student where results are shown to be promising.

## KEYWORDS

Adaptive Neuro Fuzzy Inference System (ANFIS), Mamdaini Fuzzy Inference System, Prediction, Weighted Fuzzy Decision Tree

## INTRODUCTION

Every year, lakhs of students passes the XII class in India. There are only few thousand students who are sure about their future goals. Colleges are not only a place to study but also helps the student to overcome the issues arouse at the corporate world environment. A good college is one which not only has good academic environment but also excels in extracurricular activities. A good overall environment of colleges prepares the student for better future which also helps to improve the society. Thus, in the competitive world students are worried about their colleges after schooling to choose the appropriate college.

Predicting a college on scholastic basis was a history, with more and more involvement in extra-curricular activities these days students wants college which promotes their personality development and also gives them a platform to develop their skills. The basic idea behind this research is to predict the colleges using intelligent techniques. Intelligent system for college prediction requires is to study different types of students and choices of the student for opting the colleges. The selection of various

DOI: 10.4018/IJSIR.2020010102

colleges depends on the various parameters such as courses offered, college structure, fee structure etc. In this particular request and understudies generally pick those streams for designing in which more grounds arrangement happened. Because of interest of the understudies towards building, the opposition for admission to prime designing establishment has turned out to be exceptionally mind boggling. The problem in the process of a discovering knowledge from data, in the field of educational data mining, is to identify a representative set of data, so that a classification model will be constructed. There is a lack of data on extra-curricular activities of students. Today, more and more students are involved in extra-curricular activities, thus this parameter should also be taken into consideration for college prediction. Currently College Prediction Systems in India uses only scholastic scores to predict the college; whereas other factors such as family income, extra-curricular activities, fee structure etc. were ignored. Findings show that extracurricular activities also play a vital role in knowing and taking admission in the college Hence this research work proposes an intelligent technique for College Prediction is to specifically provide students a platform which will help them to form their profile and match it to the colleges based on their academics and extra-curricular qualifications.

Data mining techniques and machine learning techniques have been used to predict the performance of a student. This paper uses a Mamdani inference method to predict the college of a student basis of score of extra-curricular and scholastic achievements. A weighted fuzzy decision tree is also proposed to calculate the performance score. This score is used to help to predict the appropriate college to student. Survey work demonstrates that grouping is the effective technique among the current strategies. The data collected from the survey via google form. A statistical method is used to analyze the result of survey. To validate the result of weighted fuzzy decision tree, an adaptive neuro fuzzy approach is applied. Further sections explain the related work, methodology and result analysis's for prediction of appropriate college to the student.

## BACKGROUND

Data mining is used in various applications to identify unique patterns directions (Zhang, W., 2008; E.Venkatesan, S.Selvaragini, 2017). Traditional data mining and machine learning techniques may not be applied directly to extract the information or predicting the performance (Borkar, S., & Rajeswari, K, 2013). Now researcher have used fuzzy logic combination with deep learning neural either for prediction or classification(Guo, B., et al., 2015) . Applications such as educational data mining is an emerging discipline, concerned with data from academic field to develop various methods and to identify unique patterns helps to explore student's academic performance. Traditional multi-layer perceptron in neural networks method greatly suffers from substantial over fitting. The other two shallow models support vector machine and Naive Bayes are not capable to be comparably discriminative as SPPN. Some researchers have used three dimensional model, i.e. preference, fuzzy logic and influence to recommend the career oriented engineering stream These measurements are integrated with relative weighted set generated using Analysis Hierarchical Process (AHP) decision system to calculate the desire score of each student (Daud, A., et al., 2017) related to different career options. Personality, skills, Influence and trust measuring parameters are also used to recommend the career option using fuzzy logic inference system (Krishnamurthi, R., & Goyal,M., 2018). Influence computes the extent to which student is likely to be influenced by his seniors, friends, parents, and mentor choice for the target career path. To compute the friends, influence on a student, the trust between student and influential entities has to be computed based on their activities and social interactions. The greater the trust persist between student and his/her other friends, the Probability that the student gets influenced by them is of high value. If trusted friends give high priority to a career option, then student will also get positive about the same career option (Daud, A., et al., 2017).

Predicting the XII grade can provide a proper idea of a student that he/she would be able to achieve in life. A predictive model requires a data set that has the essential attributes to predict the future academic performance(Ahamed, A. S., Mahmood, N. T., & Rahman, R. M., 2017). Authors

have conducted a survey of 38 questionnaires, prepared with the help of experts. The questions were dependent on socio economic-demographic factors such as sex and age, fathers and mother's education, family income, place of education. It also depends on the Psychological factors such as, parent's involvement in education, time spent with friends, health status of a student etc. The academic factors such as also contribute the academic performance of student, weekly study time, extracurricular activity, accessibility of the internet, number of absences etc. Since career option is depend on the student's performance, it is important to predict their success. Existing methods have used features which are mostly related to academic performance, family income and family assets; while features belonging to family expenditures and students' personal information are usually ignored. Some of the authors have used support vector machine (SVM), C4.5, classification and regression and Bayes network o for feature set family expenditure, family income, personal information and family assets of students to predict the performance of student in recommendation of career after XII (Verma, P., Sood, S. K., & Kalra, S., 2017).

Classification algorithms such as decision tree ID3 and C4.5 have been used for the student performance evaluation system. Using these techniques, they extract the student's knowledge that describes students' performance in end semester examination. This helps to identify the dropout students and students who need special attention and allow instructor to provide appropriate advising counseling. Student trend and behavior is also analysed towards education using data mining techniques(Ogor, E. N., 2007; E.Venkatesan, S.Selvaragini 2017). Principal Component Analysis (PCA) are able to identify a subset of student activities that are relevant to build up an orthogonal space of representation and those activities that may support prediction of learner success (measured as final grades). This PCA algorithm works on students' active involvement and students' learning styles (Giovannella, C., Scaccia, F., & Popescu, E. 2013). Artificial swarm intelligence approaches is used to predict financial market (Rosenberg, L., Pescetelli, N., & Willcox, G., 2017). Intuitionistic fuzzy ant colony optimization technique is used for course sequencing (Agarwal, S. et.al., 2016).

College choice has become a complex science with the advent of information technology and the introduction of multiple college's development in current scenario. Students are often confused while choosing a college after XII. Many students choose their college without proper guidance. This paper proposes a recommendation of colleges based on scholastic performance, extracurricular activity and family income using Mamdani Inference.

Mamdani inference has a set of operations on system such as fuzzification (Mamdani, E. H., 1976; Sugeno, M., & Yasukawa, T.,1993). Fuzzification is the process of changing a real scalar value into a fuzzy value. This can be achieved with different kind of membership function. A membership function (MF) is a curve that defines how each point in the input space is mapped to a member-ship value (or degree of membership) between 0 and 1. After that rules are developed to obtain the fuzzy output. The rule-based form uses linguistic variables as its antecedents and consequents. The fuzzy rule-based system uses IF–THEN rule-based system, given by, IF antecedent, THEN consequent. This fuzzy output can be converted into crisp value using defuzzification process.

## METHODOLOGY

This section explains approaches for prediction of colleges. A fuzzy Mamdani inference, weighted fuzzy decision tree, and fuzzy ant colony optimization is used to predict the colleges. A survey is conducted among the students of various colleges to validate the prediction results. The dataset for prediction is collected from Kaggle machine learning repository. There were 638 respondents. Following are the steps:

### Preprocessing of Data

A principle component analysis (PCA) algorithm is used for Preprocessing of data set. In PCA algorithm, to reduce the number of features to represent the data, eigenvector and eigenvalues from the

covariance matrix or correlation matrix were obtained. After sorting the eigen values in descending order, eigenvectors were chosen in such a way that corresponds to the k largest eigen values. Here k is the number of dimensions of the new feature subspace (k≤d). A projection matrix W was selected from the k eigenvectors. A Mamdani fuzzy inference system is applied to classify this cleaned data set. Figure 1 shows the architecture of prediction of colleges.

## Mamdani Fuzzy Inference Approach for Prediction

A Mamdani fuzzy inference system is applied to classify the input values. The input variables which represent the number of features are scholastic performance, extra- curricular activities and family income. These input variables are also known as linguistic variables. Such linguistic variables are {Scholastic performance, Extra-curricular, Family Income}. These linguistic variables are defined into fuzzy variables such as:

Fuzzy variables of Scholastic performance: -- {very low, low, medium, high, very high}
Fuzzy variables of extra-curricular: -- {very low, low, medium, high, very high}
Fuzzy variables of family income: -- {low, average, high}

A triangular membership function is defined to represent each linguist variable. Figure 2 represent the membership function of scholastic linguistic variable for each fuzzy variable, whereas Figure 3 and Figure 4 represent the membership function for extracurricular activity involvement and family income of the student. Here x is the crisp value, maps to the fuzzy variable. A set of fuzzy rules is defined to process the fuzzify inputs to establish a rule strength. Rules are in the form of multiple antecedents and multiple consequent. The input antecedent represents the performance of a student whereas the consequent represents the attributes of the college. 75 rules have been developed to find out the output membership functions. Following are some of the rules:

**Figure 1. Architecture of prediction of colleges using Mamdani fuzzy inference system**

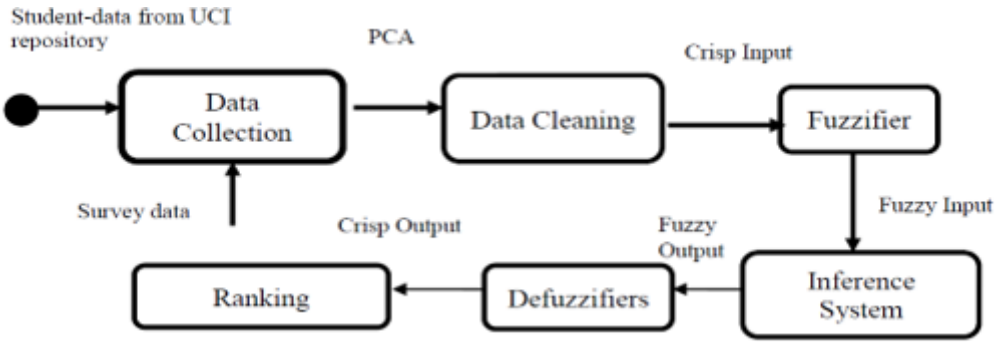



**Figure 2. Membership function for scholastic performance**

$$medium = \begin{cases} 0\ ; & x < 13.5 \text{ or } x > 5 \\ x - \frac{3.5}{1.5}; & 3.5 \le x < 5 \\ 1\ ; & 5 \le x < 5.5 \\ 7 - \frac{x}{1.5}; & 5.5 \le x < 7 \end{cases}, high = \begin{cases} 0\ ; & x < 5.5 \text{ or } x > 9 \\ x - \frac{5.5}{1.5}; & 5.5 \le x < 7 \\ 1\ ; & 7 \le x < 7.5 \\ 9 - \frac{x}{1.5}; & 7.5 \le x < 9 \end{cases}$$

**Figure 3. Membership function for extracurricular activity involvement**

$$very\ low = \begin{cases} 0\ ;\ x > 2 \\ x\ ; 0 \le x < 1 \\ 2 - x;\ \ 1 \le x < 2 \end{cases},\quad low = \begin{cases} 0\ ;\ x < 1.75\ and\ x > 4 \\ x - \frac{1.75}{1.15}; 1.75 \le x < 2.9 \\ 4 - x/1.1;\ \ 2.9 \le x < 4 \end{cases}$$

$$medium = \begin{cases} 0\ ;\ x < 3.75\ and\ x > 6 \\ x - \frac{3.75}{1.15};\ \ 3.75 \le x < 4.9 \\ 6 - x/1.1;\ \ 4.9 \le x < 6 \end{cases},\ high = \begin{cases} 0\ ;\ x < 5.75\ and\ x > 8 \\ x - \frac{5.75}{1.15};\ \ 5.75 \le x < 6.9 \\ 8 - x/1.1;\ \ 6.9 \le x < 8 \end{cases}$$

$$very\ high = \begin{cases} 0\ ;\ x < 7.75\ and\ x > 10 \\ x - \frac{7.75}{1.15};\ \ 7.75 \le x < 8.9 \\ 10 - x/1.1;\ \ 8.9 \le x < 10 \end{cases}$$

**Figure 4. Membership function for family income of the student**

$$low = \begin{cases} 0\ ;\ x > 2.5 \\ 1\ ;\ x < 1.5 \\ 2.5 - x;\ \ 1.5 \le x < 2.5 \end{cases} \quad average = \begin{cases} 0\ ;\ x < 1.5\ and\ x > 8.5 \\ x - \frac{1.5}{3.5}\ ;\ \ 1.5 \le x < 5 \\ 8.5 - x/3.5;\ \ 5 \le x < 8.5 \end{cases}$$

$$high = \begin{cases} 0\ ;\ x < 7 \\ 1\ ;\ x < 8.5 \\ x - 7/1.5;\ \ 7 \le x < 8.5 \end{cases}$$

**Rule 1:** If (scholastic performance) is very low or (extra-curricular) is very low and (family income) is low then(studies) is poor and (extra- curricular) is low and (fee structure) is low.

**Rule 2:** If (scholastic performance) is low or (extracurricular) is very low and (family income) is low then(studies) is poor and (extra-curricular) is low and (fee structure) is low.

**Rule 3:** If (scholastic performance) is medium or (extra-curricular) is very low and (family income) is average then(studies) is average and (extra- curricular) is low and (fee structure) is low.

**Rule 4:** If (scholastic performance) is high or (extra-curricular) is very low and (family income) is low then(studies) is good and (extra- curricular) is low and (fee structure) is low.

**Rule 5:** If (scholastic performance) is very high or (extra-curricular) is medium and (family income) is average then(studies) is good and (extra -curricular) is medium and (fee structure) is average.

**Rule 6:** If (scholastic performance) is high or (extra-curricular) is high and (family income) is low then(studies) is good and (extra -curricular) is high and (fee structure) is low.

Defuzzification is the process through which aggregated fuzzy values are converted into crisp value. There are several methods to defuzzify the aggregated output. Here centre of gravity method is applied to find the crisp value. The grading of the colleges is assigned by the government through ranking schemes such as NIRF. Thus, the crisp value represents the score of the student which is mapped to the college accordingly. Next section describes another approach i.e weighted fuzzy decision tree to predict the colleges.

## Weighted Fuzzy Decision Tree Approach

This section explains the weighted fuzzy decision tree algorithm. A student data set is collected through google form. The data set contain various attributes such as Name, Age, college Name, XII board percentage, extra-Curricular: Technical, extra-Curricular: Sports,extra-Curricular: Literary, extra-Curricular: Performing Art, extra-Curricular: Visual Art, Family annual Income. First step is

to calculate the weight using decision tree. In second step an algorithm weighted fuzzy decision tree is proposed for prediction. The student attributes which are considered are name, age, college name, percentage of 12th, extracurricular activities such as technical, sports, literacy, performing arts, visual art and family annual income. In this method first step is to calculate the weight. Here the weight is calculated using decision tree.

### *Algorithm for Calculating the Weight*

Figure 5a shows the algorithm for calculating weight.

After calculating the weight of the attributes next step is to find out the fuzzy values of the collected data set. A fuzzy inference system is applied to this data set and result of FIS is stored in .CSV file. Above calculate weight of the attributes is used in fuzzy decision tree algorithm. Next section describes the weighted fuzzy decision tree algorithm.

### *Proposed Weighted Fuzzy Decision Tree for Prediction*

Figure 5b explains the proposed weighted fuzzy decision tree. The score achieved through weighted fuzzy decision trees are mapped to the ranked colleges. Next section explains the validation of decision by using ANFIS algorithm, represented in Figure 6.

### *Adaptive Neuro Fuzzy Inference System Algorithm (ANFIS)*

Figure 6 shows the algorithm for validation. Different training –testing pairs have been generated to validate the results as shown in result analysis. Next section explains the prediction of colleges using fuzzy ant colony optimization.

### *Fuzzy Ant Colony Optimization*

Two of the main techniques of the swarm intelligence are Particle swarm optimization and Ant colony optimization. The Ant colony optimization algorithm is inspired by the behavior of ants to search the shortest path. In the proposed framework, students are going to act as ants and there would be a database maintained for the number of students passing through any of the paths. The path is basically the formation of the interconnected edges between the nodes. Each

**Figure 5a. Calculation of weight using decision tree**

***Algorithm-1***

- Creating a decision tree.
  1) Find target attribute.
  2) Calculates entropy for all the attributes and target attribute.
     dataEntropy += (-freq/len(data)) * math.log(freq/len(data), 2) - (1)
  3) Calculates the information gain for all attribute except the target attribute.
     Info_gain = chosen attribute entropy – target attribute entropy - (2)
     a. Chose the attribute with maximum information gain to split the dataset.
     b. Repeat step 2 and 3
        Until entropy(value) = attributes(value) or Maxheight(tree)
     c. Assign the information gain to calculate the weight of the attribute

$$W_{attribute} = \frac{\sum_{i=1}^{n} IG_{attribute}}{IG_{attribute}} \qquad (3)$$

**Figure 5b. Weighted Fuzzy Decision Tree**

***Algorithm -2***

- Divide the collected dataset into two parts i.e. training and testing data.
- Find the entropy of target attribute.
  Data Entropy += (-freq/len(data)) * math.log(freq/len(data), 2)
- Calculate the entropy of rest of the attributes
- Calculate Information gain (IG) for each attribute with respect to target attribute.
  Info gain = chosen attribute entropy – target attribute entropy
- These IG of each non targeted attribute, find the ratio among them. These ratio are used as a weight for attributes that are result of FIS.

$$F.Score = \sum_{i=1}^{n} Xi * W_{marks} + Yi * W_{ec} + Zi * W_{income} \quad -(3)$$

Where $W_{marks}$= weight of board marks
$W_{ec}$ =weight of extra-curricular activities
$W_{income}$ =weight of family income
F.Score = prediction score of student for college
Assign grades according to the F.Score using which student can get an idea about the colleges they may get.

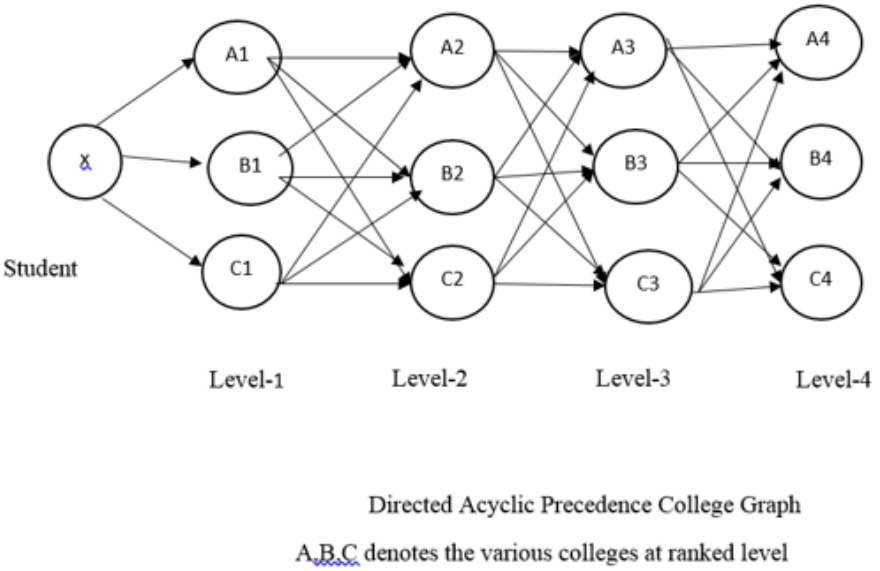
**Figure 6. ANFIS algorithm**

***Algorithm -3***

- Three 3 inputs x,y and z, and one output p are assigned.
- Layer 1 : Adaptive nodes with node function –
  $O1,i = \mu Ai(x)$ , for i=1,2
  O1 is membership grade of fuzzy set A
- Layer 2 : Every node is labeled ANFIS. Output is product of all incoming signals. $O2,I = w,i = \mu Ai(x) * \mu Bi(y)$, I = 1,2
- Layer 3 : Every node is fixed labeled N. Ith node calculates rule's firing strength. $O3,I = inv(wi) = wi / (w1 + w2)$ ; i=1,2
- Layer 4 : Output node as summation of all input signals
  Overall output = o4,I = Sum[inv(wi)*fi] = Sum(wi*fi)/Sum(wi)

node connects to every other node present in the next level as shown in Figure 7. Each level represents the number of colleges according to their rank. Let first level represent the AAA ranking colleges, second level represent the AA ranked colleges, third level represent the A ranked colleges and fourth level represent the BB ranked colleges. The membership value of each feature is calculated from Figures 2, 3 and 4. A Mamdani inference is applied to aggregate the final score of each student on the basis of this feature of the student. Each edge gets its edge weight determined whenever any student traverses that edge. The pheromone or the edge weight will be updated according to the new edge weight:

**Figure 7. College precedence graph using ant colony optimization**




$$PheromoneUpdation = \frac{\left(Edge\,weight\,old * count\right) + Edge\,weight\,new}{count + 1}$$

Here count refers to the number of students that have already traversed that path. The student performance is predicted at each level among the other students. The updated pheromone determines the change in the edge weight as the student performance are predicted at that level. The score of both the levels, current and previous, is considered in the edge weight update since the colleges are sequence in such a way that the levels are interdependent.

### *Performance Evaluation*

For performance evaluation, three standard evaluation metrics (precision, recall and F1-score) are used. 5-fold cross validation is used for comparison with baseline methods. These performance evaluation parameters are defined as:

$$\text{Precision} = \text{TP TP} + \text{FP} \tag{4}$$

$$\text{Recall} = \text{sensitivity} = \text{TP (TP} + \text{FN)} \tag{5}$$

$$\text{F1 score} = 2.\ \text{precision.recall precision} + \text{recall} \tag{6}$$

To evaluate the performance of the algorithm confusion matrix is evaluated. Results are discussed in the next section.

## PERFORMANCE ANALYSIS

This section explains the processing of collected and analysis of result using both the approaches. In approach- 1, Mamdani inference is used to predict the colleges for the student and statistical F- test is

used to compare the hypothesis. In approach – 2, weighted fuzzy decision tree is proposed to predict the colleges for the student. An adaptive neuro fuzzy algorithm is used to validate the prediction.

## Mamdani Inference Method

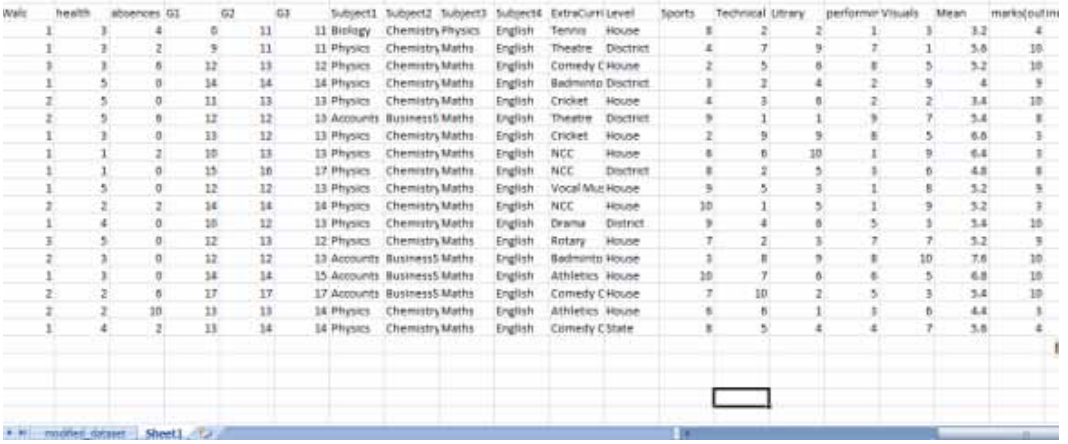
This section shows the result analysis of approach -1. Next section explains the data processing.

### *Data Cleaning*

Data is collected from https://www.kaggle.com/rmalshe/studentperformance-prediction/data. The data set contained many attributes, where only the relevant attributes has been considered whereas rest were ignored. A new attribute "Extracurricular" is added to the data set and values has been generated using pseudo random function. Figure 8 shows the snapshot of the collected data. A python language is used to implement the principal component analysis (PCA) algorithm. This considers the important attributes which dominates the result of prediction.

Figure 8 shows the snapshot of collected data with some attributes. There were 638 respondents for classification. Figure 9 shows the data having null values. Figure 10 shows the substituted values corresponding to zero after cleaning at column G2 and G3. An average is taken to assign the values in lieu of zero value. A principal component analysis (PCA) algorithm is applied to consider the important attribute. A principal component analysis algorithm is based on the variance-covariance matrix or the correlation matrix where the principal component score is used for further analyses.

**Figure 8. Snapshot of the collected data with various attributes**



| Walc | health | absences | G1 | G2 | G3 | Subject1 | Subject2 | Subject3 | Subject4 | ExtraCurri | Level | Sports | Technical | Utrary | performa | Visuals | Mean | marks(outo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 4 | 0 | 11 | 11 | Biology | Chemistry | Physics | English | Tennis | House | 8 | 2 | 2 | 1 | 3 | 3.2 | 4 |
| 1 | 3 | 2 | 9 | 11 | 11 | Physics | Chemistry | Maths | English | Theatre | Disctrict | 4 | 7 | 9 | 7 | 1 | 5.6 | 10 |
| 3 | 3 | 6 | 12 | 13 | 12 | Physics | Chemistry | Maths | English | Comedy C | House | 2 | 5 | 6 | 8 | 5 | 5.2 | 10 |
| 1 | 5 | 0 | 14 | 14 | 14 | Physics | Chemistry | Maths | English | Badminto | Disctrict | 3 | 2 | 4 | 2 | 9 | 4 | 9 |
| 2 | 5 | 0 | 11 | 13 | 13 | Physics | Chemistry | Maths | English | Cricket | House | 4 | 3 | 8 | 2 | 2 | 3.4 | 10 |
| 2 | 5 | 8 | 12 | 12 | 13 | Accounts | Business5 | Maths | English | Theatre | Disctrict | 9 | 1 | 1 | 9 | 7 | 5.4 | 8 |
| 1 | 3 | 0 | 13 | 12 | 13 | Physics | Chemistry | Maths | English | Cricket | House | 2 | 9 | 9 | 8 | 5 | 6.6 | 3 |
| 1 | 1 | 2 | 10 | 13 | 13 | Physics | Chemistry | Maths | English | NCC | House | 8 | 8 | 10 | 1 | 9 | 6.4 | 3 |
| 1 | 1 | 0 | 15 | 16 | 17 | Physics | Chemistry | Maths | English | NCC | Disctrict | 8 | 2 | 5 | 3 | 6 | 4.8 | 8 |
| 1 | 5 | 0 | 12 | 12 | 13 | Physics | Chemistry | Maths | English | Vocal Mus | House | 9 | 5 | 3 | 1 | 8 | 5.2 | 9 |
| 2 | 2 | 2 | 14 | 14 | 14 | Physics | Chemistry | Maths | English | NCC | House | 10 | 1 | 5 | 1 | 9 | 5.2 | 3 |
| 1 | 4 | 0 | 10 | 12 | 13 | Physics | Chemistry | Maths | English | Drama | District | 9 | 4 | 8 | 5 | 3 | 5.4 | 10 |
| 3 | 5 | 0 | 12 | 13 | 12 | Physics | Chemistry | Maths | English | Rotary | House | 7 | 2 | 3 | 7 | 7 | 5.2 | 9 |
| 2 | 3 | 0 | 12 | 12 | 13 | Accounts | Business5 | Maths | English | Badminto | House | 3 | 8 | 9 | 8 | 10 | 7.6 | 10 |
| 1 | 3 | 0 | 14 | 14 | 15 | Accounts | Business5 | Maths | English | Athletics | House | 10 | 7 | 6 | 6 | 5 | 6.8 | 10 |
| 2 | 2 | 6 | 17 | 17 | 17 | Accounts | Business5 | Maths | English | Comedy C | House | 7 | 10 | 2 | 5 | 3 | 5.4 | 10 |
| 2 | 2 | 10 | 13 | 13 | 14 | Physics | Chemistry | Maths | English | Athletics | House | 6 | 6 | 1 | 3 | 6 | 4.4 | 3 |
| 1 | 4 | 2 | 13 | 14 | 14 | Physics | Chemistry | Maths | English | Comedy C | State | 8 | 5 | 4 | 4 | 7 | 5.6 | 4 |

**Figure 9. Data with null values**

| tudytime | failures | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 95.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 | 395.000000 |
| .035443 | 0.334177 | 3.944304 | 3.235443 | 3.108861 | 1.481013 | 2.291139 | 3.554430 | 5.708861 | 10.908861 | 10.713924 | 10.415190 |
| .839240 | 0.743651 | 0.896659 | 0.998862 | 1.113278 | 0.890741 | 1.287897 | 1.390303 | 8.003096 | 3.319195 | 3.761505 | 4.581443 |
| .000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 | 3.000000 | 0.000000 | 0.000000 |
| .000000 | 0.000000 | 4.000000 | 3.000000 | 2.000000 | 1.000000 | 1.000000 | 3.000000 | 0.000000 | 8.000000 | 9.000000 | 8.000000 |
| .000000 | 0.000000 | 4.000000 | 3.000000 | 3.000000 | 1.000000 | 2.000000 | 4.000000 | 4.000000 | 11.000000 | 11.000000 | 11.000000 |
| .000000 | 0.000000 | 5.000000 | 4.000000 | 4.000000 | 2.000000 | 3.000000 | 5.000000 | 8.000000 | 13.000000 | 13.000000 | 14.000000 |
| .000000 | 3.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 75.000000 | 19.000000 | 19.000000 | 20.000000 |

**Figure 10. Data after cleaning**

| tudytime | failures | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 57.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 | 357.000000 |
| .042017 | 0.271709 | 3.955182 | 3.246499 | 3.098039 | 1.495798 | 2.330532 | 3.549020 | 6.316527 | 11.268908 | 11.358543 | 11.523810 |
| .831895 | 0.671750 | 0.885721 | 1.011601 | 1.090779 | 0.919886 | 1.294974 | 1.402638 | 8.187623 | 3.240450 | 3.147188 | 3.227797 |
| .000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 | 3.000000 | 5.000000 | 4.000000 |
| .000000 | 0.000000 | 4.000000 | 3.000000 | 2.000000 | 1.000000 | 1.000000 | 3.000000 | 2.000000 | 9.000000 | 9.000000 | 9.000000 |
| .000000 | 0.000000 | 4.000000 | 3.000000 | 3.000000 | 1.000000 | 2.000000 | 4.000000 | 4.000000 | 11.000000 | 11.000000 | 11.000000 |
| .000000 | 0.000000 | 5.000000 | 4.000000 | 4.000000 | 2.000000 | 3.000000 | 5.000000 | 8.000000 | 14.000000 | 14.000000 | 14.000000 |
| .000000 | 3.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 75.000000 | 19.000000 | 19.000000 | 20.000000 |

The attributes which are considered as important are Scholastic performance, Extra-curricular and Family Income.

### *Analysis of Inference Method*

A Mamdani Fuzzy inference system is applied on preprocessed data set for further processing using MATLAB. Figures 11, 12, and 13 show the graph for input variables income, Scholastic

**Figure 11. FIS input (Income)**

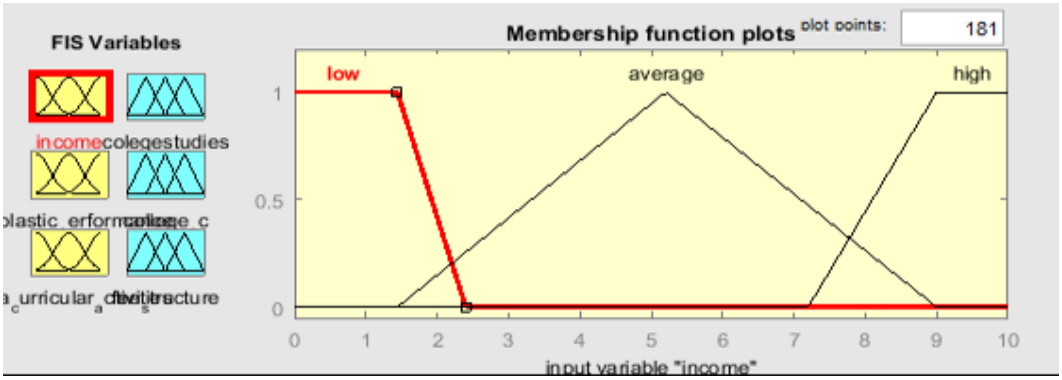



**Figure 12. FIS input (Scholastic performance)**

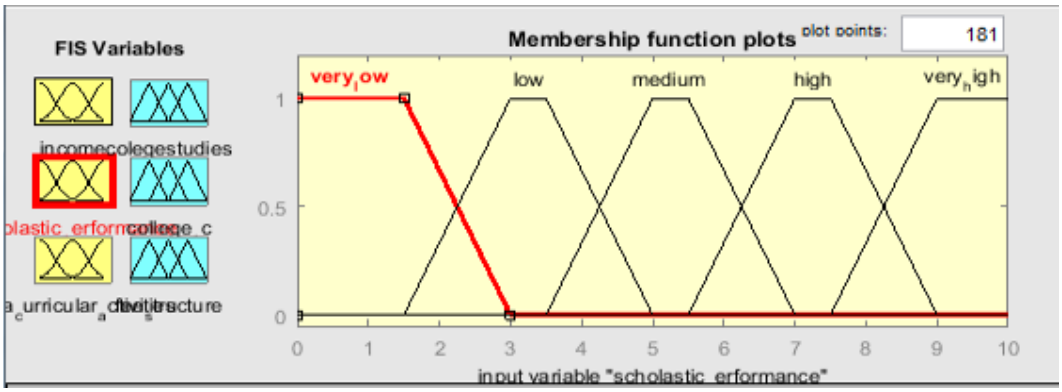
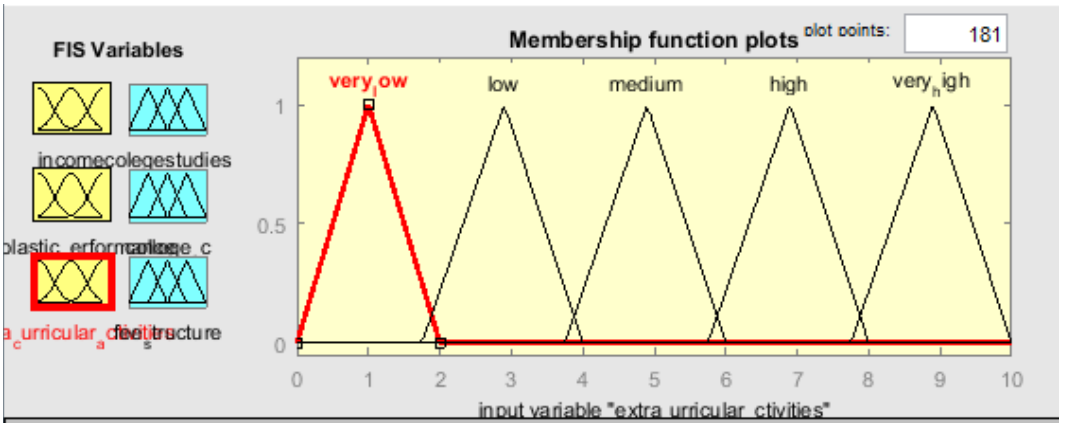


**Figure 13. FIS input (Extra-curricular activities)**




performance, extra-curricular activities respectively. Figures 14, 15, and 16 show the graph for output variables fee structure, college studies and college respectively. According to the rule if the student scholastic performance is very low, extra-curricular activity is very low and family income is low then recommended college studies will not very good, college fee will be low and the college will be low category. Figure 17 shows the defuzzified output. The output distribution in Figure 17 shows after combining the output rules.

FIS Model

- **Input membership function:** See Figure 11-13:

Linguistic variables: -- {Scholastic_performance, Extra-curricular_activites, Income}
Linguistic values of Scholastic performance: -- {very low, low, medium, high, very high}
Linguistic values of extra-curricular: -- {very low, low, medium, high, very high}
Linguistic values of family income: -- {low, average, high}

**Figure 14. FIS output (Fee structure)**

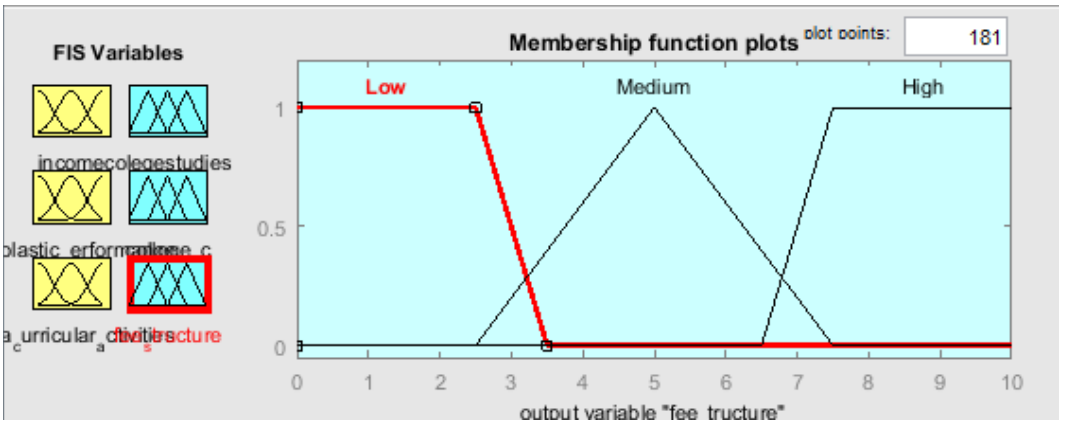


**Figure 15. FIS output (College studies)**

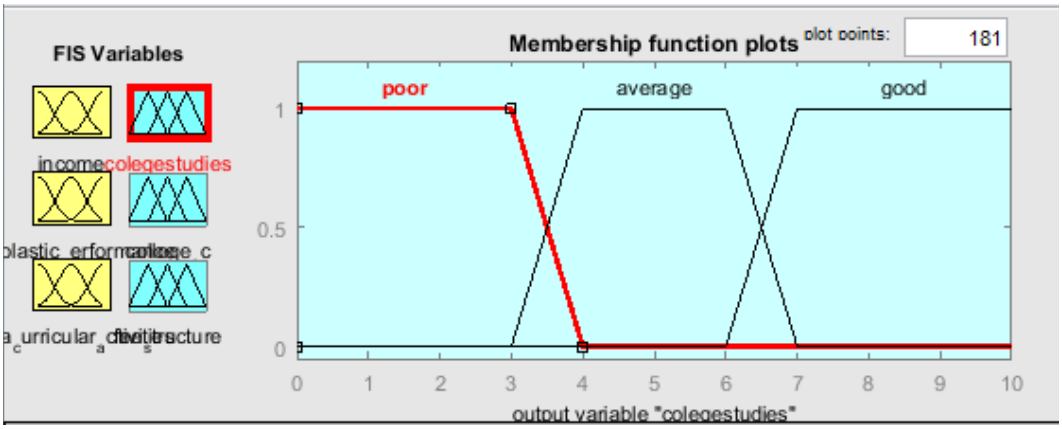



**Figure 16. FIS output (College)**

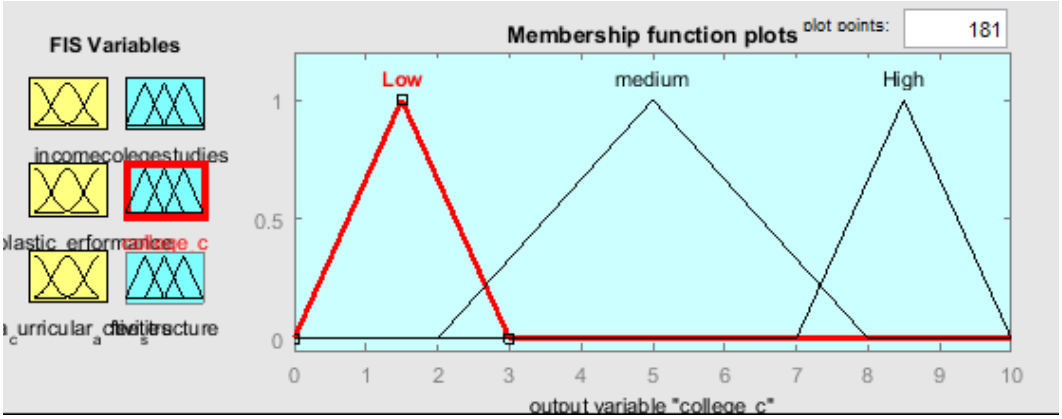
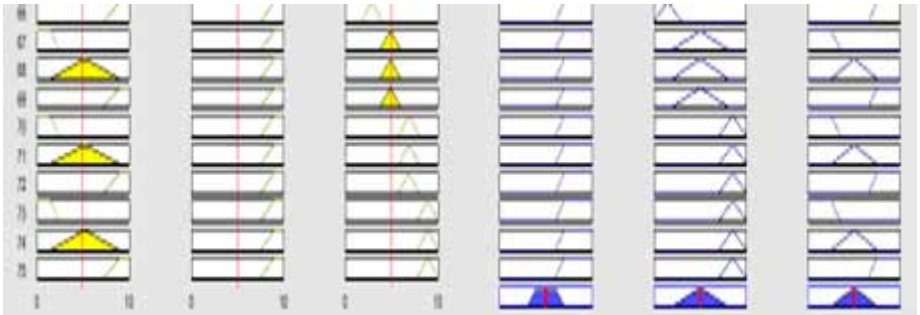



**Figure 17. Defuzzified output after combining rules**



- **Output membership function:** See Figures 14-16

A centre of gravity method is applied to find out the crisp value. Figure 18 shows the crisp output values. Algorithm is validated through statistical test. A data is collected from the various colleges of the students with attributes such as percentage of XII, extracurricular technical, sports, literacy, performing art, visual art and family income via Google form. Figure 19 shows the snapshot of the collected data from the student via Google form.

A statistical F-test is done to compare the mean of predicted value and survey data. We presume the hypothesis:

**H0:** Data collected from https://www.kaggle.com/rmalshe/studentperformance-prediction/data and Surveyed data mean are equal.
**H1:** Both data set means are not equal.

Table 1 shows the result of statistical F-test. In the table we can easily see that the mean of both the data set are different with some margin. Alpha value is considered to be less than 0.05. It can be seen that p value is achieved as 0.03 which is less than 0.05. F- Critical value is 1.59 shows that confidence level at 95 percent. Thus, we reject the null hypothesis that mean from predicted values and surveyed data should be equal.

**Figure 18. Crisp output values**

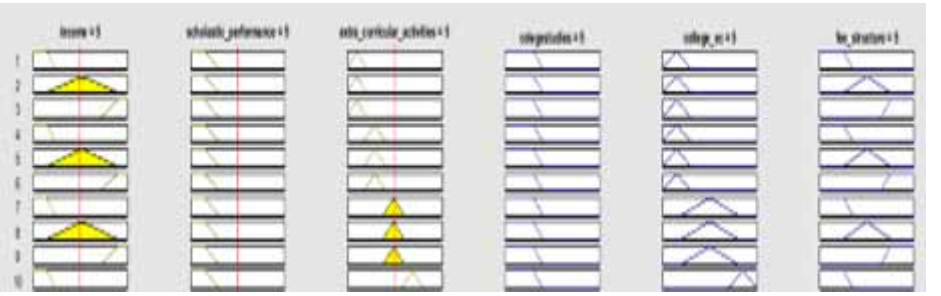


**Figure 19. Snapshot of survey data from student**

| Timestamp | Name | Age | 12 Board Percentage | Extra-Curricular Technical | Extra-Curricular Sports | Extra-Curricular Literary | Extra-Curricular Performing Art | Extra-Curricular Visual Art | mean of extra curricular | Family Annual Income |
|---|---|---|---|---|---|---|---|---|---|---|
| 2017/10/11 3:59:03 p | Monish kumar bajwa | 19 | 80.4 | 9 | 9 | 9 | 9 | 9 | 9 | 4 lakh to 6 lakh |
| 2017/10/11 4:00:00 p | Sm Kaushik | 18 | 81.2 | 6 | 7 | 7 | 6 | 8 | 6.8 | 4 lakh to 6 lakh |
| 2017/10/11 4:01:11 p | SHREENE | 18 | 89 | 5 | 6 | 7 | 9 | 7 | 6.8 | 2 lakh to 4 lakh |
| 2017/10/11 4:03:41 p | Gaurav udhwan | 18 | 74 | 5 | 8 | 6 | 5 | 3 | 5.4 | 6 lakh to 8 lakh |
| 2017/10/11 4:06:06 p | Praveen kumar singh | 18 | 88.9 | 3 | 3 | 5 | 4 | 3 | 3.6 | 2 lakh to 4 lakh |
| 2017/10/11 4:08:51 p | MD ASHRAF JAMI | 18 | 68.8 | 1 | 8 | 5 | 4 | 4 | 4.4 | Below 2 lakh |
| 2017/10/11 4:13:09 p | Hardik choudhary | 18 | 85.2 | 4 | 3 | 3 | 4 | 3 | 3.4 | Below 2 lakh |
| 2017/10/11 4:14:14 p | Raghvendra singh | 19 | 80 | 5 | 9 | 10 | 9 | 7 | 8 | 4 lakh to 6 lakh |
| 2017/10/11 4:15:00 p | Prakhar Bhatt | 18 | 86 | 6 | 8 | 8 | 8 | 5 | 7 | Above 8 lakh |
| 2017/10/11 4:15:21 p | Divyansh Somvanshi | 18 | 94 | 6 | 2 | 6 | 9 | 5 | 5.6 | Above 8 lakh |
| 2017/10/11 4:17:01 p | Ayush Varshney | 17 | 91 | 8 | 4 | 6 | 4 | 4 | 5.2 | Below 2 lakh |
| 2017/10/11 4:18:20 p | Vipul kumar | 17 | 84 | 10 | 10 | 10 | 10 | 10 | 10 | Below 2 lakh |
| 2017/10/11 4:19:50 p | Sagar Tiwari | 17 | 86 | 5 | 8 | 2 | 8 | 6 | 5.8 | Above 8 lakh |

**Table 1. Statistical F-test**

| | Predicted Value Using Inference Method | Surveyed Data |
|---|---|---|
| Mean | 57.42815 | 68.46663462 |
| Variance | 176.8503 | 103.8576315 |
| Observations | 638 | 638 |
| df | 637 | 637 |
| F | 1.702815 | |
| P(F<=f) one-tail | 0.030027 | |
| F Critical one-tail | 1.591972 | |

Figure 20 shows the students allotted to the college based on their ranking. Here AAA is considered as the highest-ranking college whereas the BB is the least ranking college. A student who cores greater than 70 will be assigned to AAA ranking college whereas a student who scores less than 50 can be assigned as BB College. Mediocre students are assigned to AA and A colleges.

Next section describes the results for weighted fuzzy decision tree.

## Weighted Fuzzy Decision Tree

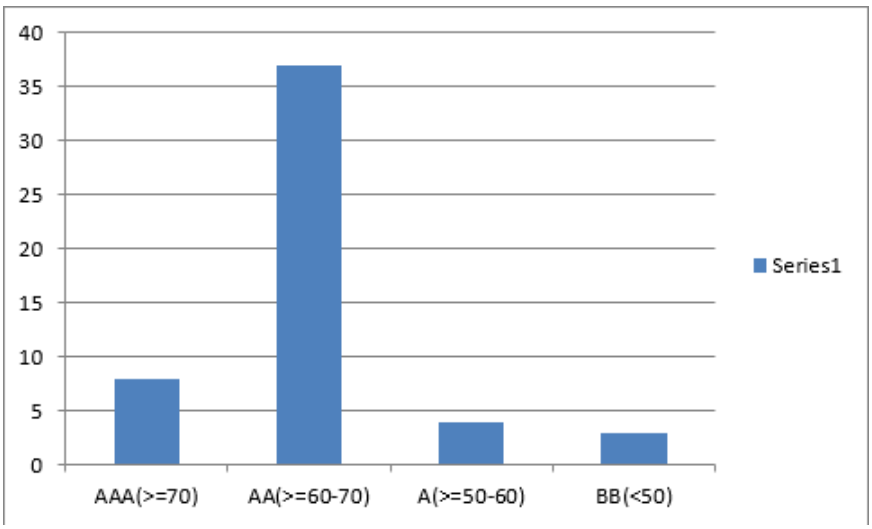
This section explains the result analysis of proposed algorithm and validates the result through ANFIS approach.

### *Result Analysis of Weighted Fuzzy Decision Tree*

This section explains the analysis of result using proposed approach i.e weighted fuzzy decision tree. Firstly, data set is collected as explained in section 3.3. As discussed in algorithm -1, the Information gain is calculated for the attributes marks, extracurricular activities and income are:

$$IG_{marks} = 1.03,\ IG_{ec} = 0.42,\ IG_{income} = 0.39$$

**Figure 20. Recommendation of colleges according to the grade**



According to equation 3 in algorithm-1, the weight of the attributes are calculated. Table 2 shows the weight of each attribute. Further these weights are applied to fuzzy decision tree.

A weighted fuzzy decision tree is used to predict the score. Algorithm execute for different runs for 218 respondents. Table 3 shows the confusion matrix for different colleges for different runs. The performance is measured on the factors of precision, recall, F1- score and support.

It can be seen that (table -3) in run -1, it is achieved as 81% as it decreases to 74% for another three runs, again it reaches to 79.6%. Thus, it can be concluded that weighted fuzzy decision tree achieves the accuracy between 75-80%.

Figure 21 shows the two sets of respondents. As no of respondent increases, accuracy increases. One set of respondents is 62 and the other set of respondents is 218. In the graph blue color represents the 218 respondents whereas orange color depicts the 62 respondents for different colleges.

Figure 22 shows the mean accuracy achieved for different runs for different size of respondent. Here size of respondent is 62 and 218. Blue color represents the 218 respondents whereas orange color represents the 62 respondents.

**Table 2. Weight allocation table using decision tree**

| No. of Entry | Weight of 12th Marks | Weight of Extra Curricula Activities | Weight of Income | Accuracy |
|---|---|---|---|---|
| 218 | .55 | .25 | .20 | 69% |

**Table 3. Confusion matrix for different colleges**

| | Colleges/ Run | Amity | DTU | IIT Kanpur | JIIT, Noida | MNIT Jaipur | Avg / Total | Accuracy |
|---|---|---|---|---|---|---|---|---|
| Run-1 | precision | 0.86 | 0.89 | 0.91 | 0.8 | 0.67 | 0.8 | 81.3559322 |
| | Recall | 0.8 | 0.89 | 0.91 | 0.86 | 0.67 | 0.81 | |
| | F1-score | 0.83 | 0.89 | 0.87 | 0.83 | 0.67 | 0.81 | |
| | Support | 5 | 9 | 11 | 14 | 9 | 59 | |
| Run-2 | precision | 0.69 | 0.78 | 0.8 | 0.86 | 0.62 | 0.76 | 74.57627119 |
| | Recall | 0.9 | 0.78 | 0.89 | 0.63 | 0.67 | 0.75 | |
| | F1-score | 0.78 | 0.78 | 0.84 | 0.73 | 0.64 | 0.74 | |
| | Support | 10 | 9 | 9 | 19 | 12 | 59 | |
| Run-3 | precision | 0.92 | 0.44 | 0.67 | 1 | 0.83 | 0.81 | 72.88135593 |
| | Recall | 0.8 | 0.88 | 1 | 0.75 | 0.36 | 0.73 | |
| | F1-score | 0.86 | 0.58 | 0.8 | 0.86 | 0.5 | 0.73 | |
| | Support | 15 | 8 | 10 | 12 | 14 | 59 | |
| Run-4 | precision | 0.83 | 0.62 | 0.8 | 0.73 | 0.77 | 0.75 | 74.57627119 |
| | Recall | 0.83 | 0.67 | 0.67 | 0.73 | 0.83 | 0.75 | |
| | F1-score | 0.83 | 0.64 | 0.73 | 0.73 | 0.8 | 0.75 | |
| | Support | 12 | 12 | 12 | 11 | 12 | 59 | |
| Run-5 | precision | 0.88 | 0.88 | 0.92 | 0.8 | 0.5 | 0.82 | 79.66101695 |
| | Recall | 0.94 | 0.7 | 0.92 | 0.62 | 0.75 | 0.8 | |
| | F1-score | 0.91 | 0.78 | 0.92 | 0.7 | 0.6 | 0.8 | |
| | Support | 16 | 10 | 12 | 13 | 8 | 59 | |

**Figure 21. Graph shows the difference between number of respondents**

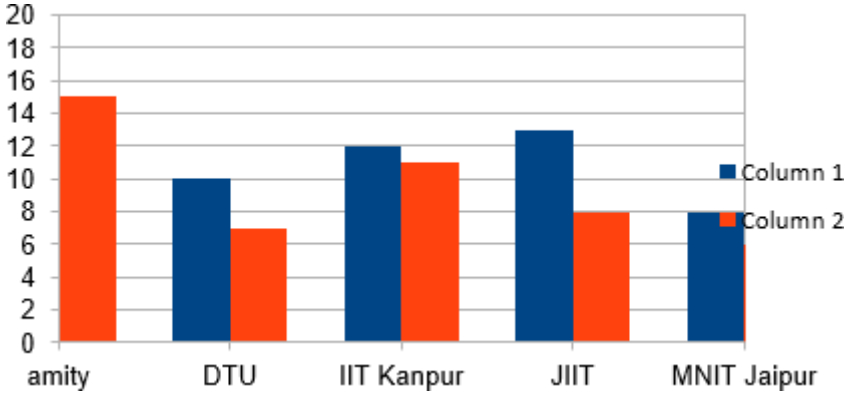
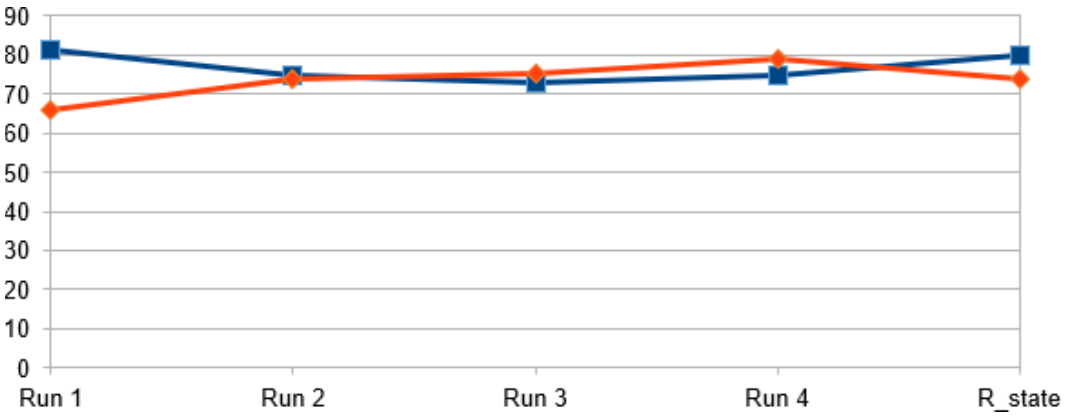



**Figure 22. Mean accuracy for different runs for different size of respondents**




*Validation*

Adaptive Neural based fuzzy system is used to validate the above result on the same no. of respondent. Matlab is used to run the algorithm. On different Epoch. Figures 23, 24, and 25 show the result on 100, 500, and 1000 epochs.

The error rate for different epoch was achieved is shown in Table 4. As the number of epochs are increased error rate decreases.

It shows that the proposed algorithm achieved the accuracy between 75 to 80 percent with weighted fuzzy decision tree.

## Fuzzy Ant Colony Optimization Result Analysis

This section analysis the path traversed by the student in various ranked colleges. Figures 26, 27, and 28 show the pheromone graph of edges for various levels. Figure 19 shows the pheromone graph of edge for level 0 to level 1. These graphs monitor the change in the pheromone levels of the paths due to the selection of the students in various colleges at each level. The fluctuations in the level of pheromone depict the varying student selection. Graph 26, 27, and 28 contain the count of the student traversing from level 0 to level 1, level 1 to level 2 and level 2 to level 3, respectively. X-Axis denotes the number of students who have already traversed that path: Y-Axis denotes the pheromone levels

**Figure 23. ANFIS outputs (100 epochs)**


**Figure 24. ANFIS outputs (500 epochs)**

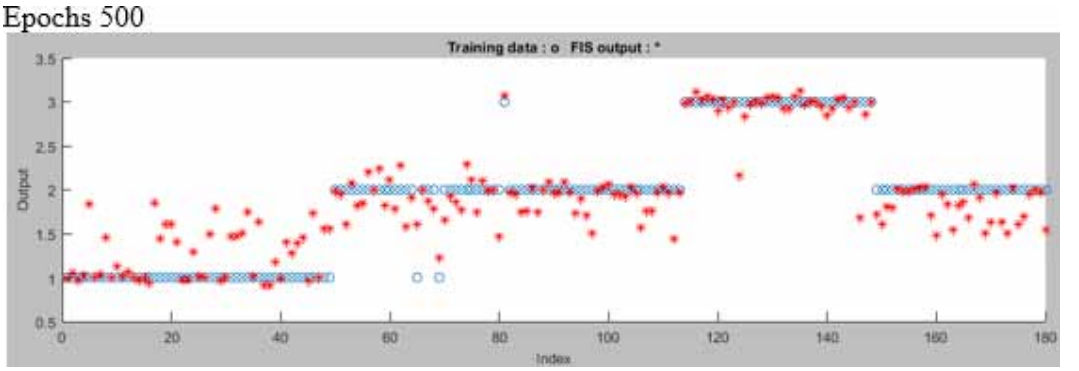
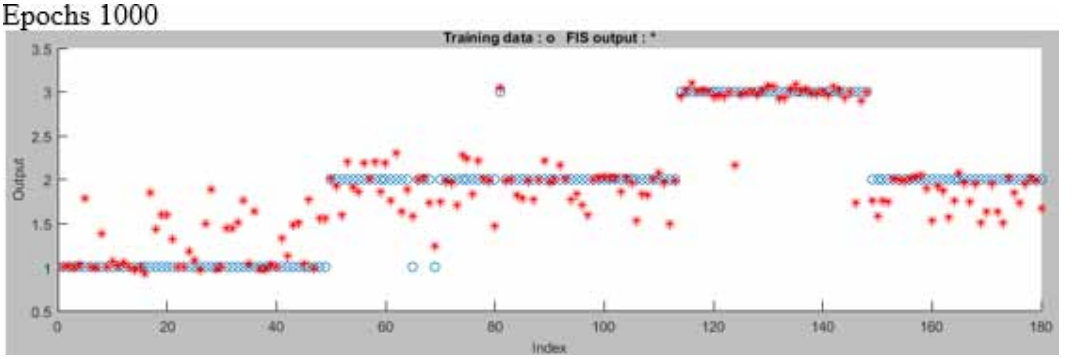



**Figure 25. ANFIS outputs (1,000 epochs)**




The decline in the graph shows the declining selection of the student in various colleges. As the pheromone goes below the threshold value, the path will be changed.

The stability in the graph shows the constant selection of the student on that path. Thus, the swarm intelligent are one of the efficient approaches to predict the colleges to the students on the basis of their various feature. Table 5 shows the comparative analysis of other prediction model of student performance.

**Table 4. Error between different epochs**

| Epochs | Error |
|---|---|
| 100 | 0.35129 |
| 500 | 0.2896 |
| 1000 | 0.280 |

**Figure 26. Pheromone graph of edges for level 0 to level 1**

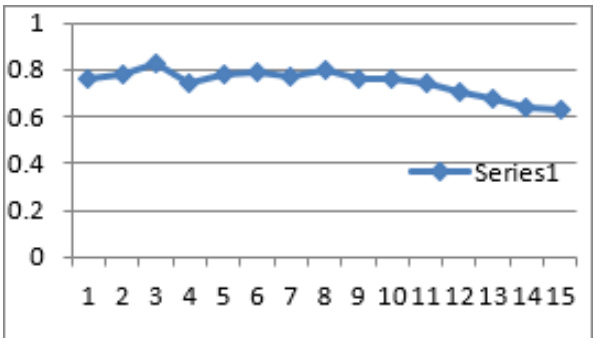



**Figure 27. Pheromone graph of edges for level 1 to level 2**

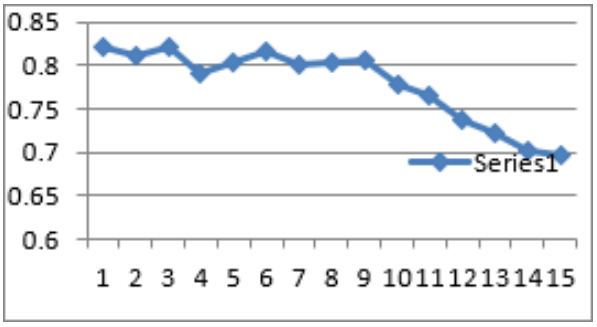



**Figure 28. Pheromone graph of edges for level 2 to level 3**

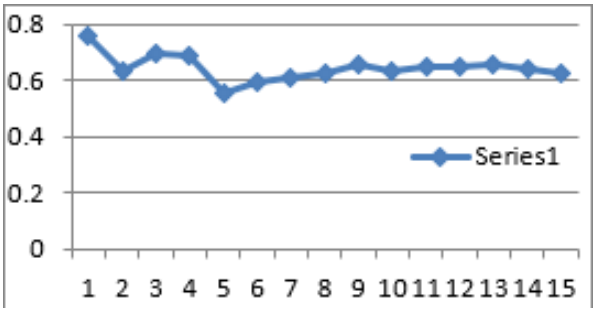


**Table 5. Comparative analysis with other approaches**

| Author | Technique | Accuracy/ Mean Absolute Error |
|---|---|---|
| Zhang. W (2008) | Relational CF | MAE = 0.697 |
| Guo, B., et.al., (2015) | Deep learning | Accuracy-77.2% |
| Proposed model | Weighted fuzzy decision tree | MAE =0.28 |

Hence Table 5 shows that the proposed algorithm achieves mean absolute error is 0.28 which is improved in comparison of relational CF. If the algorithm compares with the deep learning it has achieved the similar accuracy with extra-curricular attributes as additional attribute.

## CONCLUSION

This paper proposed the Mamdani fuzzy inference method to predict the suitable colleges for a student on the basis of percentage of XII, family income and extracurricular activities. It also proposes a weighted fuzzy decision tree to predict the suitable colleges. To validate the result of the Mamdani inference method, statistical F-test is done. The P value is achieved 0.03 which less then alpha value 0.05 at the confidence level of 95 percent. To validate the result of another approach i.e. weighted fuzzy decision tree, an adaptive Neuro-fuzzy inference system (ANFIS) algorithm is used. The ANFIS gives mean absolute error is 0.28, run on 1000 epochs. The error came out to be constant after 1000 epochs. Future work can include the prediction of colleges on the basis of student profile and increasing the number of colleges. The other factors also can be taken in the consideration. It may include the bigger data set to reduce the error. Though the algorithm has been tested for few colleges, increasing number of colleges will be able to classify the data more properly into more number of classes. This paper suggested for the recommendation of engineering colleges; future work can include different skill type colleges and other swarm intelligent techniques can apply to predict the colleges.

## REFERENCES

Agarwal, S., Goyal, M., Kumar, A., & Rajalakshmi, K. (2016). Intuitionistic fuzzy ant colony optimization for course sequencing in e-learning. In *Contemporary Computing (IC3), 2016 Ninth International Conference on* (pp. 1-6). IEEE. doi:10.1109/IC3.2016.7880248

Ahamed, A. S., Mahmood, N. T., & Rahman, R. M. (2017). An intelligent system to predict academic performance based on different factors during adolescence. *Journal of Information and Telecommunication*, *1*(2), 155–175. doi:10.1080/24751839.2017.1323488

Borkar, S., & Rajeswari, K. (2013). Predicting students academic performance using education data mining. *International Journal of Computer Science and Mobile Computing*, *2*(7), 273–279.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 415-421). International World Wide Web Conferences Steering Committee. doi:10.1145/3041021.3054164

Giovannella, C., Scaccia, F., & Popescu, E. (2013). A PCA study of student performance indicators in a Web 2.0-based learning environment. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on* (pp. 33-35). IEEE.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In *Educational Technology (ISET), 2015 International Symposium on* (pp. 125-128). IEEE. doi:10.1109/ISET.2015.33

Krishnamurthi, R., & Goyal, M. (2018). Automatic Detection of Career Recommendation Using Fuzzy Approach. *Journal of Information Technology Research*, *11*(4), 99–121. doi:10.4018/JITR.2018100107

Mamdani, E. H. (1976). Application of fuzzy logic to approximate reasoning using linguistic synthesis. In *Proceedings of the sixth international symposium on Multiple-valued logic*, (pp. 196-202). IEEE Computer Society Press.

Ogor, E. N. (2007). Student academic performance monitoring and evaluation using data mining techniques. *Electronics, Robotics and Automotive Mechanics Conference,* 354-359. doi:10.1109/CERMA.2007.4367712

Rosenberg, L., Pescetelli, N., & Willcox, G. (2017). Artificial Swarm Intelligence amplifies accuracy when predicting financial markets. In *Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017 IEEE 8th Annual* (pp. 58-62). IEEE. doi:10.1109/UEMCON.2017.8248984

Sugeno, M., & Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, *1*(1), 7. doi:10.1109/TFUZZ.1993.390281

Venkatesan, E., & Selvaragini, S. (2017). A Study On The Result Based Analysis Of Student Performance Using Data Mining Techniques. *International Journal of Pure and Applied Mathematics*.

Verma, P., Sood, S. K., & Kalra, S. (2017). Student career path recommendation in engineering stream based on three-dimensional model. *Computer Applications in Engineering Education*, *25*(4), 578–593. doi:10.1002/cae.21822

*Mukta Goyal is presently working as an Assistant Professor at Jaypee Institute of Information Technology (Deemed to be University), Noida. She has more than 18 years of teaching experience. Her research interest includes Soft computing, E-learning, Artificial intelligence, Compiler design, Theory of Computation. She published more than 25 research papers in reputed International journals and conferences. She is also reviewer/committee member of various International Journals and Conferences.*