# Hybrid Feature Extraction Approach for Enhancing Classification of Education Data Mining

Priyanka Devi[1], Gaganpreet Kaur[2]
*[1]Sri Sukhmani Institute of Engineering & Technology*
*[2]Sri Sukhmani Institute of Engineering & Technology*
*(E-mail: pjangra9105@gmail.com, gagan682@gmail.com )*

***Abstract*—** The learning algorithms are widely adopted by educational domain so as to calculate the overall performance of the students depending on academic data of the students. The usage of automatic mechanism can ease the decision deriving process regarding the reputation of the organizations as per the student's performance. The traditional student data mining mechanisms uses the feature extraction and classification mechanism for upgrading system's performance. A scheme namely LDA and PCA for the extraction of features is discovered as better as compared to existing techniques. Therefore, the author in the work develops a novel student data mining concept by using the ANFIS classifier and a hybrid feature extraction mechanism by using PCA and LDA technique. The proposed work is implemented in MATLAB simulation platform in the terms of F-Measure, Recall and Precision. This proposed work was assessed for various extraction schemes like cfs Attribute, Principal component, Relief Attributes, Gain Ratio Attributes, Chi Squared and Filtered Attributes. As per the observed facts, the proposed work is confirmed more efficient than the other mechanisms.

***Keywords*—** *Educational Data Mining; Knowledge Discovery; Student's performance; Feature Extraction; Classifiers*

## I. INTRODUCTION

With the advancements in the technology, the modern education also gets advanced as the e-learning is widely adapted by the institutes to teach the students. It becomes significant to collect the student's data to evaluate their performance in the organization and to discover the learning process. The incremental growth of educational data results leads to the requirement for establishing the research in educational data mining (EDM). In other way, the EDM can described as the rising discipline related to the developing mechanism for discovering the exclusive types of data that is gathered from educational setting and then to utilize it to understand the overall performance of students.

A huge number of researches by several authors have done in this field. For this purpose, feature extraction mechanism was implemented for extract the principal features from gathered student's data. After this the classification was done by classifiers by taking the extracted data. From the observations from previous work, it's been concluded that the mechanism implemented for student data mining was less efficient and hence led to the lower accuracy rate. Thus, in this work a novel concept has been designed for enhancing the accuracy for the predicted data. For this purpose, the extraction of feature is performed by collaboration the two of the most prominent feature extraction mechanism i.e. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA).

The Hybrid feature extraction is applied for extraction of the relevant data efficiently. Along with this, Adaptive Neuro Fuzzy Inference System (ANFIS) is applied for classifying the extracted features so that the more accurate decisions can be derived.

## II. FEATURE EXTRACTION

### *A. LDA (Linear Discriminant Analysis)*

LDA approach is employed in face recognition process. It is a numerical approach used for comparison of unknown patterns with known patterns. This method use variables like continuous independent and category based dependent variable. This approach also used PCA for low dimension representation. LDA use classes which are based on database by dividing the database into number of classes. According to segmented classes LDA perform various operations. The classes are randomly created by using sample database. LDA is utilized in the cases where the unequal frequencies are present and the requirement is to evaluate dynamically generated information. It provides maximum distinction among within-class and between-class variance. The fundamental difference between PCA and LDA: PCA performs feature classification; LDA have major contribution in data classification. Hence, it provides greater understanding about the feature data distribution.

Mechanisms followed by LDA are:

1. *Class dependent transformation*: An increase in ratio of between-class variance and within-class variance is obtained in order to provide maximum attainable value. The objective behind this enhancement is that ratio is directly concerned with separation factor. The higher value of ratio provides greater distinction.

2. *Class independent transformation*: In this approach, between-class variance is not considered. Its aim is to make the ratio overall change optimum within-class variance. It utilizes a unique optimizing mechanism on the data. Such that all the data points are transformed, without considering their class identity. Every class is treated as a distinct class corresponding to other classes.

### *B. PCA (Principal Component Analysis)*

It is basically a conversion method that converts number of incorporated variables into unassociated variables. In the classification of the image this technique operates in smooth and effective way along with its compression. Mathematical calculations and functions are taken to convert the incorporated variables. This helps in briefly describing the datasets. Its basic functions are implemented in form of principal components. To hold the data as flexible form, 1st principal component is to be considered and second principal component is maintained vertical to the first component's subspace. However, the 3rd principal component is used for maximizing the divergence of subspace ⊥ to 1st and 2nd component. Principal Component Analysis basically acts as backup to IHS based conversion method. The basic reason behind this same working is mutual relationship between both techniques of MS band as are known as PC1, PC2 and PC3 and so on. Its working is given below:

- Initially, the IR is provided to the PAN and MS value, further re-sampling the MS.
- Then, conversion of MS bands to components such as PC1, PC2, and so on is carried out.
- Histogram links are provided between PAN and PC1.
- Then restore the PC1 with PAN.
- At the end PAN is converted to left principal components.

## III. CLASSIFIER

ANFIS utilizes the collaborated characteristics of fuzzy system with neural network under an umbrella. ANFIS refers to Adaptive Neuro Fuzzy Interference System. It is technique which implements various learning method for training an ANN to Fuzzy model as name shows. FIS is implements non-linear mapping of the inputted data. The mapping is achieved with utilization of several if-then rules of fuzzy system. Each and every rule of system describes the mapping nature. Parameters used in if-then rules of fuzzy network are termed as input space. The output of input space is termed as output space. Hence the working and the output generated by fuzzy is sensitive to the selected parameters. The parameter selection is not dependent on any available procedures itself. Hence ANFIS is a solution to this problem. ANFIS facilitates the Fuzzy network for which the membership functions or parameters can be tuned by using various algorithms such as least square method and back propagation etc. ANFIS enables a Fuzzy system to be trained from data which is being modeled. Fuzzy network or rules should be chosen efficiently when they are going to incorporate as ANFIS.

Sugeno: Sugeno is a fuzzy model. It works as follows:

- Let's consider that fuzzy inference two inputs: x & y respectively and generate corresponding single output z and A and B be fuzzy sets in antecedent.
- First-order Sugeno model comprise following rules:
- Rule 1:

  $x$ is $A_1$ ; $y$ is $B_{1;}$ Implies:

$$z_1 = p_1x + q_1y + r_1 \qquad (1)$$

- Rule 2:

  $x$ is $A_2$; $y$ is $B_2$; Implies:

$$z_2 = p_2x + q_2y + r_2 \qquad (2)$$

## IV. PROBLEM FORMULATION

Data mining is the procedure of extracting important and meaningful values from the database. For the purpose of fetching the data as required in the close future a data warehouse is type of data storage where data is accumulated. Data warehouses can store all kinds of data. In particular the data type is depended on the industry type from which it is taken.

In most industries, all types of data are recorded, but in some companies only that information is stored. This information is useful and meaningful. The data stored in the warehouse are useful for system used for decision making. Based on historical data, decision making about future schemes can easily or effectively be made.

Previously, much work is done for predicting the performance of student using different techniques of feature selection. In recent studies, researchers use different feature selection techniques and the combination of classifiers to produce efficient prediction models. A research is required to discover the performance analysis with respect to prediction accuracy in combination of various feature selection mechanism with differently classifiers. Moreover, the classifiers which are advanced needed for classification. The techniques employed in the existing work provide less prediction accuracy, efficiency, and effectiveness. Considering these issues in the existing work, a fresh approach is demanded for identifying the prediction accuracy of different available feature selection algorithm in the era of classifiers being used on educational data.

## V. PROPOSED WORK

After reviewing the issues in the existing work, a novel approach is proposed. Initially, feature selection will be performed using two schemes i.e. Principal Component Analysis and Linear Discriminate Analysis. The collaboration of these techniques can perform effectively for making prediction of accuracy. The advantage of PCA and LDA is that it extracts the features and also normalizes the data in an effective manner. Once the features are extracted, next phase is classifying the data which is extracted. ANFIS classifier is taken in proposed work to fulfill this objective. ANFIS is discovered to be more advantageous in comparison to other classifiers because it is a grouping of two most prominent mechanisms i.e. ANN and Fuzzy Inference System. Thus it is supplementary efficient to perform even in the case of uncertainties.

The methodology of proposed work consists of several steps which clearly describes the working of the projected work and is given below:

Step 1 Start

Step 2 Initially the dataset is formed from the accessible knowledge of the students.

Step 3 After that, as per the students' information, normalization of the collected data is done.

Step 4 From the original dataset the tested data is chosen after normalizing the data. For the testing purpose the tested dataset is utilized.

Step 5 After that, for extorting the principle components from the dataset the PCA feature extortion is applied. Afterward the LDA method is applied to PCA extracted features.

Step 6 Now the training and testing of the extorted data is initiated for which the ANFIS is utilized. For this, initially the training of ANFIS is accomplished and tested data related to the trained data was classified.

Step 7 Ultimately, the estimation of the proposed work is accomplished as the precision, Recall and F-measure.

The methodology of proposed work is represented in fig 1.

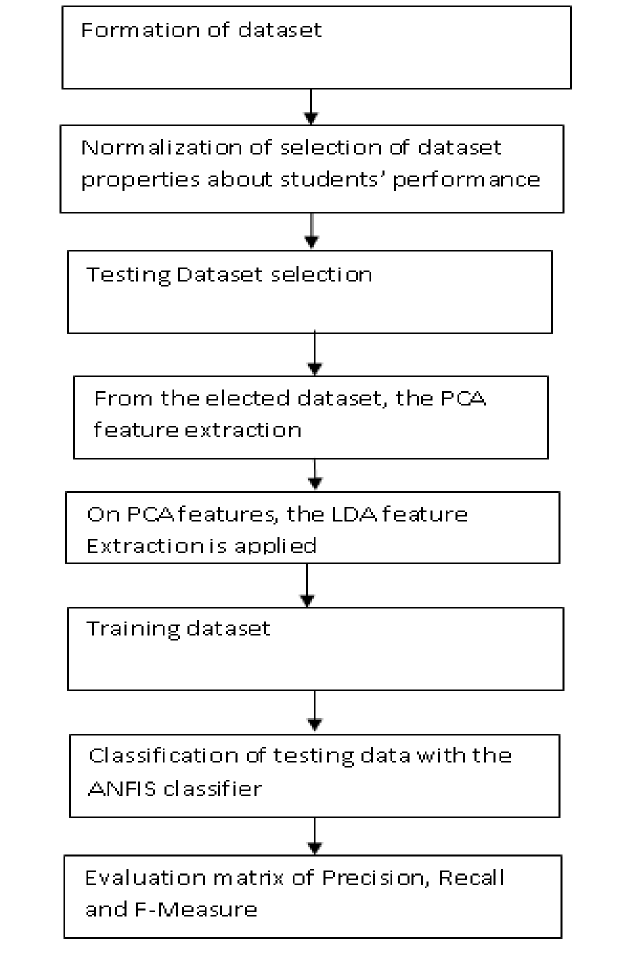
Formation of dataset
Normalization of selection of dataset properties about students' performance
Testing Dataset selection
From the elected dataset, the PCA feature extraction
On PCA features, the LDA feature Extraction is applied
Training dataset
Classification of testing data with the ANFIS classifier
Evaluation matrix of Precision, Recall and F-Measure

Fig. 1. Methodology of Proposed Work

## VI. Experimemtal Results

In this study, the ANFIS classifiers are for classifying the student's data. Before classification, the LDA and PCA are applied for extraction of feature. Proposed work's performance is assessed corresponding to Recall, F-measure, and Precision. Precision is regarded as positive predictive value. It is measured as follows:

$$Precision = TP/(TP + FP) \qquad (3)$$

Where TP is true positive, FP is false positive.

Recall defines that how many relevant items are elected. The formulation given below is used for evaluating the recall for proposed work:

$$Recall = TP/(TP + FN) \qquad (4)$$

F-Measure is a performance matrix which is used for evaluating the harmonic mean of precision and recall. The formulation is as follows:

$$Fmeasure = 2 * (precision * recall/(precision + recall)) \qquad (5)$$

The comparison of proposed work is done with traditional classification algorithms in terms of F-Measure, Precision and Recall. Principal Component Analysis, Gain Ratio, Cfs Subset, Chi Squared, Relief Attribute, etc, are the traditional feature selection algorithms.

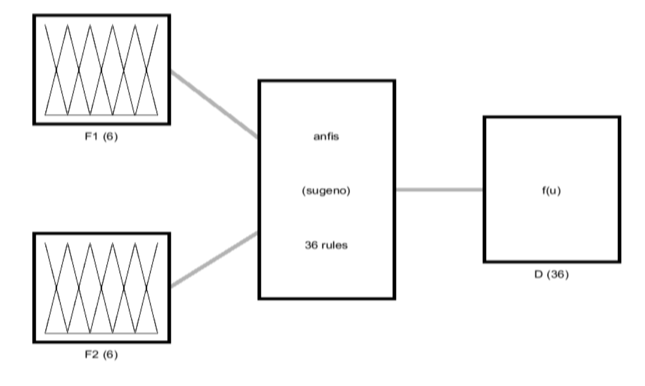

Fig. 2. Proposed ANFIS System

The graph in the fig. 2 shows the proposed ANFIS system. In this figure two inputs having six degree of memberships in each are given to the Sugeno model of the ANFIS system on which 36 rules are applied to provide single output with 36 membership functions. ANFIS which is the grouping of the ANN and FIS is used in the study for offering accuracy in the work.

The graph in the fig. 3 shows the Degree of membership for input F1. The input F1 is shown on the x-axis whereas the Degree of membership is represented along the y-axis. There are also six ranges of the degree of membership for the input F1 that are NB, NM, NS, ZE, PS and PM that ranges from -6 to -5.5, -6 to -4.5, -5.5 to -3.5, -4.5 to -3, -3.5 to -2 and -3 to -2.

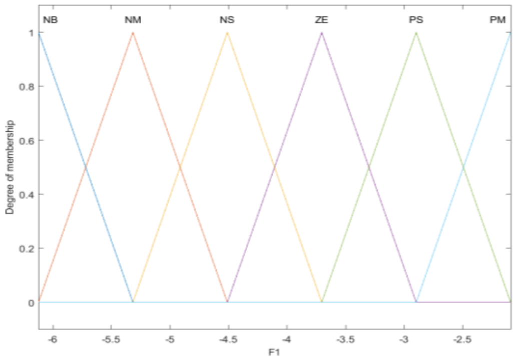

Fig. 3. Degree of membership for input F1

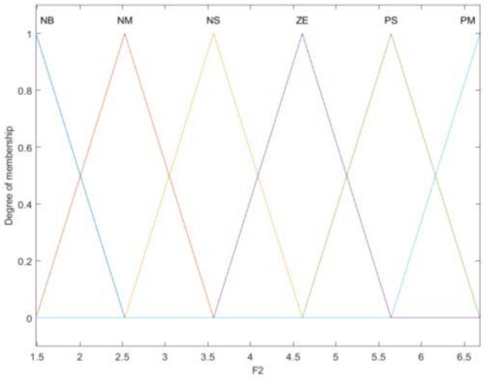

Fig. 4. Degree of membership for input F2

The graph in fig. 4 shows the Degree of membership for input F2. In this graph the input F2 is represented along x-axis that ranges from 1.5 to 6.5 and Degree of membership is demonstrated on the y-axis that ranges from 0 to 1. The range of degree of membership of input F2 varies in six levels that are NB, NM, NS, ZE, PS and PM that are ranges from 1.5 to 2.5, 1.5 to 3.5, 2.5 to 4.5, 3.5 to 5.5, 4.5 to 6.5 and 5.5 to 6.5 .

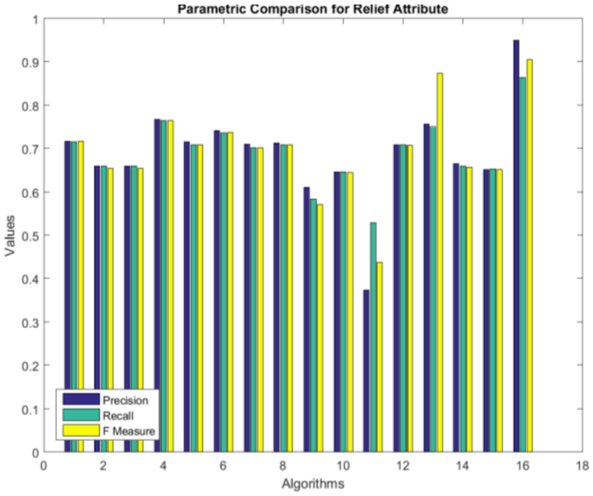

Fig.5. Comparison Analysis of Relief Attribute.

The graph in fig. 5 shows the comparison investigation of the Relief Attribute. In this graph it is shown that the Recall, Precision and F-measure of the projected work are higher comparative to the traditional mechanisms. The algorithms are shown along x-axis and values are shown along y-axis.

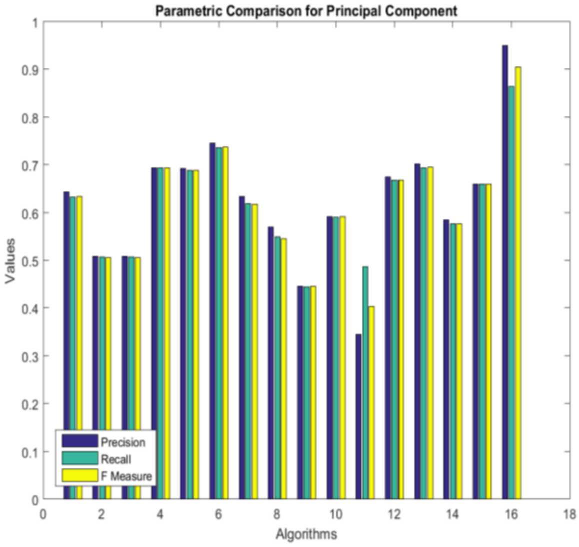

Fig. 6. Comparison Analysis of Principal Component Attribute

The graph of fig. 6 depicts comparison Analysis of Principal Component. The algorithms are shown on the x-axis that is 16 in number and the values are represented along y-axis that ranges from 0 to 1. It is represented in the graph that the values of the projected mechanism for PCA are higher than the conventional algorithms.

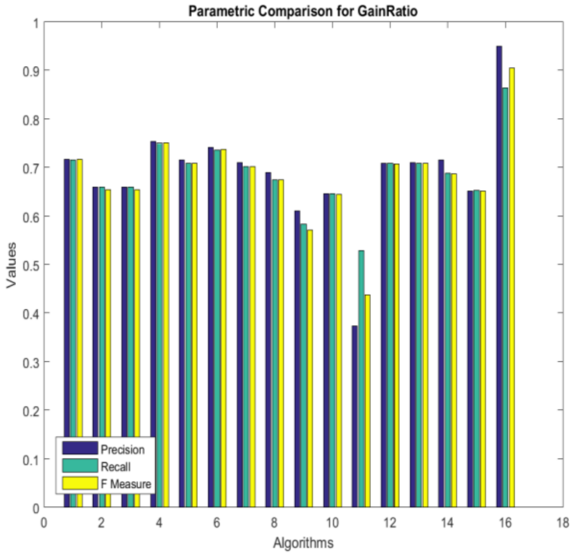

Fig.7. Comparison Analysis of Gain Ration Attributes

The graph in fig. 7 demonstrates the comparison Analysis for the Gain Ration Attributes. Here the comparison of classification algorithms with gain ratio and the observations of the proposed work are better than others.

The graph in the fig. 8 shows the comparison analysis of Filtered attribute. Three features of the proposed system i.e. F-measure, Precision, and Recall and have higher value for the filtered attribute whose values are 0.98, 0.88 and 0.91.

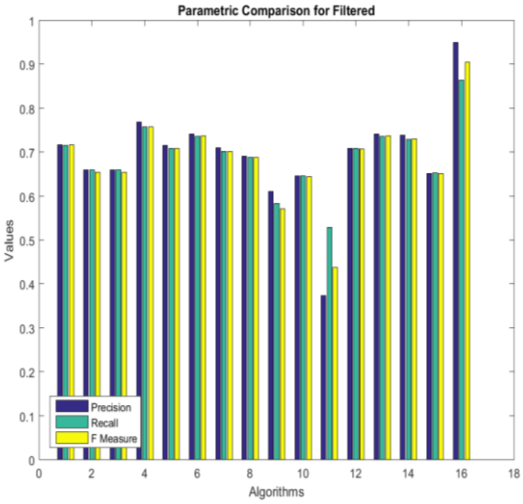

Fig.8. Comparison Analysis of Filtered Attributes

The proposed system is having values higher than conventional schemes in the graph of fig. 9. The figure shows the comparison analysis of Chi Squared Attribute. There are total 16 algorithms represented in the graph along x-axis.

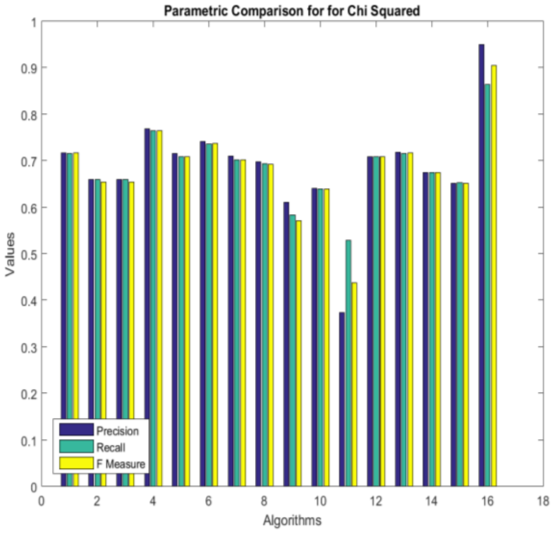

Fig.9. Comparison Analysis of Chi Squared Attributes

The graph of fig. 10 depicts comparison Analysis of Cfs Subset Attribute. The algorithms are shown on the x-axis that is 16 in number and the values are shown on the y-axis that ranges from 0 to 1. It is shown in the graph that the values for the proposed algorithm for Cfs Subset Attribute are higher than the conventional algorithms which indicates that proposed algorithm outperforms the traditional approach.

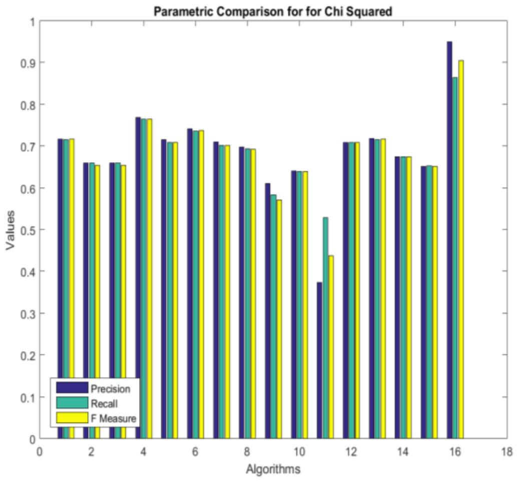

Fig.10. Comparison Analysis of Cfs Subset Attributes

Summation of the facts and figures that are gathered from above defined graphs is done in the table 1.

TABLE I. OVERALL PERFORMANCE EVALUATION OF PROPOSED WORK CORRESPONDING TO OTHER CLASSIFICATION ALGORITHMS

| Techniques | Precision | F-Measure | Recall |
|---|---|---|---|
| PCA | 0.5995 | 0.6032 | 0.5975 |
| Gain Ratio | 0.6707 | 0.6745 | 0.6670 |
| Cfs Subset | 0.6669 | 0.6731 | 0.6662 |
| Chi Squared | 0.6697 | 0.6759 | 0.6683 |
| Filtered Attribute | 0.6707 | 0.6745 | 0.6670 |
| Relief Attribute | 0.6707 | 0.6745 | 0.6670 |
| Proposed Work | 0.95 | 0.8636 | 0.9048 |

The data in table describes the overall performance of the various considered features selection mechanisms.

## VII. CONCLUSION

The education data mining or student data mining has a key role for evaluation of the students' performance an institute. This is done for finding the overall students' performance according to the given parameters. Data mining concept is used in the student data mining for measurement of the performance. Various classification and feature extraction schemes were used for the purpose in past. PCA, ANN and KNN are some extensively used schemes for this task.

This study develops a mechanism for student data mining by use of hybrid feature extraction mechanism and ANFIS classifier. PCA and LDA are used here for the hybridization purpose. Along with this the ANFIS as the classifier is used for classification of data which is extracted. The performance evaluation is done in comparison with the traditional feature extraction process and it is observed that the projected work outperform the traditional classifiers. The overall precision, F-measure and recall of the proposed work are 0.95, 0.8636 and 0.9048 respectively. The proposed work provides better results but in future the more research can be done on the techniques used for the selection of features. These techniques of feature selection can be taken for selecting the optimum features and give them to the classifiers which in turn can enhance the system performance significantly.

## REFERENCES

[1] Raheela Asif, Agathe Merceron, Syed Abbas, Alic Najmi, Ghani Haidera, "Analyzing undergraduate students' performance using educational data mining", ELSEVIER, vol 113, Pp 177-194, 2017.

[2] Charoula Angeli, Sarah K.Howard, Jun Mab Jie, Yangb Paul, A. Kirschnercd, "Data mining in educational technology classroom research: can it make a contribution?", ELSEVIER, vol 113, Pp 226-242, 2017.

[3] Evandro B. Costaa Baldoino , Fonsecaa Marcelo, Almeida Santanaa Fabrísia, Ferreirade Araújob, Joilson Regod, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", ELSEVIER, vol 73, Pp 247-256, 2017.

[4] AlejandroPeña-Ayala, "Educational data mining: a survey and a data mining-based analysis of recent works", ELSEVIER, vol 41,Pp 1432-1462, 2014.

[5] ManolisChalaris,StefanosGritzalis,ManolisMaragoudakis,CleoS gouropoulou,AnastasiosTsolakidis, "Improving quality of educational processes providing new knowledge using data mining techniques", ELSEVIER, vol 147, Pp 390-397, 2014.

[6] Surjeet Kumar Yadav, "Data mining applications: a comparative study for predicting student's performance", ijitce, vol 1(12), Pp 13-20.

[7] Surjeet Kumar Yadav, "Data mining: a prediction for performance improvement of engineering students using classification", WCSIT, vol 2(2), Pp 51-56, 2012.

[8] Sayali rajesh suyal,"Quality improvisation of student performance using data mining techniques", vol 4(4), Pp 1-4, 2014.

[9] Paulo Cortez, "Using data mining to predict secondary school student performance", 2008.

[10] Muluken Alemu Yehuala, "Application of data mining techniques for student success and failure prediction", (The Case Of Debre Markos University), vol 4(4), Pp 91-95.

[11] Brijesh Kumar Baradwaj, "Mining educational data to analyze students' performance", ijacsa, vol 2(6), Pp 63-70, 2011.

[12] Ajinkya Kunjir," Recommendation of data mining technique in higher education", IJCER, vol 5(3), Pp 29-35, 2015.

[13] Jayashree M Kudari," Survey on the factors influences the students‟ academic performance", IJERMT, vol 5(6), Pp 30-37, 2016,

[14] K. Amarendra "Research on data mining using neural networks" Special Issue of International Journal of Computer Science & Informatics (IJCSI), , Vol.- II, Issue-1, 2, Pp2231–5292.

[15] Anand V. Saurkar, "A Review paper on various data mining techniques", International Journal of Advanced Research in Computer Science and Software Engi neering, Volume 4, Issue 4,Pp 98-101, 2014G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.

Priyanka Devi is currently pursuing masters degree in Computer Science Engineering in Sri Sukhmani Institute of Engineering & Technology, Derabassi, Punjab

Gaganpreet Kaur is an assistant professor in Computer Science department in Sri Sukhmani Institute of Engineering & Technology, Derabassi, Punjab