IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Analysis and Comparison of Classification Algorithms for Student Placement Prediction

M. Shukla[1*], A. K. Malviya[2]

[1] CSE Dept., Kamla Nehru Institute of Technology, A.K.T.U., Sultanpur, India
[2] CSE Dept., Kamla Nehru Institute of Technology, A.K.T.U., Sultanpur, India

Corresponding Author: mekhlashukla1012@gmail.com, Mob.: +91-9616634921

Available online at: www.ijcseonline.org

Accepted: 20/Jun/2018, Published: 30/Jun/2018

***Abstract*—** Educational data mining has gained importance for discovering the useful information from the student databases. It is observed that there is a lack of performance of the students during campus selection in technical institutions. Hence the problem highlighted in this research work is: “What factors are responsible for placement of some students but why not others during campus selection of technical institutions?” The objective of this research work is related to the prediction and discovery of the factors for student placement using the data mining techniques and tool. The methodology used in this research work involves four main stages to achieve the required objectives. They are Data Collection, Pre-processing, Classification and Interpretation of Result. The Classification algorithms used in this research paper include decision tree, Naive Bayes, Neural Network (Multilayer perceptron) and Sequential Minimal Optimisation. It has been found that Naive Bayes algorithm works best in student placement prediction with maximum accuracy. The identification of attributes is done using output decision tree model. After such findings, a classification system model is proposed which depicts the stages of pre-processing, attribute selection, classification, factor identification, factor improvement and placement prediction. It may also be applied at any institute where placement prediction is required before-hand to increase the chances of campus selection irrespective of courses. The classification model can be applied to the problems related to student placement at technical institutions.

***Keywords*—** Educational Data Mining, Placement chance prediction, Classification Algorithms, Attribute selection, Student Performance.

I. INTRODUCTION

Data Mining is a field of study which enables us to discover the hidden knowledge in the form of patterns. Within this field, a new sub-field of study has emerged in recent years known as educational data mining (EDM). EDM is a time efficient process for discovering new relationships between datasets provided [10]. It concentrates in mining patterns from educational information systems like registration, admissions, course management and various other systems from school, colleges and universities. Many researchers have provided novel approaches for educational data mining thereby helping it make an independent research area [6]. Researchers working in this field try to help the educational institutes in better student management or to help student in improving their performance [12].

Educational data mining (EDM) focuses on finding the patterns from the information related to the students profile. These patterns helps in answering various sorts of questions related to the progress of students. Information related to the past performance of students from the students profile database provides an insight into the future performance in most of the cases. The performance of student is the most challenging factor in academic [5]. It has been identified that there is a weak learning approach in students and therefore the solution to this problem is being focused in EDM [5]. EDM helps in discovering, analyzing and predicting the probability of student’s future performance based on useful patterns discovered.

The main motivation factor for this research work is the opportunities of job in the form of campus selection for students. Educational Data Mining has gain importance in the recent year. Academic data analysis has been done till now but there arises a need for analyzing the placement of students so that their performance can be improved during campus selection. In this research work, the study is presented related to finding the useful patterns for campus selection. The aim of this study is to provide an effective model which will help the students to improve their performance during placement.

Academic institutions’ success is measured on the basis of the quality of education it provides to the students. Quality of education can be observed from the performance of students.

But there is inefficiency of existing system to evaluate and analyze the students' performance and progress. This is happening due to inaccuracy in predicting students' performance and because of lack of consideration of required vital factors affecting their performance. This task of prediction is challenging as academic database is usually very large [2]. Therefore, the model will be used in mining patterns from students' academic performance. It will provide solution to the problems related to the placement chance during campus selection.

Educational institutions primary objective is the enhancement of excellence of students in order to create human resources [5].Therefore, the students as well as the education managements can use the mined patterns obtained with the help of proposed model to predict the factors to be addressed. Hence, they can work towards improvement of those factors resulting in better placement chance of students. The objective of this study is to determine the factors which influence the selection procedure as well as the cause of selection up to some rounds of placement but rejection at the end.

EDM Classification helps in categorizing the students in order to improve their learning styles and inclination [1]. EDM can be implemented using various techniques namely decision tree, neural networks, K-nearest neighbor, Naive Bayes, Support Vector Machine and many others [1]. Classification is a supervised machine learning algorithm. It predicts the categorical attribute value using predicting attribute's values [8]. Classification discovers the prototype (or function) for recognizing important features of data classes or concepts. It outputs a model to predict class of object whose class label is unknown [1]. Naive Bayes classification algorithm outperforms complex algorithms as it works on a simple and intuitive concept by observing variables independently of each other [10]. Decision tree is most efficient algorithm for decision analysis and accurately identifies the most likely path to reach a goal for each leaf node is labelled with a class to which an instance should belong [8]. Neural Network can work excellent with noisy data and therefore it is used for many complex classification problems [6]. To implement these algorithms weka tool is used in this work. Weka is an open source data mining tool (written in java) which is widely used as it helps in solving real-world data mining problems [1].

Rest of the paper is organized as follows, Section II contains the related work in the area of education data mining, Section III contains the proposed work for the discovery of patterns during the student placement prediction using classification techniques, Section IV explains the methodology of the proposed work with flow chart, Section V describes the data collection process in detail, Section VI explains the preprocessing steps of data cleaning, attribute construction and transformation applied on collected student dataset, Section VII contains the comparative analysis of Naïve Bayes, Support Vector Machine and Neural Network algorithms with the experimental result of each, Section VIII explains the proposed classification and prediction model with figure, Section IX contains the results and discussion and Section X concludes research work with future directions.

II. RELATED WORK

C. Anuradha and T. Velmurugan [1] presented a study with the aim to predict how the student will perform in end semester of university examinations. For this purpose, the data is collected from three private colleges and analysis of the final year results of UG degree students is done. The analysis was done using data mining techniques namely decision tree algorithm C4.5, Bayesian classifiers, k Nearest Neighbor algorithm and two rule learner's algorithms namely OneR and JRip techniques. The study concludes that the tested classifier have sixty percent overall accuracy. Different classes have the classification accuracy which reveals that the predictions are good enough for the first class but worst for the distinction class.

Chaudhari K. P., Sharma R., et al. [2] discussed a study which aims at predicting GPA of students so as to improve their academic performance. The Naive Bayes, C4.5 Decision Tree of Data mining method and k-means Data Clustering algorithms are used to give a hybrid procedure for this purpose. The students are grouped into different segments using cluster analysis. This study also predicts the student year down and backlog based on the rule. It also compares different algorithms used for prediction.

Hamsa H. et al. [3] presented a study related to the performance of students for each subject during bachelor and master's degree. The prediction of performance has been done independently for each degree. Two algorithms namely decision tree and fuzzy genetic algorithms are used. It has been found that fuzzy genetic algorithm gives better result than decision tree algorithm for this study.

Kabra R.R. and Bichkar R.S. [4] described the decision tree models in this study for predicting the performance of Engineering Students in academic. For this study, Genetic Algorithm (GA) have been used to obtain better decision trees as it is one of the best search and optimization technique. GA and Evolutionary algorithms have been used to generate decision trees. Weka tool is used to find the accuracy and size of the decision tree models. The results shown that GA induced trees have less accuracy than J48 but improved results can be obtained by varying GA parameters and GA types as GA is a strong optimization technique.

Katare A. and Dubey S. [5] presented a study with objective of comparing 2-level classification with J48 (decision tree algorithm) and predicting student performance by classifying new grade classes. It includes the entropy and information gain based feature selection and then performing the normalization on student dataset. It uses 2-level Classification methods in which SVM is used in first step and KNN based classification in second step. It results in the accuracy, sensitivity, specificity of different grade classes thereby predicting student's performance. For this purpose, MATLAB 2012a tool is used.

Mueen A., Zafar B., Manzoor U. [6] also discussed a study which aims at predicting academic performance of students. It uses Multilayer Perceptron, Naive Bayes and C4.5 decision tree techniques. It analyses the student forum participation and academic record data using the classification models. It then predicts the student's performance based on those models. The result shows that the best prediction accuracy of 86 percent is obtained by Naive Bayes classifier among other classifiers.

M. Mayilvaganan and D. Kalpanadevi [7] presented a study that describes the improvement of techniques which play a role in analyzing the skill expertizes of students. These techniques are Prediction/Classification techniques. In this paper the comparative analysis of AODE, C4.5, Multi Label K-Nearest Neighbor and Naive Bayesian classifier algorithms is done to find the algorithm having better accuracy. To do the performance analysis of students, decision tree algorithm is used. All these analysis tasks have been carried out in weka.

Nichat A., Raut A.B. [8] presented a study which proposes a methodology to build a classification model that would classify a student's academic performance. This study includes a proposed systems for students to give some subjects related tests. The data obtained from the system is analysed using C4.5 decision tree algorithm and Generalized Sequential Pattern mining algorithm. Therefore, the variables are identified which distinguishes the student performance into "satisfactory" and "not satisfactory". This in turn helps in identifying the lack of students in particular field or subject.

Rana S. and Garg R. [9] have studied the clustering algorithms namely Hierarchical and K-means clustering to calculate the performance of students. A comparative study is presented in the paper where these two unsupervised algorithms are compared using weka tool. It has been found that K-means clustered instances are more effective than Hierarchical clustered instances. Also, K-means algorithm takes less time (0.12 seconds) to build a model for student performance evaluation.

Revathy P., Kalaiarasi P., Kavitha J., Madhumita D.A. [10] have collected the real time data of 60 students and served it as input to the clustering algorithm namely k-means clustering and Naive Bayes classifier is used in grouping the students who were willing to pursue higher studies. They used the RStudio tool for this purpose. From the analysis, the interest in various fields of each and every student can be identified. This is combined with the clustering output so as to predict the specialization a student can look for in his/her higher studies. Main aim is to provide a model which can guide students clearly in selecting their master's degree field.

Romero C. [11] described the educational data mining and its use in solving the educational data related problems. This paper explains about EDM to be a new research area which deals with computational methods to explore educational data. It also explains the types of Educational Environments, Educational data and different group of people in education field. Then it explains the tasks that are resolved through data mining techniques.

Saa A.A. [12] presented a study which aims at finding a model to predict the student's academic performance. The personal and social factors are taken for collecting the student data. The data is then analyzed using C4.5, ID3 and Naive Bayes classification algorithms. These algorithms help in building the classification models for finding interesting patterns. The main objective of this study is to find how the personal and social factors affect a student academic performance. All the data mining tasks were done using Rapid Miner and Weka tools.

III. PROPOSED WORK

As per the related work, only academic performance has been taken into account. The work has been done to predict end semester performance, Grade Point Average (GPA), student year down, Backlog, Degree Performance based on subjects and specialization a student can look for in his/her higher studies. Also the work has been carried out in state-of-art for classifying academic performance. The models given by state-of-art have less accuracy for example, C. Anuradha and T. Velmurugan [1] has concluded in their work that tested classifiers' model have 60 percent overall accuracy.

As an extent to the state-of-art, the next step to academic performance is placement performance which is another problem that is identified by the technical institutions. The placement performance makes use of the academic performance of students. Hence, in order to provide solution in this direction, a model is being proposed in this research work. Here, the proposed work aims at providing better accuracy model. It accomplishes the future scope of state-of-art in classifying and predicting placement chance of students.

The proposed work is to classify the student placement performance for discovering useful patterns and predict the placement chance. For this, a classification and prediction model is prepared. Firstly, manual forms and google forms are designed for collecting student basic, academic and placement details. The collected raw data is pre-processed using Ranker filter method, simple mathematical formulae and defined attributes' values with ranges. The mathematical formulae are created for attribute construction and are applied using Microsoft Excel application software on student dataset before applying filters on student dataset in weka software tool. The dataset is then transformed into appropriate form having nominal and numerical values for classification and prediction.

Further, the dataset of known class values is used to find the important attributes. With the help of identified attributes, Decision tree (J48 in WEKA) classification model is prepared and analyzed to find pattern of attributes (factors). Then identified attribute pattern is used to find factors that need improvement so as to improve student performance during campus placement. For the prediction of placement chance of students, three classification algorithms namely Naive Bayes, SVM (SMO technique in WEKA) and Neural Network are used. These algorithms are applied on the pre-processed student dataset and then are compared in terms of accuracy, precision, mean absolute error and recall. The algorithm with the best performance is selected for placement prediction using the student pre-processed data set with unknown class values. All the classification, prediction, experiment and analysis are done using the explorer interface, experimenter interface and knowledge flow environment interface of Weka 3.9 data mining tool. The proposed work has been described in detail in further chapters.

IV. METHODOLOGY

The methodology involves generating a database for the proposed work and using classification techniques and Weka Tool for mining useful patterns. The method is required to discover the patterns and analyze them to predict the placement chance of students. There are four main stages in this method, Data collection, pre-processing, classification and interpretation. Figure 1 depicts flow-chart of the methodology to achieve the objective of the proposed work.

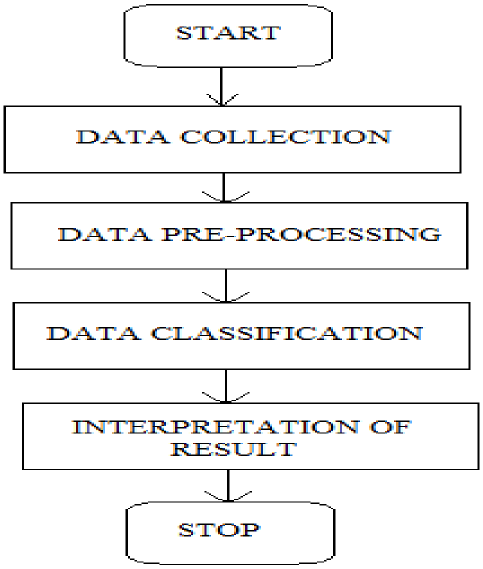


Figure 1. Methodology of the proposed Work

A. Data Collection: The academic progress information and placement progress information are collected from the technical institutes to form a large database. In this process, the data will be collected from the students of various courses in computer science department.

B. Data pre-processing: In this step, dataset is analyzed so that attributes are identified which have greater impact on the output variables. Weka provides several feature selection algorithms. Data will be balanced by solving the imbalanced problem of class instances. Data will be transformed by converting the format of the source data files into the destination data files.

C. Data classification: To predict placement chance data classification techniques will be used.

D. Interpretation of Result: The Weka tool will be used to visualize the patterns. Weka is an open source data mining tool developed at the University of Waikato, New Zealand, which is a free software available under the GNU General Public License.

V. DATA COLLECTION

Data collection is an important task for any case study. Student dataset is prepared by collecting details from final year B.Tech (CS/IT) and MCA students of KNIT, Sultanpur. The data has been collected from students of session 2017-18. The methodology is followed to generate the required database. The dataset includes the basic details, secondary details and placement details of students. The above information fields are required to help identify the factors having an effect on placement chance of students. The classification of the students is based upon attributes which are responsible for selection during placement.

Based on the details collected, the attributes are defined. The attribute's definition and values are formulated in the Table 1 below for data collection.

Table 1. Student attribute's definition and values based on data collection.

Attributes	Definitions	Values
SSC _Performance	Information about senior secondary academic excellence.	Excellent, good, average, poor
HSC/D/(HSC+D) _Performance	Information about academic excellence in higher secondary, diploma or both.	Excellent, good, average, poor
DEGREE _Performance	Information about the academic excellence in B.Tech. or during BSC/BCA and MCA taken together.	Excellent, good, average, poor
Backlog(s)	Back papers given	Extreme, high, medium, low
Prep_hrs	Average study hours per day for campus interview preparation.	Positive, neutral, negative
APTITUDE	The ability of solving problems faster with accuracy	Rich, poor
GD	Ability to work in a team by possessing leadership skills, listening skills, interpersonal skills and confidence.	Good, bad
TECHNICAL	Good knowledge about technical content	Rich, poor
MANAGEMENT	Ability to perform managerial roles	Good, bad
COMMUNICATION	Ability to express oneself and interaction skills	Active, medium, passive
Training	Information about attending the training programme which is a key aspect for enhancing the employability skills and personality development.	Yes, no
WorkshopORSeminar	Information about number of workshops or seminars attended	Excellent, very good, good, average, below average, poor
Vocational _Courses	Information about attending the application based study related to specific vocation or occupation.	Yes, no
Prog_Lang	Information about the number of programming languages known.	Excellent, good, average, poor
ACHIEVEMENTS	Information about the performance in extra-curricular and technical activities at school and college level	Excellent, good, average, poor
Coding_platforms	The online learning platforms for accessing one's programming strengths and weaknesses and thereby improving coding skills.	Coding platform groups(CPG) from CPG1 to CPG32
AreaofInterest	One or more core subjects of student's interest	Area of interest groups(AIG) from AIG1 to AIG74
Hobby	An interest or activity undertaken to utilize the spare time.	Reading, fishing, computer, gardening, writing, etc.
Interview _attempts	The information about the number of times a student appeared for campus interview	Positive, neutral, negative
Is_Placed	Finally, got selected in a company	Yes, no

VI. PREPROCESSING

Data is pre-processed to improve the efficiency and ease of the mining process. Data cleaning, data integration, data transformations and data reduction are the types of data pre-processing techniques.

The first step for pre-processing involves cleaning. It means finding and filling values, removing noisy data and correcting inconsistent data. The data collected through has no inconsistent data. The second step for pre-processing is data integration. The main task is the schema integration which is not required for this research work.

The third step is data transformation. It transforms the data into a form which appears to be appropriate for mining. It involves smoothing, aggregation, generalization, normalization and attributes construction. Attribute construction is a step required before generalization and normalization in this study. The data collected from students contained the attributes having semester marks. The semester marks are added and percentage of respective academic qualification is calculated. Similar process is carried out for SSC and HSC percentages. Therefore new attributes having percentage values are constructed from attributes having marks values. Also the APTITUDE, GD, TECHNICAL, MANAGEMENT and COMMUNICATION attributes are constructed from the company rounds given and rounds qualified attributes. These are basically ratios and hence in normalized form. The attribute ACHIEVEMENTS is constructed from school achievement, college extra-curricular achievement and college technical achievement attributes by calculating average of these attributes' values. The mathematical formulae for attributes construction are given as follows:

- Numerical value of APTITUDE, TECHNICAL, MANAGEMENT and COMMUNICATION (taken as decimal form of the ratio for the attributes for example: 0.67)

$$ratio = \frac{Number\ of\ Round(s) qualified}{Number\ of\ Round(s) given} \quad (1)$$

- Percentage of SSC, HSC, Diploma & BSC/BCA are calculated as

$$percentage(\%) = \frac{Total\ Marks\ Obtained}{Grand\ Total} \times 100 \quad (2)$$

- Percentage of MCA is calculated as

$$percentage(\%) = \frac{Sum\ of\ obtained\ Marks\ of\ all\ semester}{Sum\ of\ Total\ Marks\ of\ all\ semester} \times 100 \quad (3)$$

- Percentage of B.Tech is calculated as

$$percentage(\%) = \frac{a+b+c+d}{x} \times 100 \quad (4)$$

Where,
a=25%of obtained marks in first year
Or
a=0 (for diploma students)

b=50% of obtained marks in second year

c=75% of obtained marks in third year

d=0 (for marks upto third year)
Or
d=100% marks obtained in seventh semester

x= Total marks upto third year (i.e., 3000 for students without diploma and 2500 for diploma students)
Or
x=Total marks upto fourth year 7^{th} *semester (i.e., 4000 for students without diploma and 3500 for diploma students)*

- Numerical value of ACHIEVEMENTS (calculated as average and value converted in whole number form, for example: 2)

$$average = \frac{x+y+z}{3} \quad (5)$$

Where,
x= number of school achievements
y= number of college extra-curricular achievements
z= number of college technical achievements

The Backlog(s) attribute is constructed from the semester backlogs by counting the total number of backlog. The function (formulated in formula bar of MS-Excel application software) for calculating Backlog(s) attribute value is as follows:

$$f(x) = COUNTIF(X{:}Y, "YES") \quad (6)$$

Where,
x=cell reference under Backlog(s) attribute column in which value is counted
X=starting cell reference in the same row as x
Y=ending cell reference in the same row as x
X:Y=range of cell references in the same row as x

The percentage attributes HSC/D/(HSC+D) and DEGREE are constructed from old percentage attributes such as HSC, Diploma, B.Tech and MCA. This makes the attributes more specific having the percentages to be calculated at the same level. The functions (formulated in formula bar of MS-Excel application software) for calculating new attributes are as follows:

- Percentage of HSC/D/(HSC+D) is calculated as

$$f(x) = IF\ (X == "NIL", Y, (IF(Y == "NIL", X, ((X+Y)/2)))) \quad (7)$$

Where,
x=cell reference under HSC/D/(HSC+D) attribute column in which value is calculated
X= cell reference under HSC percentage attribute column in the same row as x
Y= cell reference under Diploma percentage attribute column in the same row as x

- Percentage of DEGREE is calculated as

$$f(x) = IF\ (X == "NIL", ((Y+Z)/2), X) \quad (8)$$

Where,
x=cell reference under DEGREE attribute column in which value is calculated
X= cell reference under B.Tech percentage attribute column in the same row as x
Y= cell reference under BSC/BCA percentage attribute column in the same row as x
Z= cell reference under MCA percentage attribute column in the same row as x

The student data collected for this study required generalization for mapping numeric attribute values to higher level concepts. That is, transforming the academic percentages calculated to academic performance values. Quartile range method is used for scaling the percentages to new labels. Numeric values of Backlog(s), Prep_hrs, APTITUDE, GD, TECHNICAL, MANAGEMENT, COMMUNICATION, WorkshopORSeminar, Prog_Lang, ACHIEVEMENTS and Interview_attempts attributes are transformed to high level nominal values. For this purpose, Interval scale of measurement with equal intervals is used for scaling the range of values to new labels. The attributes Training, Vocational_Courses, Coding_platforms, AreaofInterest, Hobby and class attribute “Is_Placed” required no transformation. The transformation process for dataset from numeric values to nominal values is depicted by the Table 2.

Table 2. Student dataset attributes with nominal values and corresponding numeric values (range) for pre-processing student dataset.

Attributes	Nominal Values	Numeric Values (range)
SSC_Performance	Excellent	82.6%-100%
	Good	65.1%-82.5%
	Average	47.6%-65%
	Poor	30%-47.5%, NIL (null value). (30% is the minimum required % to pass)
HSC/D/(HSC+D) _Performance	Excellent	82.6%-100%
	Good	65.1%-82.5%
	Average	47.6%-65%
	Poor	30%-47.5%, NIL (null value). (30% is the minimum required % to pass)
DEGREE _Performance	Excellent	82.6%-100%
	Good	65.1%-82.5%
	Average	47.6%-65%
	Poor	30%-47.5%, NIL (null value). (30% is the minimum required % to pass)
Backlog(s)	Extreme	3
	High	2
	Medium	1
	Low	0
Prep_hrs	Positive	9-12
	Neutral	5-8
	Negative	1-4
APTITUDE	Rich	0.51-1.00
	Poor	0.00-0.50
GD	Good	0.51-1.00
	Bad	0.00-0.50
TECHNICAL	Rich	0.51-1.00
	Poor	0.00-0.50
MANAGEMENT	Good	0.51-1.00
	Bad	0.00-0.50
COMMUNICATION	Active	0.67-1.00
	Medium	0.34-0.66
	Passive	0.00-0.33
Training	Yes	-----
	No	-----
WorkshopORSeminar	Excellent	6
	Very good	5
	Good	4
	Average	3
	Below average	2
	Poor	<=1
Vocational _Courses	Yes	-----
	No	-----
Prog_Lang	Excellent	6,7
	Good	4,5
	Average	2,3
	Poor	0,1
ACHIEVEMENTS	Excellent	7,8
	Good	5,6
	Average	3,4
	Poor	<=1,2
Coding_platforms	Coding platform groups(CPG) from CPG1 to CPG32	-----
AreaofInterest	Area of interest groups(AIG) from AIG1 to AIG74	-----
Hobby	Reading, fishing, computer, gardening, writing, etc.	-----
Interview_attempts	Positive	3
	Neutral	2
	Negative	<=1
Is_Placed	Yes	-----
	No	-----

VII. EXPERIMENT AND ANALYSIS OF ALGORITHMS

For the purpose of analysing the dataset, four classification algorithms have been used. They are decision tree (J48), Naive Bayes, Neural Network (Multilayer perceptron) and SVM (SMO i.e. Sequential Minimal Optimisation). These are used to derive the model for which training data provides the class label. Decision tree algorithm is used in further chapters for finding patterns in the student dataset. Therefore, only Naive Bayes, Neural Network and SVM are analysed and compared in this chapter to find the best algorithm for student placement chance prediction. Weka 3.9 tool is used for analysis purpose.

A. *Naive Bayes algorithm*

Naive Bayes technique has been applied on the student datasets. The analysis is carried out using the knowledge flow through the various design tools in weka as shown by Figure 2 below.

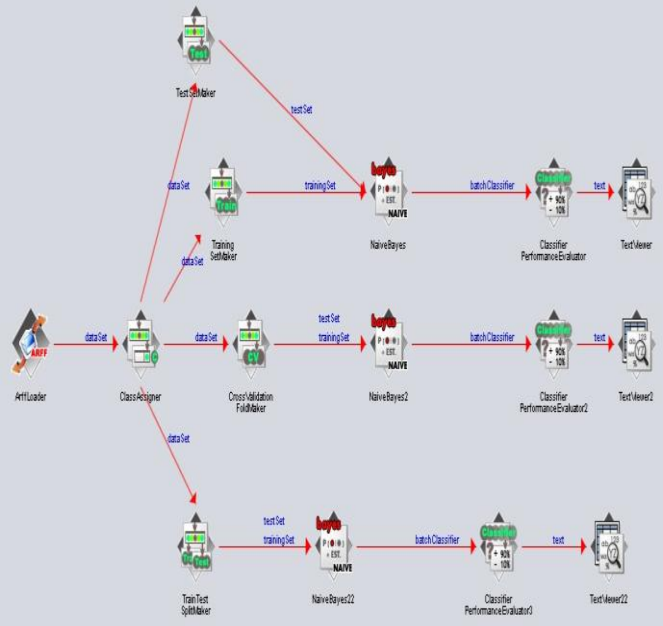


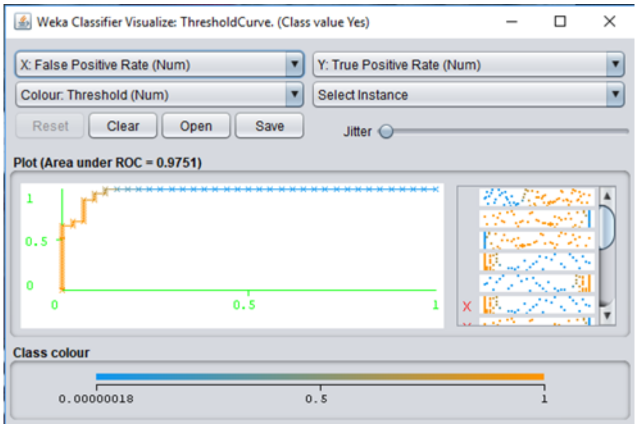
Figure 2. Showing the knowledge flow for the analysis of Naive Bayes algorithm

The various performance criteria are calculated using different test modes in weka, see Table 3.

Table 3. Performance criteria of various test modes applied on student dataset.

Test mode	Accuracy	Mean absolute Error	Precision (average)	Recall (average)
10-fold cross validation	95%	0.0731	0.954	0.950
Split 66.0% train, remainder test	88.8889 %	0.0984	0.906	0.889

The ROC (Receiver Operating curve) and AUC (Area under ROC) for the Naïve Bayes algorithm applied on student dataset are shown in Figure 3 below.



Weka Classifier Visualize: ThresholdCurve. (Class value Yes)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9751)

Figure 3. Showing AUC and ROC of Naïve Bayes algorithm when applied on student dataset

From the Figure 3, it is observed that the value of AUC is 0.9751 and the ROC curve is towards the top left corner. Hence it is more close to value 1.0000 which indicates that this algorithm performs well in a cost sensitive manner for student dataset.

B. *SVM (Support Vector Machine) algorithm*

SMO (Sequential Minimal Optimisation) technique has been applied on the student dataset for this study. The analysis is carried out using the knowledge flow through the various design tools in weka as shown by Figure 4 below.

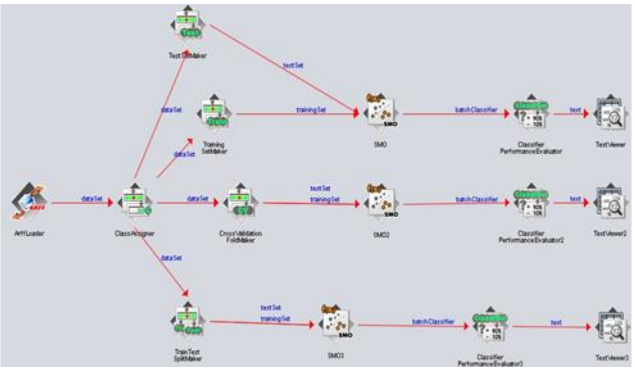


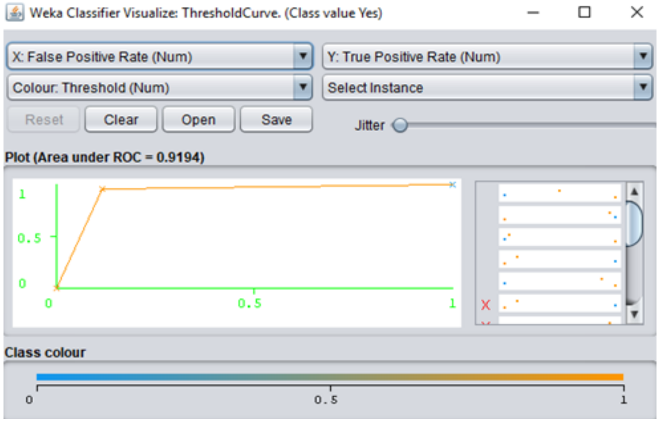
Figure 4. Showing the knowledge flow for the analysis of SVM algorithm

The various performance criteria are calculated using different test modes in weka, see Table 4.

The ROC (Receiver Operating curve) and AUC (Area under ROC) for the SVM algorithm applied on student dataset are shown in Figure 5 below.

Table 4. Performance criteria of various test modes applied on student dataset

Test mode	**Accuracy**	**Mean absolute Error**	**Precision (average)**	**Recall (average)**
10-fold cross validation	93.75 %	0.0726	0.939	0.938
Split 66.0% train, remainder test	92.5926 %	0.0786	0.934	0.926



Weka Classifier Visualize: ThresholdCurve. (Class value Yes)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9194)

Figure 5. Showing AUC and ROC of SVM algorithm when applied on student dataset

From the Figure 5, it is observed that the value of AUC is 0.9194 and the ROC curve is towards the top left corner. Hence it is more close to value 1.0000 which indicates that this algorithm performs well in a cost sensitive manner for student dataset.

C. *Neural Network (NN) Algorithm*

The Multilayer perceptron technique of Neural Network algorithm is used for data classification in weka. It has been applied on the student dataset for this study. The analysis is carried out using the knowledge flow through the various design tools in weka as shown by Figure 6 below.

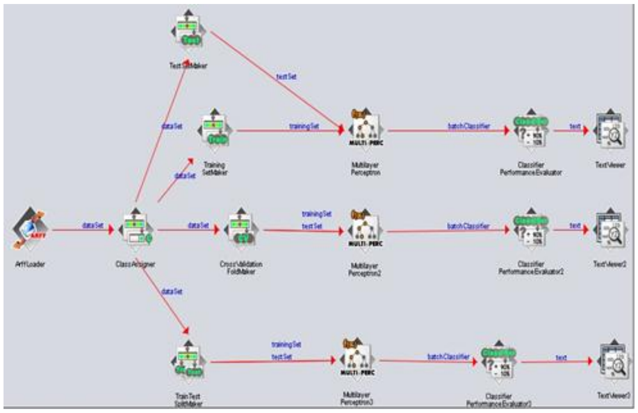


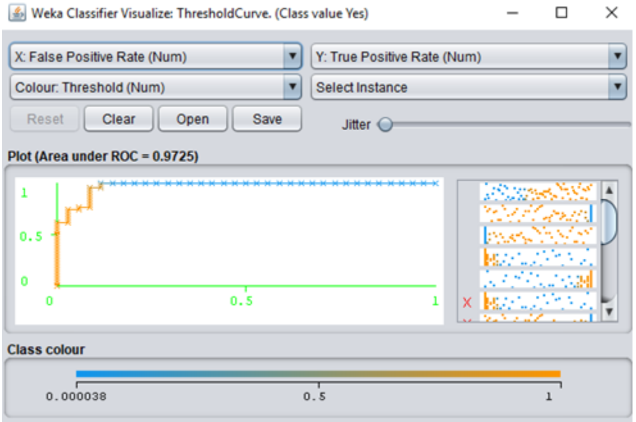
Figure 6. Showing the knowledge flow for the analysis of Neural Network algorithm

The various performance criteria are calculated using different test modes in weka, see Table 5.

Table 5. Performance criteria of various test modes applied on student dataset.

Test mode	Accuracy	Mean absolute Error	Precision (average)	Recall (average)
10-fold cross validation	92.5 %	0.075	0.926	0.925
Split 66.0% train, remainder test	92.5926 %	0.0741	0.934	0.926

The ROC (Receiver Operating curve) and AUC (Area under ROC) for the Neural Network algorithm applied on student dataset are shown in Figure 7 below.



Weka Classifier Visualize: ThresholdCurve. (Class value Yes)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9725)

Figure 7. Showing AUC and ROC of Neural Network algorithm when applied on student dataset

From the Figure 7, it is observed that the value of AUC is 0.9725 and the ROC curve is close to the top left corner. Hence it is close to value 1.0000 which indicates that this algorithm performs well in a cost sensitive manner for student dataset.

D. Comparison of Naive Bayes, SVM and Neural Network algorithms

In this section, the performance of Naive Bayes, SVM and Neural Network are compared so as to select the best algorithm for predicting placement chance of students.

Table 6. Comparison of performance criteria of Naive Bayes, SVM and Neural Network classification algorithms

Performance Criteria	**Naive Bayes algorithm**	**SVM algorithm**	**Neural Network algorithm**
Accuracy	95%	92.5926 %	93.75 %
Mean Absolute Error (MAE)	0.0731	0.0741	0.0726
Precision	0.954	0.934	0.939
Recall	0.950	0.926	0.938

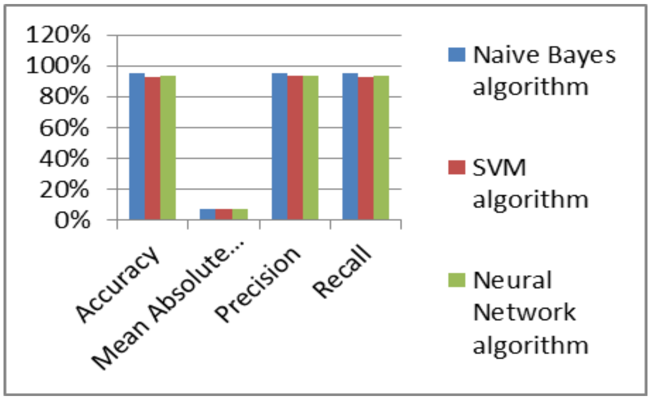


Figure 8. Plot for comparison of classification algorithms used.

Hence, from Table 6 and Figure 8, it is found that Naive Bayes algorithm works best to classify with maximum accuracy of 95%, a reliable evaluation result. It has mean absolute error of 0.0731, higher precision of 0.954 and higher sensitivity (recall) of 0.950.

The comparison of ROCs of Naive Bayes, SVM and Neural Network algorithms is carried out using the knowledge flow through the various design tools in weka is shown by Figure 9 below.

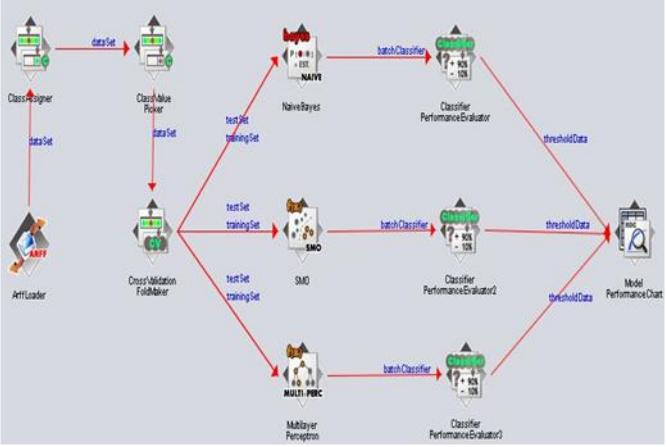


Figure 9. Knowledge flow for comparison of ROCs of classification algorithms used

Table 7. Comparison of Area under ROC of different classifiers for offline dataset (D1) and student dataset (D2)

Classifiers	**AUC values**
Naive Bayes Classifier	0.9751
SVM Classifier	0.9194
Neural Network Classifier	0.9725

From Table 7, the best value observed for AUC is of Naive Bayes classifier i.e. 0.9751. Through the observations of performance criteria, AUC and ROC analysis, Naive Bayes is found to be a good classifier among the three classifiers used. Hence, Naive Bayes is used in the model for student placement prediction.

VIII. CLASSIFICATION AND PREDICTION MODEL

For this study, it is observed that Naive Bayes algorithm and Decision tree algorithm (J48) are useful algorithms. The classification model is built using these two algorithms. The model is used to identify factors responsible for selection/rejection of students as well as predicting the placement chance using the patterns of identified attributes. Figure 10 shows the suggested model for this case study.

The first stage of Raw Data is the student database formed by collecting student details. It is then pre-processed using pre-processing techniques. Now, the processed data may or may not have the class value but the class attribute is defined in the pre-processing stage. In the Processed Student Data stage, the data file (for this study is in .arff file format) is divided into two sets based on the class value known and class value unknown. If the class value is known then the dataset is sent to the attribute selection stage. If the class value is unknown, it is sent to the Naive Bayes algorithm stage where Naïve Bayes algorithm is applied on the dataset to predict the class value.

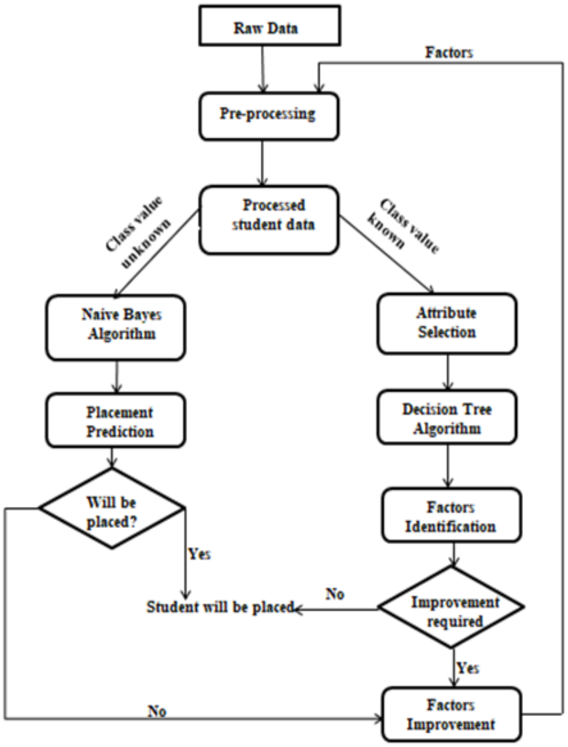


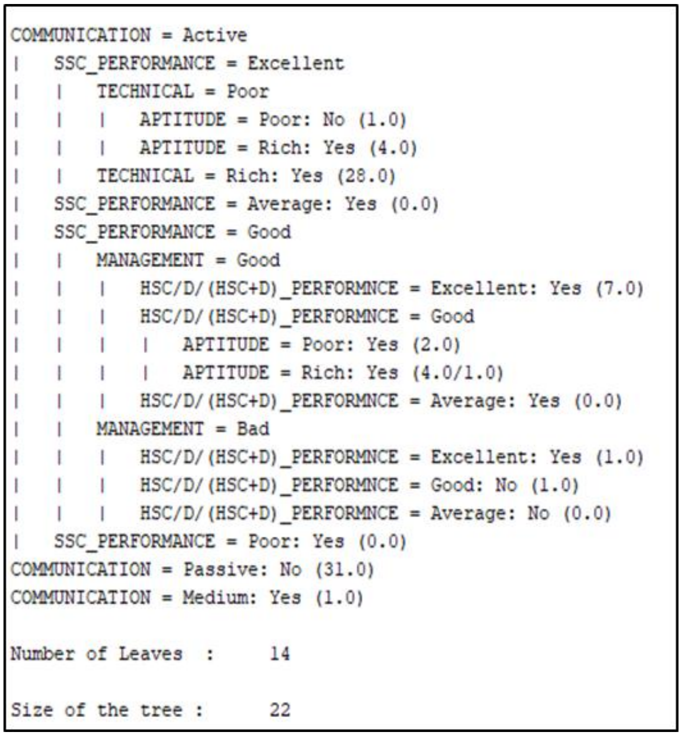
Figure 10. Classification and Prediction model for Improving Student Placement

In the attribute selection stage, the attribute evaluator with the search method is applied to find the significant attributes. Then, the unwanted attributes are removed using appropriate filters and only the selected attributes are taken for the classification. The dataset file is then saved with updated changes and sent to the next stage. The next stage is the application of the decision tree algorithm, to build a tree that describes the role of input factors in placement. The decision tree algorithm technique such as J48 (in weka) is used. This tree model is sent to the next stage of Factor Identification. The attributes from the roots to the leaf nodes are analysed to identify those attributes (or factors) which have significant higher gain ratio than other attributes. Thus, it analyses and identifies the factors which require improvement. Then a decision is made whether there are any factors for improvement. If decision is no, then it is found that the student will be placed if it performs with the same attributes (or factors) value. If decision is yes, then those factors are sent to the Factors Improvement stage. In the Factor Improvement stage, the institute management may help the students in improving their performance in those factors that affects their placement. Then the improved factors are then sent to the previous stage of pre-processing for forming the new dataset and doing it's pre-processing. Thus, the cycle continues until no factor improvement is required.

In the case of prediction when the class value is unknown, Naive Bayes algorithm is used. Then the output from the Naive Bayes algorithm stage is used for placement prediction i.e., whether the student will be placed or not. If decision is yes, then the student will be placed and if no, then again factor improvement is required.

IX. RESULTS AND DISCUSSION

The purpose of this study is to discover useful patterns (set of factors) in student placement and predict the placement chance. For this purpose, the useful patterns discovered in the student placement dataset using the decision tree algorithm (J48) are given in the figure 10.1 below.



```
COMMUNICATION = Active
|   SSC_PERFORMANCE = Excellent
|   |   TECHNICAL = Poor
|   |   |   APTITUDE = Poor: No (1.0)
|   |   |   APTITUDE = Rich: Yes (4.0)
|   |   TECHNICAL = Rich: Yes (28.0)
|   SSC_PERFORMANCE = Average: Yes (0.0)
|   SSC_PERFORMANCE = Good
|   |   MANAGEMENT = Good
|   |   |   HSC/D/(HSC+D)_PERFORMNCE = Excellent: Yes (7.0)
|   |   |   HSC/D/(HSC+D)_PERFORMNCE = Good
|   |   |   |   APTITUDE = Poor: Yes (2.0)
|   |   |   |   APTITUDE = Rich: Yes (4.0/1.0)
|   |   |   HSC/D/(HSC+D)_PERFORMNCE = Average: Yes (0.0)
|   |   MANAGEMENT = Bad
|   |   |   HSC/D/(HSC+D)_PERFORMNCE = Excellent: Yes (1.0)
|   |   |   HSC/D/(HSC+D)_PERFORMNCE = Good: No (1.0)
|   |   |   HSC/D/(HSC+D)_PERFORMNCE = Average: No (0.0)
|   SSC_PERFORMANCE = Poor: Yes (0.0)
COMMUNICATION = Passive: No (31.0)
COMMUNICATION = Medium: Yes (1.0)

Number of Leaves  :     14

Size of the tree :      22
```

Figure 11. Showing the useful pattern of attributes discovered in student placement dataset

From the Figure 11, the resultant patterns give the useful information which is summarized in Table 8 below.

Table 8. The summary of useful information obtained from the discovered patterns.

Attributes	Values	Is Placed (YES/NO)
COMMUNICATION	Medium, Active	YES
	Passive	NO
TECHNICAL	Rich	YES
	Poor	NO
MANAGEMENT	Good	YES
	Bad	NO

HSC/D/(HSC+D) _PERFORMANCE	Excellent, Good	YES
	Average, Poor	NO
APTITUDE	Rich	YES
	Poor	NO
SSC_PERFORMANCE	Excellent, Good	YES
	Average, Poor	NO

In order to predict the placement chance, classification algorithms Naive Bayes, SVM and Neural Network are applied on the student placement datasets and the performance of the classifiers are compared. As a result of comparison, it is found that the Naïve Bayes outperforms the other two classifiers with the following resultant values in Table 9 below.

Table 9. The various performance criteria with resultant respective values of Naive Bayes Classifier

S. No.	Performance criteria	Values
1	Accuracy	95.99%
2	Mean Absolute Error (MAE)	0.0547
3	Precision	0.9625
4	Recall	0.9600
5	Area under ROC	0.9827

Therefore, the Naive Bayes classifier is used for the prediction of placement chance.

In order to help the technical institute managements to improve the performance of students during placement process, a classification and prediction model is also proposed. It helps to discover the useful patterns such as the useful patterns discovered in this case study as given in Figure 11 and resultant information in Table 8. Also, to predict the placement chance and thereby improving the factors (attributes) that does not meet the required values or may result in rejection during placement. The various stages and the respective tasks proposed are summarized in Table 10 below.

Table 10. The various stages and respective tasks proposed for the classification and prediction model

S. No.	Model Stages	Tasks
1	Raw Data	Collection of student details to form student database.
2	Pre-processing	Application of pre-processing techniques on student database.
3	Processed Student data	Form two datasets based on the class value known and class value unknown
4	Attribute Selection	Application of attribute evaluator and respective search method to find significant attributes
5	Decision Tree Algorithm	Build the decision tree classification model
6	Factor Identification	Identify the attributes (or factors) having significant higher gain ratio with respect to the class
7	Improvement Required	Making a decision whether there are any factors for improvement or not.
8	Factor Improvement	Taking actions to improve student performance in those factors that affects their placement.
9	Naive Bayes Algorithm	Build the prediction classifier
10	Placement Prediction	Predict the placement chance

X. CONCLUSION and Future Scope

Campus selection i.e. placement of students is one of the important topic of concern for the management of technical institutions. The model presented by this study can efficiently predict the placement chance of a student and may help in improving performance during campus selection. The data collection process is carried out using a survey of student's placement performance of Computer Science department of KNIT technical institute. The data set is pre-processed to transform it into the useful form so that important pattern of attributes can be identified. Decision tree algorithm (J48) creates an inverted tree structure of attributes giving a pattern for classification of a student data instance. Hence, the resulted useful pattern found by this study is the set of attributes which includes COMMUNICATION, TECHNICAL, MANAGEMENT, HSC/D/(HSC+D)_PERFORMANCE, APTITUDE and SSC_PERFORMANCE.

The most efficient algorithm is Naive Bayes for the data set collected for this study because the efficiency of Naive Bayes classifier is observed to be 95.99 percent, which is best among other classifiers used. The students as well as the education managements can use the model to improve students' performance during placement process by working towards the refinement of the factors (attributes) in which the students are lacking. The work presented in this paper is beneficial to institute management because it provides a model that can identify the cause of selection and rejection of students during placement. Hence, removing the cause of rejection of students may help to increase more placements at the institute.

Only the classification algorithms are considered for this research work, no cluster analysis or association rule are used. It is simply a case study so modifications are not made in classification techniques. Student dataset used is small and confined to one department of the technical institute.

The future work can include applying other data mining techniques such as association rule mining and clustering on the student dataset. Next, the dataset can be increased by performing this study on the entire institute's student dataset. It can be done by collecting data from various other departments or courses. The bigger dataset may reveal some other interesting patterns that are not mined in this case study. For further research, new training sessions can be conducted and some tests can be scheduled to check student's performance. These tests can be additional useful attributes. Similarly, interviews can be conducted by the teachers of the institute itself as a mock interview for students. The result of these interviews can be another attribute for student performance analysis to improve placement chance.

REFERENCES

[1] C. Anuradha, T. Velmurugan, "*A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance",* International Journal of Science and Technology, Vol. 8, Issue 15, 2015.

[2] K.P. Chaudhari, R.A. Sharma, S.S. Jha, R.J. Bari, "*Student Performance Prediction System using Data Mining Approach*", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 3, 2017.

[3] H. Hamsa, S. Indiradevi, J.J. Kizhakkethottam, "*Student Academic Performance Prediction Model Using Decision tree and Fuzzy Genetic Algorithm*", Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science and Technology, 2016.

[4] R.R. Kabra, R.S. Bichkar, "*Student's Performance Prediction Using Genetic Algorithm*", International Journal of Computer Engineering and Applications, Vol. VI, Issue III, 2014.

[5] A. Katare, S. Dubey, "*A Comparative Study of Classification Algorithms in EDM using 2 Level Classification for Predicting Student's Performance*", International Journal of Computer Applications, Vol. 165, Issue 9, 2017.

[6] A. Mueen, B. Zafar, U. Manzoor, "*Modeling and Predicting Students' Academic Performance Using Data Mining Techniques*", I. J. Modern Education and Computer Science, 2016.

[7] M. Mayilvaganan, D. Kalpanadevi, "*Comparison of Classification Techniques for predicting the performance of Students Academic Environment*", International Conference on Communication and Network Technologies, 2014.

[8] A. Nichat, A.B. Raut, "*Predicting and Analysis of student Performance Using Decision Tree Technique*", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5,Issue 4, 2017.

[9] S. Rana, R. Garg, "*Evaluation of Student's Performance of an Institute Using Clustering Algorithms*", International Journal of Applied Engineering Research, Vol. 11, 2016.

[10] P. Revathy, P. Kalaiarasi, J. Kavitha, D.A. Madhumita, "*Data Mining Approach for Suggesting Higher Education Courses Based on Student's Performance*", International Journal of Science and Technoledge, Vol. 3, Issue 3, 2015.

[11] C. Romero, "*Educational Data Mining: A Review of the State of the Art*", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 40, 2010.

[12] A.A. Saa, "*Educational Data Mining and Students' Performance Prediction*", International Journal of Advanced Computer Science and Applications, Vol. 7, 2016.

[13] R. Anupriya, P. Saranya, R. Deepika, "Mining Health Data in Multimodal Data Series for Disease Prediction", International Journal of Scientific Research in Computer Science and Engineering, Vol. 6, Issues 2, pp. 96-99, 2018.

[14] M. Fernandes, "Data Mining: A Comparative Study of its Various Techniques and its Process", International Journal of Scientific Research in Computer Science and Engineering, Vol. 5, Issue 1, pp. 19-23, 2017.

Authors Profile

Miss Mekhla Shukla pursued Bachelor of Technology from Feroze Gandi Institute of Engineering and Technology, Raebareli, India in 2016. She is currently pursuing M.Tech. from Kamla Nehru Institute of Engineering and Technology (KNIT), Sultanpur, India. Her research interest is in Data mining.

Dr. Anil Kumar Malviya is currently working as a Professor in Computer Science and Engineering Department at Kamla Nehru Institute of Technology (KNIT), Sultanpur, India. He received his B.Sc.(Hons.) and M.Sc. both in Computer Science from Banaras Hindu University (BHU) Varanasi respectively in 1991 and 1993 and Ph.D. degree in Computer Science from Dr. B.R. Ambedkar University, Agra in 2006. He is a life member of Computer Society of India (CSI) and Indian Society for Technical Education (ISTE). He has published about 55 papers in International/National Journals, conferences and Seminars. His research interests are Software Engineering, Data Mining and Cryptography and Network Security. He has experience of approximately 20 years in Teaching.