A Comparative Study of Classification Algorithms in EDM using 2 Level Classification for Predicting Student's Performance

Ankita Katare
Department of Computer Science and Engineering
RITS, Bhopal-INDIA

Shubha Dubey
Department of Computer Science and Engineering
RITS, Bhopal-INDIA

ABSTRACT
In higher education the performance of students is a most challenge work day by day in academic as well as in other curricular activities. As they all know that internet technology is growing as much as faster, but the learning approach of students are not up to the mark. The emerging research community which helps to find the solution to the said problem is Educational Data Mining. In present scenario, the huge students' data is stored in educational database. That type of database contains widely open or secret information to improve student performance. In our proposed work, we will have tested it on reputed dataset, which can be downloaded from a well known organization UCI repository and dataset name is student-mat.csv. This work has been investigated the process of classification of plethora of student's data. Classification plot data into pre-determined groups of classes. It is often mentioned to as supervised learning because the classes are determined before analyzing the data. The work will to be divided into two parts. The first part will be the entropy based feature selection, after that classification process has to be performed. For the classification, we would have used 2 level classification method i.e, SVM and KNN. Later than observe the performance prediction of students based on parameters like accuracy, sensitivity, specificity of proposed method and is to be compared with some previous methods results.

Keywords
Data Mining, EDM, Classification Algorithms, Entropy, Performance Prediction.

1. INTRODUCTION
For higher education institutions whose objective is to provide to the enhancement of excellence of higher education, the accomplishment of creation of human resources is the subject of a continuous study. Therefore, the prediction of students' success is critical for higher education institutions, because the quality of teaching process is the capability to meet students' needs. In this sense on a regular basis important data and information are gathered, and they are considered at the suitable authorities and standards in order to sustain the quality.

Data mining is measured to be a new standard, but due to its implication in decision making, it has been successfully applied to a variety of domains including education. The data gathered from various applications require proper method of extracting knowledge from big repositories for making better decision. Knowledge discovery in databases (KDD), also called data mining, resulting useful information from large amount of data is its purpose. The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [1]. Data mining and knowledge discovery applications have got a rich center due to its importance in decision making and it has become an essential component in various organizations. Such promising fields of Machine learning, Statistics, Databases, Pattern Reorganization, Computation capabilities and Artificial Intelligence the data mining techniques have been used.

There are rising research interests in using data mining in education. Educational Data Mining is this rising field, deals with mounting methods that discover knowledge from data available in educational environments. Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor and all these techniques used by Educational Data Mining. Some classification algorithms mostly used in EDM are as follows:

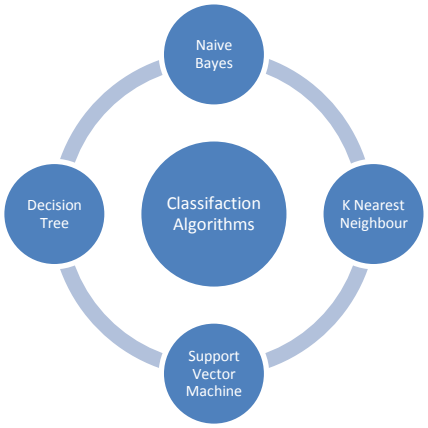

Figure 1: Different Classification Algorithms

A tree like structure is a decision tree in which between a number of alternatives each branch node represents a choice, and each leaf node represents a decision. By gaining information decision trees are commonly used for the purpose of decision –making. Decision tree is for users to take actions as it starts with a root node. From this node, According to decision tree learning algorithm users split each node recursively. A decision tree in which each branch represents possible circumstances of decision and its outcome is the final result. The most widely used decision tree learning algorithms are ID3, ASSISTANT and C4.5 [18].

SVM is a differentiate classifier which is examined by a separating hyperplane. SVM create a Hyperplane, which is used for classification and regression tasks. It determines closest data vectors called support vectors (SV), to the decision detention in the training set and a given new test vector can be separated by using only these closest data vectors [19].

Both distance metric ("nearest") and number of neighbors can be changed is a nearest-neighbor classification object,. Using the predict method the object classifies new observations. The data used for training contains by object, so resubstitution predictions can be computed. KNN is an *non parametric* algorithm. When you say a technique is non parametric, it means that on the underlying data distribution it does not make any assumptions. As in the real world , this is useful because the typical theoretical assumptions made (eg gaussian mixtures, linearly separable etc) are not followed by most of the practical data. It is also a lazy algorithm.

Many kinds of knowledge can be discovered using these techniques such as association rules, classifications and clustering. The recognized knowledge can be used for prediction regarding enrolment of students in a particular course, isolation of traditional classroom teaching model, detection of unfair means used in online examination, detection of irregular values in the result sheets of the students, prediction about students‟ performance and so on[2].The main purpose of this paper is to use data mining methodologies to analyze students‟ performance. To analyze the student performance data mining provides many tasks. For student performance prediction as there are many techniques that are used for data classification. In this process, Educational Dataset embedded into MATLAB simulation tool and results are interpreted and evaluated using SVM and KNN classifiers. Educational Dataset consist of attributes contains information to predict their class result using classification technique.

The rest of the text is ordered as follows: section II signify background study and related work; section III signify the problem definition; the proposed work describe in section IV; section V present the experimental result; finally, the conclude the survey work and future work being in section VI.

2. RELATED WORK

For analysis of data available at educational institutions data mining methods and tools are implemented, defined as Educational Data Mining (EDM) is a relatively new field in the data mining research. Much attention needed to be paid on student performance, student behavior analysis, faculty performance and result of this on student final performance. Many research papers work in educational Data mining are by:

Three different data mining classification algorithms (Naïve Bayes, Neural Network, and Decision Tree) were used on the dataset. The prediction performance of three classifiers are measured and compared. A result shows that Naïve Bayes classifier outperforms other two classifiers by obtaining the overall prediction accuracy of 86%. This study will help teachers to improve student academic performance. In this data mining techniques is to be applied to analyze academic performance of student's based on their academic record and forum participation. Ahmed Mueen et.al. [3]

Predicting student's performance in placement in an education system has become more difficult due to huge amount of data and inaccurate data with uncertainty in educational databases. Some of the existing methodologies and their problem for the student analysis have been discussed in this survey. In this paper, a survey is made for college activities were students are needed to be focused to improve their placements and different ways to find the solution. Where many authors have proposed a method to improve the performance of staff but there is no related work for student's placement activity. S. Indhu Priya et.al.[4].

Ruhi R. Kabra et.al.[5] This study is about to obtain the decision tree models that predict the academic performance of the engineering students in contact education system. Genetic algorithm is a great search and optimization technique that has shown assurance in obtaining good decision trees. Decision trees are generated using greedy as well as evolutionary algorithms. The results are discussed with respect to the accuracy and size of the tree induced using genetic algorithm and J48 (from WEKA). In this it shows that the GA induced trees observe the accuracy slightly less than J48. However GA is a powerful optimization technique and it is quite possible to obtain further improvements in result by using different GA parameters and GA types.

Important clustering algorithms are discussed after that which can be applied to calculate student's performance. By applying Hierarchical Clustering and K-Means Clustering Algorithms the performance of students is calculated. K-Means and Hierarchical clustering an Unsupervised Learning Algorithms are discussed here. Using WEKA further a comparison is made between two unsupervised algorithms. According to the output, clustered instances in Hierarchical clustering are less effective than in K-Means. For the given data set the time taken to build a model in K-Means is less (0.12 seconds) than that in hierarchical clustering (0.49 seconds). Shiwani Rana et.al.[6]

Prediction of student's performance in bachelor's and master's degree for each subject was done independently using decision tree algorithm and fuzzy genetic algorithm. Result from fuzzy genetic algorithm gives more satisfactory result as compared to decision tree algorithm Hashmia Hamsa et.al.[7]

This work presents an analysis of final year results of UG degree students using data mining technique. In this classification techniques are applied to prediction of the performance of students in university examinations. Mainly, the decision tree algorithm C4.5 (J48), Bayesian classifiers, k Nearest Neighbor algorithm and OneR and JRip two rule learner's algorithms are used for examining the performance of students as well as a model of student performance predictors to be developed. The results showed to be satisfactory. C. Anuradha et.al.[8]. Amirah Mohamed Shahiri et.al.[9] In this they focus on how the prediction algorithm can be used to identify the most important attributes in a student's data . It could bring the benefits and impacts to students, educators and academic institutions. It provide an overview on the data mining techniques that have been used to predict students performance. Previous work has been reviewed by them on predicting student's performance with various analytical methods.

3. PROBLEM STATEMENT

It is challenging to pre-process the dataset which contains various attributes. In inadequacy of application of proper data processing technique, the classification process will not be satisfactory. Handling missing and continuous attributes J48 need entire data to fit in memory. What this means is that it does not use the training data points to do any *generalization*. Lack of generalization means that J48 keeps all the training data. More exactly, all the training data is needed during the testing phase. This is in distinction to other techniques like SVM where All non support vectors without any problem can be discarded. J48 – makes decision based on the entire training data set (in the best case a subset of them). In this

survey we analysed the data and predict student performance to overcome such limitation of J48 algorithm, we further proceed this work using other data mining techniques in analysing data using other trends [12]. In our work after the entropy based feature selection, 2 level classification i.e, SVM and KNN classification are used to evaluate and compare the performance prediction based on parameters like accuracy, sensitivity and specificity. The features include both problem content and a performance feature, which predicts that the similar records must be associated to the same class and performance of the records must be homogenous. The students are required to learn each data which corresponds to a class.

4. PROPOSED WORK

Here read student learning dataset [17] and pre-process student dataset by converting into xlxs format. As shown in figure (2) the Entropy based feature selection is to be done by finding Entropy and information gain of each attribute separately. Information gain measure is to be calculated using decision tree algorithm. Entropy is commonly used to calculate impurity. Information content is more when impurity is higher. Information Gain is metric for how well one attribute Ai classifies the training data. Decision Tree recursively partitions the training set Parameter which best classifies data is Entropy (H). Entropy is a good measure of the information carried by an ensemble of events. Entropy of set S is denoted by H(S). If S=Sample of n training events and Pi is the probability of occurrence of event, then entropy is given by:

$$H(S) = \sum_{i=1}^{n} - P_i \log_2 P_i$$

For each attribute calculate the information gain. Information gain is a statistical quantity measuring how well an attribute classifies the data.we have calculated the information gain (Gain(S, Ai)) for each attribute using Algorithmic Approach and in the end attribute with the highest information gain will be chosen for decision-making. *Sv* is the subset of *S* for which attribute *A* has value *v*.

$$Gain(S, A_i) = H(S) - \sum_{v \epsilon Values(Ai)} P(A_i=v)\, H(S_v)$$

Information gain is our metric for how well one attribute Ai classifies the training data. The process is normalized using min-max algorithm to provide linear transformation on original range of data. This specifically fit the data by finding new range from an existing one range. Here we use new approach of level 2 classification using SVM in level 1 classification and KNN in level 2 classifications. Start level 1 classification process on training and test data using Support vector Machine classification. After prediction, now predicted classes will passed for training. On training data KNN based Level 2 classification will be applied. Classified result of new classes Gh1, Gh2, Gh3 and Gh4 will be calculated. Performance of student's will be predicted by calculating standard parameters like accuracy, sensitivity and specificity of all four classes. All these parameters are examined and compared with J48 classification algorithm.

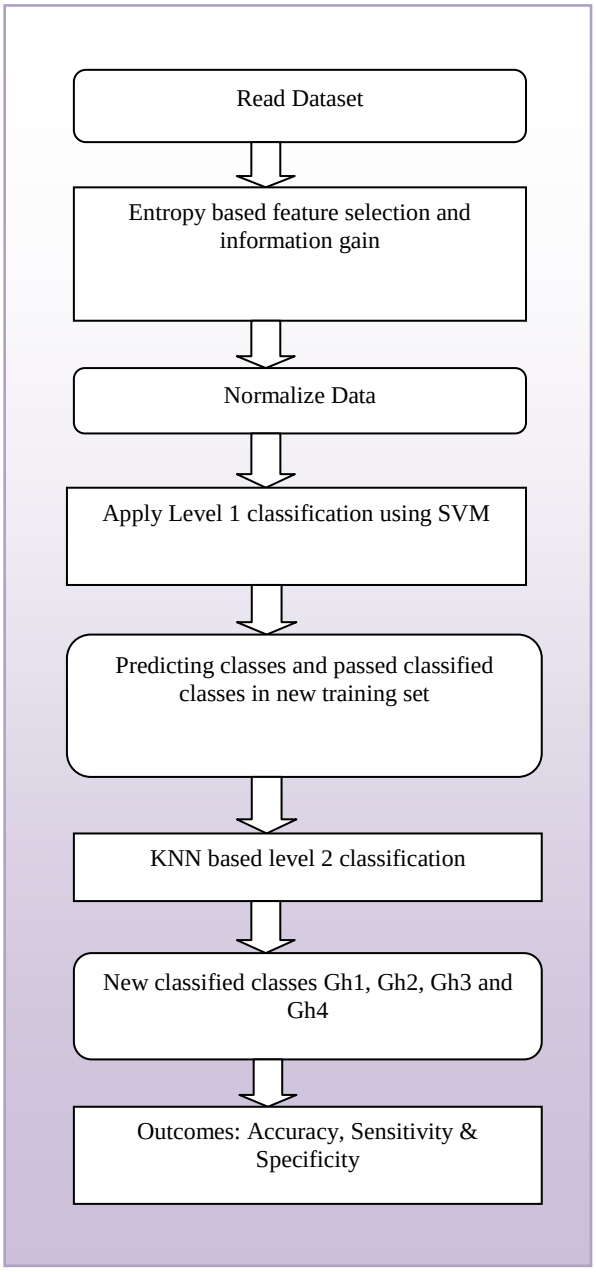

Figure 2: Data Flow Diagram

5. EXPERIMENTAL RESULT

All experiments are done in java using MATLAB 2012a tool. This tool supports so many programming languages. MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation. As compared to traditional programming languages, such as C, C++, and Fortran problems you can solve technical computing problems faster using MATLAB. Signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology is a wide range of applications where MATLAB can be used. MATLAB is the language of technical computing for a million of engineers and scientists in industry and academics. Here System configuration was intel I5 processor 2.2 GHZ, 4GB RAM.

Student's dataset can be categorized in different ways such as, it can be grouped for engineering, schools, non technical and also by country wise. Here for predicting performance of students based on attributes classification algorithm is applied on student performance dataset [17]. This is a multivariate dataset used for classification and regression tasks. In this dataset number of instances are 649 and number of attributes are 33.

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

After applying our proposed work as explained above to predict student performance the data analyzed based on commonly used measures such as accuracy, sensitivity and specificity. The performance of classification model is measured by evaluating the correctness of the classification decision of the classifier. The table shows these counts are commonly known as confusion matrix. The terms used in confusion matrix are:

(TP): Number of True Positives (Classifier correctly labeled record as positive).

(TN): Number of True Negatives (Classifier correctly labeled record as negative).

(FP): Number of False Positives (Classifier incorrectly labeled record as positive).

(FN): Number of False Negatives (Classifier incorrectly labeled record as negative).

To evaluate and compare classifier performance we used accuracy, sensitivity, and specificity. Using confusion matrix they are measured as:

Accuracy (proportion of total number of correct prediction): TP+TN/ TP+TN+FP+FN

Sensitivity (proportion of positives correctly predicted as positive): TP/TP+FN

Specificity (proportion of negatives correctly predicted as negative): TN/TN+FP

After computing these measures using 2 level classification techniques are compared with one of the classification algorithm J48.The comparison is shown in form of graphical representation based on values calculated as shown in tables given below.

Table 1: Accuracy Comparison

Method/Class	J.48	Proposed:2-level classification
Gh1	97.53	98.46
Gh2	87.03	92.76
Gh3	64.81	84.28
Gh4	79.93	90.29

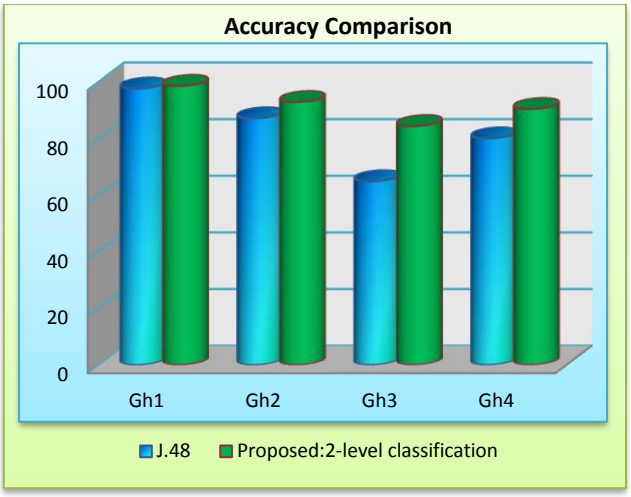

Figure 3: Accuracy Comparison

Table 2: Sensitivity Comparison

Method/Class	J.48	Proposed:2-level classification
Gh1	1	0.98
Gh2	1	0.95
Gh3	0.86	0.78
Gh4	1	0.94

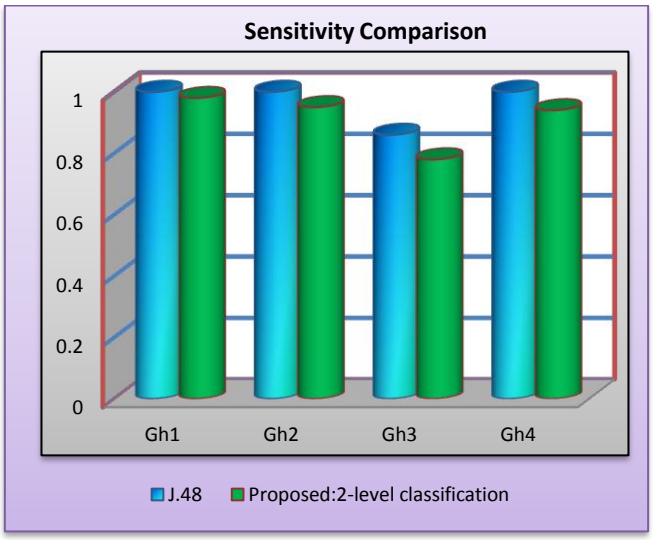

Figure 4: Sensitivity Comparison

Table 3: Specificity Comparison

Method/Class	J.48	Proposed:2-level classification
Gh1	0	80
Gh2	0	74.66
Gh3	100	87.44
Gh4	0	75

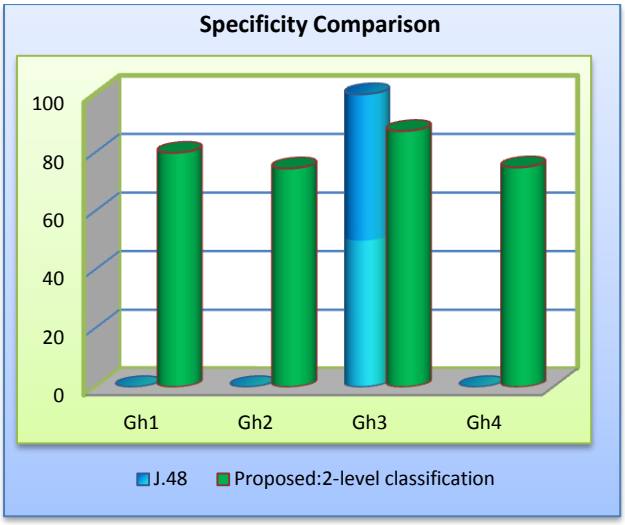

Figure 5: Specificity Comparison

Here table 1, 2 and 3 shows that the experimented result analysis of proposed and J48 method, and similarly figure 3, 4 and 5 shows that the comparisons result if the proposed and existing method, and where they found that proposed method has given more efficient result as compared to J48method, and produces up more than 98% accuracy rate, 98% sensitivity rate and 87% specificity rate. Here, the classes from G0 toG19 are divided into categories Gh1,Gh2,Gh3 and Gh4 such that from G0 to G5 comes in Gh1, G6 toG9 comes in Gh2, G10 to G14 comes in Gh3 and G15 to G19 comes in Gh4. From experimental result we analyze that using 2 level classifications in educational dataset the measures to predict student performance is to be evaluated for classes categorized as Gh1, Gh2, Gh3 and Gh4 with respect to the values in attributes of predicted classes. The accurate prediction of student academic performance is of importance for making decision as well as providing better educational services. The proposed classification model predicts with higher accuracy the Gh1, Gh2, Gh3 and Gh4 classes. So we analyzed that proposed work has given best result for the each predicted classes as compared to J48.

6. CONCLUSION AND FUTURE WORK

In this paper, 2 level classification methods is used in educational data mining area on student's dataset to predict the student's performance on the basis of student's database. We use some attribute were collected from the student's database to predict the performance. Predicting student's performance is mostly useful to help the educators and learners improving their learning and teaching process. In this the entropy of attributes is calculated to find highest information gain attribute based on feature selection. From above result it is clear that predicting student performance using 2 level classification algorithm gives more accurate result as compared to other previous classification algorithm J48 used in previous researches. For student's this study will be helpful to recover the student's performance, to find out those students which required special attention to reduce failing ration and taking suitable action at right time. To carry out further researches the analysis on predicting student's performance has provoked us. It will be beneficial for the educational system to monitor the student's performance in a systematic way. Classification Model in this paper can be further analyzed for predictions and for more accurate results by suspicious selection of values.

7. REFERENCES

[1] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby."Data Mining: A prediction for Student's Performance Using Classification Method'' World Journal of Computer Application and Technology, 2(2):pp. 43-47, 2014.

[2] Brijesh Kumar Baradwaj, Saurabh Pal." Mining Educational Data to Analyze Students" Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[3] Ahmed Mueen, Bassam Zafar, Umar Manzoor." Modeling and Predicting Student's Academic Performance Using Data Mining Techniques" I.J. Modern Education and Computer Science, 11, pp.36-42, November 2016.

[4] S. Indhu Priya, P. Devaki." Evaluating Students Performance in Placements Activity" International Journal of Innovations & Advancement in Computer Science (IJIACS) ISSN 2347 – 8616 Volume 6, Issue 1,January 2017.

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

[6] Shiwani Rana,Roopali Garg." Evaluation of Student's Performance of an Institute Using Clustering Algorithms" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 5,pp. 3605-3609,2016.

[7] Hashmia Hamsa, Simi Indiradevi, Jubilant 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 (RAEREST), Procedia Technology 25, pp.326 – 332,2016.

[8] C. Anuradha, T. Velmurugan." A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance" Indian Journal of Science and Technology, Vol 8(15), DOI: 10.17485/ijst/2015/v8i15/74555, July 2015.

[9] Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid." A Review on Predicting Student's Performance using Data Mining Techniques" The Third Information Systems International Conference,Procedia Computer Science 72, pp.414 – 422,2015.

[10] Xin chen, Mihaela Vorvoreanu, Krishna Madhavan."Mining Social Media Data for Understanding Students Learning Experiences" IEEE Transactions on Learning technologies,Vol.7,No.3,September 2014.

[12] Pratiyush Guleria, Niveditta Thakur, Manu Sood. "Predicting Student performance Using Decision Tree Classifiers and information Gain" International Conference on Parallel, Distributed and Grid Computing, DOI: 10.1109/PDGC.2014.7030728,February 2015.

[13] Kiran parmar, Dinesh kumar Vaghela, Priyanka Sharma. "Performance Prediction of Students Using Distributed Data Mining" IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems, DOI: 10.1109/ICIIECS.2015.7192860 ,August 2015.

[14] Alana M.de Morais, Joseana M.F.R. Araujo, Evandro B.Costa. "Monitoring Students Performance Using Data Clustering and Predictive Modelling" Frontiers in Education Conference, DOI: 10.1109/FIE.2014.7044401 , February 2015.

[15] Ruchi Jain." Application of KNN-Genetic Algorithm for Analysing Student Learning in Educational Data Mining paradigm" International Journal of Innovation Research in Computer and Communication Engineering,Vol.4,Issue 6,pp.10319-10323,June 2016.

[16] Susan Bergin, Aidan Mooney, John Ghent, Keith Quille." Using Machine Learning Techniques to Predict Introductory Programming Performance" International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 12, pp.323-328,December 2015.

[17] Student Performance Data Set https://archive.ics.uci.edu/ml/datasets/Student+Performance

[18] Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal." Data Mining Applications: A comparative Study for Predicting Student's performance**"** International Journal of Innovative Technology & Creative Engineering (ISSN:2045-711) Vol.1 No.12 December.

[19] Guleria Pratiyush, Sood Manu." Classifying Educational Data Using Support Vector Machines:A Supervised Data Mining Technique" Indian Journal of Science and Technology, Vol 9(34), DOI: 10.17485/ijst/2016/v9i34/100206, September 2016.