STUDENTS' PERFORMANCE EVALUATION USING MACHINE LEARNING ALGORITHMS

[1]P. Krishna Reddy*, [2]S. Bavankumar, [3]Y. Peer Mohaideen, [4]G. Kishore
[1,2,3,4]Assistant Professor, Department of Computer Science and Engineering,
[1,2,3,4]St. Martin's Engineering College, Secunderabad, Telangana, India.
*Corresponding Author E-mail: pkrishnareddycse@smec.ac.in

Abstract

Student's performance is a major problem for the society. Rapid growth of technologies and the application of different machine learning methods in present years, the development of good models increase the progress of student's performance progress have become more and more accurate. Therefore, development of machine learning techniques, which can effectively predict student's performance, is of vast importance. In this research paper, we apply five different data mining techniques Passive Aggressive Classifier (PAC), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Radius Neighbour Classifier (RNC) and Extra Tree (ET) and then compare the results of five machine learning algorithms to choose the best performing algorithm. We use educational data to analysis different machine learning techniques to evaluate the performance of student.

The results obtained by different machine learning algorithms are discussed in this paper and we get the highest accuracy in the case of Support Vector Machine (SVM).Various metrics are also evaluated to verify the results of accuracy like sensitivity, specificity and precision. These results can be applied on the new coming students to check whether they perform well or not and by knowing the non-performing students, higher educational institutions can pay attaint ion for improving student's performance.

Keywords: Educational Data Mining; Support Vector Machines, Radius Neighbor Classifier, Linear Discriminant Analysis, Passive Aggressive Classifier.

1. Introduction

The quality of an academic institution is depend on the performance of student and dropout rate between the enrolled students in a course and finally completed the course. The dropout rate is high because students do not know whether the course in which they are going to take admission is suitable for their study or not. In India parents forced the student to take admission in Engineering or professional courses without knowing their interest and this is the main reason of the dropout and low performance.

Educational Data Mining (EDM) is an area focusing to use technologies and data mining techniques in the teaching environment. EDM relates to the machine learning for identifying hidden patterns within huge academic data, to develop data mining and statistical methods, research and implementation, which would provide fruitful results

for the students. Data mining is a complicated procedure and needs multi-step for identifying hidden patterns. Data mining has a cross-cutting regulation which requires for merging knowledge from all walks of life [1]. Higher Education uses this information to help to focus on those poor students, who have a higher risk of failure by providing classification features [2-6]. Data searched by mining techniques, knowledge, institutions of higher learning will not be limited to make better decisions in a variety of ways, Students make more advanced plan to the instructions, will be able to predict individual behavior with high accuracy and organization allocated more effectively the resources and staff. . This results in improved effectiveness and efficiency of the processes [3]. Data is a form of classification data mining Analysis that could be critical data used to describe the classes or remove models to predict the set of data for the future. Classification process has two steps, the first step learning process; Training data will be followed by the classification algorithms. Learned models or classification rules will be represented as. Next, the second stage classification process where the classification model used test data to estimate the accuracy of the classifier.

The keypurpose of this research paper is to develop an efficient predictive modelwith the help of Passive Aggressive Classifier (PAC), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Radius Neighbour Classifier (RNC) and Extra Tree (ET) to predict student's performance into performer or not-performer students.The performer and non-performer students are predicted by student's dataset.

2. Related Work:

The literature related to the use of machine learning technology in field of higher education mainly focuses on the application of technologies such as clustering, association rules, classification, regression and statistics to predict, the performance ofthe developed model. Educational data mining (EDM) researchers provide other aspects related to academic activities, including identifying factors related to student success, failure, and intention to drop out [7-9], institutional planning and strategies [10-11], and understanding Teacher support and administrative decision-making.

The applications of machine learning techniques in the field of higher education is still in primary stage and it's need more consideration. Educational Data Mining inthe field of education mainly to improvestudents'performance with the help of learning process by identifying, extracting and evaluating attributes related to the students characteristics[1]. With the help of educational data mining we can improve the decision making and implement policies for student that helps institutions of higher education today [12-13].

Ashwin Satyanarayana, and Mariusz Nuckowski conducted a survey over 788 school student dataset which includes 37 questions each. In this research multiple classifiers were used like Decision Trees, Naïve Bayes and Random Forest to get accuracy over student's data by eliminating noisy instances. From this study association rules were identified that affect student's results using a set of rule based techniques like Apriori, Filtered Associator and Tertius. In this regard the result was found that prior work there was no filtering on student data has been performed and focused on using single classifiers. So in this study comparison of single filters and ensemble filters was done and it is concluded that ensemble filters works better for identifying and removing noisy instances [14].

Another comparative study was done by Bhrigu kapur, Nakin Ahluwalia and Sathyaraj (2017). They compared six algorithms like J48, Random Forest, Naïve Bayes, Naïve Bayes Multinomial, K-Star and IBK. They used 480 entry of data set and implemented through Weka tool. The Survey conducted based on seven attributes and found Random Forest algorithm provides more accuracy compared to otheralgorithms [15].

Various previous works has been done by Pal et. al [16-19] to improve the performance of the prediction using different data mining techniques and they provide a better results which are also applicable at various institution to find the weakstudents.

K. Prasada Rao et. al [20] conducted a survey over 200 college students. In this research classification techniques were used on student database to predict the learning behavior of student's. From this research, the researcher identified the slow learners, and effectively the action taken to rectify the failures and take appropriate action to qualify the weaker students in perfect manner. In this study the performance of J48, Naïve Bayes and Random forest algorithms were compared. Finally theresearcher got accuracy using Random forest algorithm when the data set is in massive size.

A research carried out by the team [21] (2016), and the performance of the student's were predicted. Classification techniques were used to create prediction module of the system to predict the future values. Various parameters like previous academic performance were considered to predict student's academic results and placement.The dashboard is the module which describes the whole overview of the institution in a graphical representation of data. Decision tree algorithms ID3 and C4.5 were implemented to generate reports based on structured database. From this research ID3 algorithm provided the best accuracy of 95.33%.

3. Methods

Machine Learning is the technique for developingnew algorithms, which provides computer the capabilityto learn from previously stored information's. In this research paper different machine learning classifiers are used. (i) PAC
(ii) SVM, (iii) LDA (iv) RNCand (v)ET. A brief description of the classifiers used are described below.

- **Passive Aggressive Classifier (PAC):**

  PAC algorithm is a set of algorithms which are used for comprehensive learning. Passive-aggressive algorithms are very similar to Multilayer Perceptron except learning rate is not required. But, converse to Perceptron, Passive-aggressive algorithmscomprise a regularization variable C.

- **Support Vector Machine (SVM):**

  Support Vector machine is discriminative classifier used in supervised learning problems i.e. given labeled training data, and finds out the line (or hyperplane) in a multidimensional space, which separate outs classes

- **LinearDiscriminantAnalysis (LDA):**

  This algorithm is also known as attribute reduction method. LDA is supervised machine learning technique. This method reduced the attributes as less as possible in a dataset without affecting the results of the classification. Linear Discriminant Analysis, or LDA, uses the information from all reduced features to create a new axis and projects the data on to the new axis in such a way as to minimizes the variance and maximizes the distance between the means of the two classes.

- **RadiusNeighborsClassifier (RNC):**
  RNC is a type of KNeighborsClassifier (Radius based learning algorithms). RNC returns the indices and distances of each data points from the dataset lying in a ball with size radius around the points of the query array. Points lying on the boundary are included in the results.
- **Extra Tree (ET):**
  This method is an ensemble method which stands for Extremely Randomized Trees. The main objective of this algorithm is to further randomizing tree building in the context of numeric input features, where the choice of the optimal cut-point is responsible for a large proportion of the variance of the

induced tree. It often leads to increased accuracy when compared to theordinary random forest.

Fig. 1 shows the structure of methodology used in this research paper.

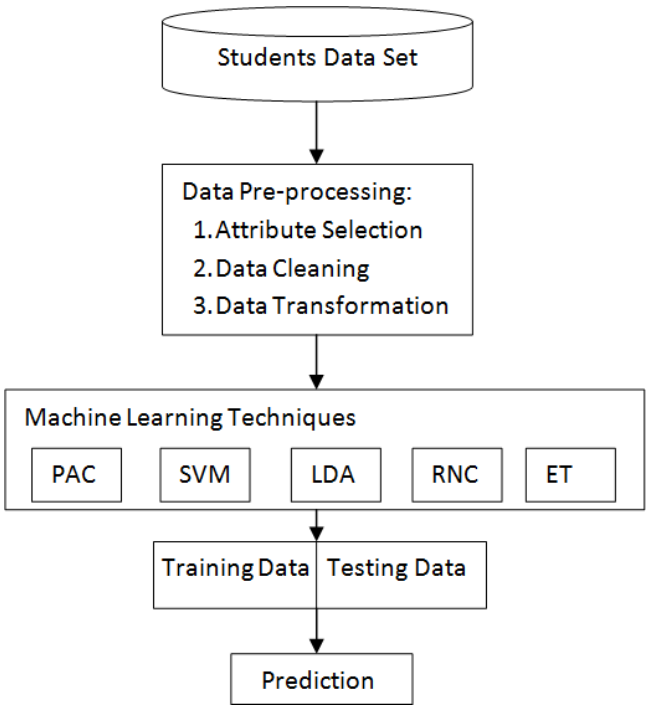

Figure. 1. Methodological approach for Performance Prediction

3.1 Dataset Analysis:

The data used in this study is of Bachelor of Computer Applications programme, which has been collected from United Institute of Management, Prayagraj. The BCA course is divided in 3 years which consist of two semester per year, therefore total six semester examination completes the whole BCA course. In this research paper we have taken count of only final semester results. The data is collected with the permission of examination and admission departments from the year 2014 to 2019 and total number of students passed from the institution is 1000, therefore total 1000 instances are available with 22 attributes; these attributes are collected from theregistration as well as examination form. The target and other variables discussed in this study are listed in table 1.

Table 1: Student Dataset

Feature	Attribute	Domain

S1	Sex of Students	1= Female, 2=Male
S2	Students category	1= General, 2=OBC, 3=SC, 4=ST, 5=Minority
S3	Discussion at home	1=Always, 2=Almost Always, 3=Sometimes, 4= Never
S4	Own Computer /Laptop	1=Yes, 2=No
S5	Laptop shared with family	1=Yes, 2=No
S6	Study desk at home	1=Yes, 2=No
S7	Own mobile phone	1=Yes, 2=No
S8	Own Gaming system	1=Yes, 2=No
S9	Heating/Cooling systems at	1=Yes, 2=No
S10	Absent from school	1=Once a week or more, 2=Onceevery two weeks, 3=Once a month, 4=Never or almost never
S11	How often use computer/Laptop at home	1=Every day or almost every day, 2=Once or twice a week, 3=Once or twice a month, 4=Never or almost never
S12	How often use computerat School	1=Every day or almost every day, 2=Once or twice a week, 3=Once or twice a fifteen days, 4= Once or twice in a month, 5=Never or almost never
S13	Access textbooks	1=Yes, 2=No
S14	Completed assignments	1=Yes, 2=No
S15	Collaborate with classmates	1=Yes, 2=No
S16	Communicate with	1=Yes, 2=No

	teacher	
S17	Students grade in Senior Secondary education	1 =90% -100%, 2= 80% - 89%, 3= 70% - 79%, 4= 60% - 69%, 5= 50% - 59%, 6= 40% - 49%, 7= < 40%
S18	Fathers qualification	1=elementary, 2=secondary, 3=graduate/ost-graduate, 4=doctorate
S19	Mother's Qualification	1=elementary, 2=secondary, 3=graduate/ost-graduate, 4=doctorate
S20	Father's Occupation	1=Service, 2=business, 3=not-applicable
S21	Mother's Occupation	1=House-wife , 2=Service, 3=business, 4=not-applicable
S22	Grade obtained in B.C.A	1= > 60%, 2= >45 &<60%, 3= >36 &<45%, 4= < 36%}

3.2 DataPreprocessing:

The methodology proposed in this research paper starts with data preprocessing. Data preprocessing step includes (i) a data driven method to select students' records and selecting important variables for analysis and (ii) The collected data from students records are not clean and may include noise, incorrect, missing values, or inconsistent data. So we have to apply different method of data cleaningto clean such anomalies. The Experimental investigation is made using the students data set collected from the United Institute of Management, PrayagRaj. The dataset had descriptions, 1028 instances, in 22 dimensions. The noise and missing values present in the dataset may impact the predictive ability of the machine learning model. Hence students dataset is extensively pre-processed using a normal scalar using the equation

$$x_N = \frac{(x - x_{mean})}{SD}$$

Where

x_N= Normalized value of x,
x=Original value of x,

x_{mean} =Mean value of x and

SD=Standard deviation of the given population.

4. Results and Discussion

Before conducting the experiment, we first visualize the values of attributes shown in Fig 2. Figure shows the histogram of all attributes related to student dataset which consists 1000 instances and 22 attributes. Each attribute represent the bar of frequency of different values excluding the feature S22 (which is target attribute Result).

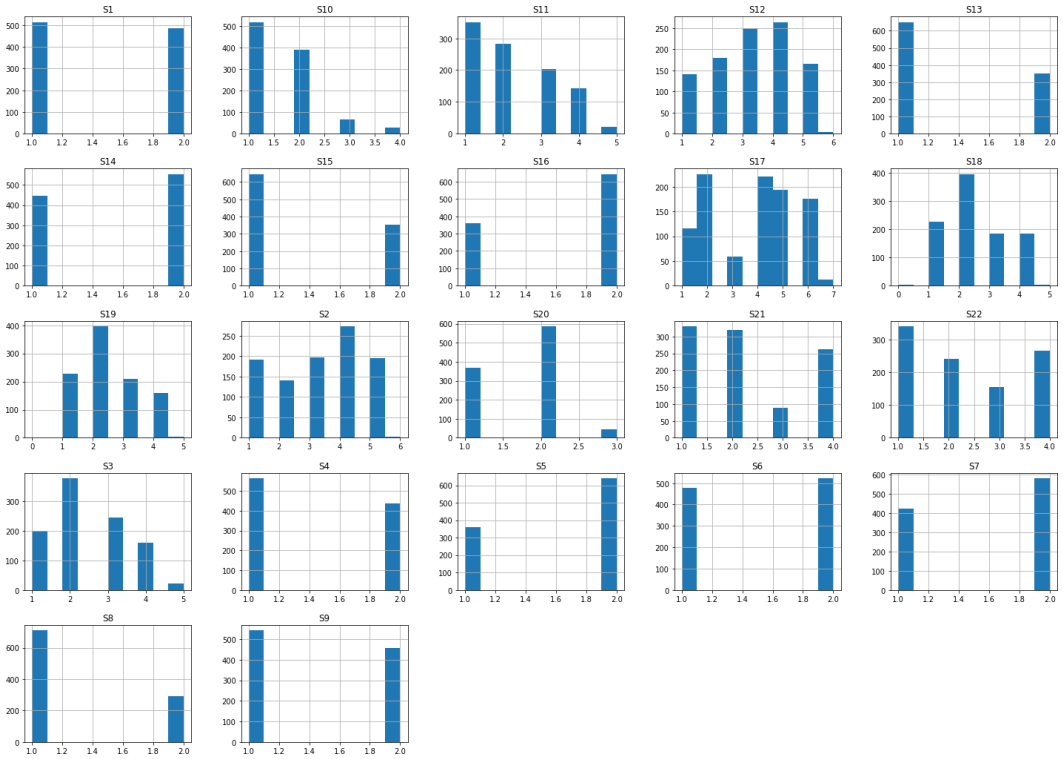

Figure2. Dataset visualization using histogram

The experiment is performed on student's dataset using Python code along with the supporting packages such as Scikit-learn, Pandas and Numpy etc. The student dataset is divided into 80% training set and 20% test dataset.

Another diagram that helps summarize the observed distribution is the box and the whisker. The plot draws a 25th and 75th percentile around the data that captures the middle 50% of the observations. Draw a line at the 50th percentile (median) and drawwhiskers above and below the box to summarize the general range of observations. Draw points for outliers outside the data or for outliers outside the range. The box andwhisker plot of data set is shown in Fig.3.

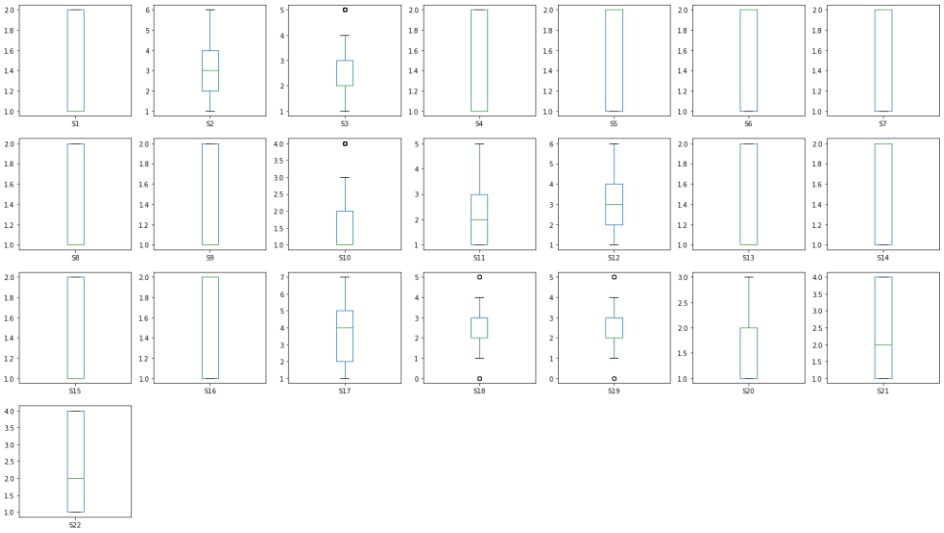

Figure3. Dataset visualization using Box and Whisker Plot

In this paper Python code is used to represent the different graphs and to evaluate accuracy, precision, recall and sensitivity of the different machine learning techniques initially.Python programming is chosen because the codes for different classifiers have been defined in the form of predefined modules.

The Performance of classifiers is the most important metrics for any predictive model especially when the model is built for the performance prediction. A wrong predictionmay have to pay a heavy cost of student. Hence, the selection of a performance metrics plays a very crucial role in performance prediction. In data analysis system, there are a number of performance metrics such as accuracy, sensitivity, precision andspecificity which are shown in table 2.

Table 2: Formulas

Sr. No.	Performance Metrics	Formula
1.	Accuracy	$\frac{(TP+TN)}{(TP+FP+TN+FN)}$
2.	Sensitivity	$\frac{TP}{(TP+FN)}$
3.	Specificity	$\frac{FP}{(FP+TN)}$
4.	Precision	$\frac{TP}{(TP+FP)}$

The value calculated by five classifiers is shown in Table 3.

Table 3:Output of Evaluating Algorithms

Classifier	Accuracy	Precision	Sensitivity	Specificity
PCA	89.51	84.95	62.38	97.25
SVM	94.86	89.17	63.1	98.65
LDA	93.21	88.26	59.23	98.25
RNC	87.23	91.23	66.27	92.89
ET	91.27	90.04	70.23	93.44

A high accuracy score of a classifier does not ensure that the classifier correctly predicts the desired results. A high accuracy may be the result of more number of correct predictions of true negative cases. Therefore, only achieving high accuracy of a classifier cannot be considered as a good measure for a classification algorithm.

In a predictive analysis system a wrong prediction may be a false-negatives or false- positives. The cost of these two wrong predictions may vary from one system to othersystem. In one system, a false-negative result may incur more cost from a false- positive result. For example in a performance prediction system such as good performance prediction, classifier cannot afford a wrong prediction about a student which is actually bad performer (TRUE- POSITIVE) and is predicted as non-performer (FALSE-NEGATIVE). So we need a model in which the chance of false- positives and false-negatives is less. In other words, its precision should be high since number of false-positives is less, similarly recall should be also high because it shows lower number of false-negatives. A high precision and high recall of classifier ensuresthat it predicts less number of False-Positive and False-Negative results. In case of high false-positives cost, precision will be good measure and recall is for high false- negative cost. Accuracy will be a good measure if cost of false positives and false negatives are nearly same but when it is different precision and sensitivity both are considered.

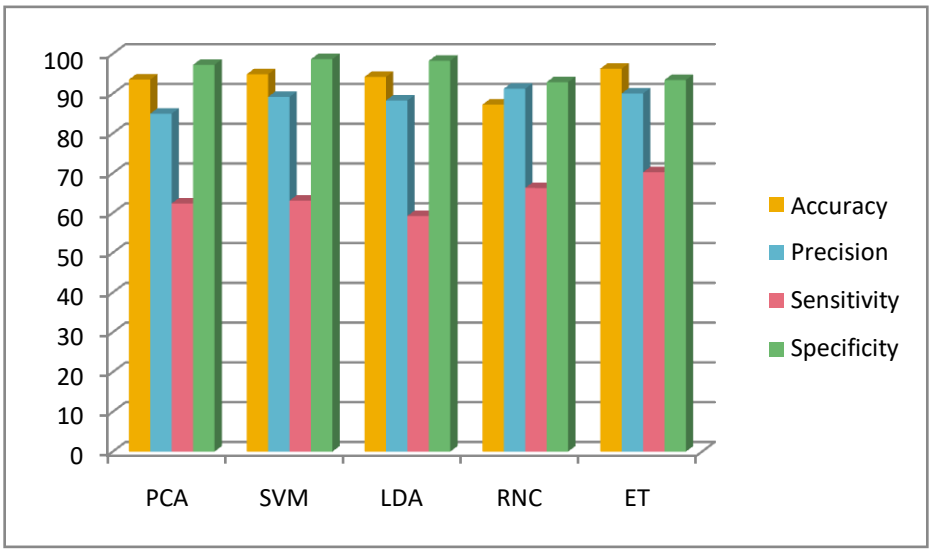

Figure4. Accuracy of different algorithms

5. Conclusion

The aim of the proposed work is to build efficient framework which can extensively improve Performance accuracy of students. Machine learning techniques are widely used in student's performance prediction. Knowledge gained with the help of machinelearning techniques can be used to make successful and effective decisions that improve and develop student's performance.This paper describes different machine learning techniques for evaluating the performance of students.Five machine learning techniques PCA, SVM, LDA,RNC and ET are used to classify the prediction of students. The best accuracy find among these different techniques is 94.86% from SVM. The second highest accuracy obtained is 93.21% in the case of LDA.

We get the highest accuracy in the literature available on student's performance prediction. The machine learning-based method reduces generation errors and obtains more information by using the first-stage prediction as a feature rather than a separate training. In addition, by using machine learning, the complex relationships between classifiers are automatically learned, enabling the collection method for better predictions.

These results can be used to pay a more attention on the non-performing students to improve their performance and the quality of higher educational institute.

References

[1]. B.Baradwaj, S.pal, "Mining Educational Data to Analyze Students' Performance" (IJACSA) International Journal Of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[2]. Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. In: Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12.

[3]. D. Michie, D.J. Spiegelhalter, and C.C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood Series in Artificial Intelligence, 1994.

[4]. H. E. Erdem, "A cross-sectional survey in progress on factors affecting students' academic performance at a Turkish university," Procedia-Social and Behavioral Sciences, vol. 70, pp. 691-695, 2013.

[5]. G. Elakia and N. J. Aarthi, "Application of data mining in educational database for predicting behavioural patterns of the students," International Journal of Computer Science and Information Technologies, pp. 4649-4652, 2014.

[6]. S. Parack and F. Z. Zahid, "Application of data mining in educational databases for predicting academic trends and patterns, in: Technology Enhanced Education (ICTEE)," IEEE International Conference on, IEEE, pp. 1-4, 2012.

[7]. Cambruzzi, W.L., Rigo, S.J., Barbosa, J.L., 2015. Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach. J. UCS 21 (1), 23–47.

[8]. Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., Loumos, V., 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput. Educ. 53 (3), 950–965.

[9]. Marquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H., Ventura, S., 2016. Early dropout prediction using data mining: a case study with high school students. Exp. Syst. 33 (1), 107–124.

[10]. Caputi, V., Garrido, A., 2015. Student-oriented planning of e-learning contents for Moodle. J. Network Comput. Appl. 53, 115–127.

[11]. Mankad, S.H., 2016. Predicting learning behaviour of students: Strategies for making the course journey interesting. In: Paper presented at the Intelligent Systems and Control (ISCO), 2016 10th International Conference on.

[12]. Sacin, C.V., Agapito, J.B., Shafti, L., Ortigosa, A., 2009. Recommendation in higher education using data mining techniques. In: Paper presented at the Educational Data Mining 2009.

[13]. Ji, H., Park, K., Jo, J., Lim, H., 2016. Mining students activities from a computer supported collaborative learning system based on peer to peer network. Peer-to-Peer Netw. Appl. 9 (3), 465–476.

[14]. Ashwin Satyanarayana, Mariusz Nuckowski, "Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance" Spring '2016' Mid. Atlantic 'ASEE' Conference, 'April' 8.9,'2016' GWU.

[15]. Bhrigu Kapur, Nakin Ahluwalia and Sathyaraj R, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms", International Journal of Advanced Research in Computer Science, 8 (3), March-April 2017,632-636.

[16]. Pandey, U.K. and Pal, S., 2011. Data Mining: A prediction of performer or underperformer using classification. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (2), 2011, 686-690.

[17]. Bhardwaj, B.K. and Pal, S., 2012. Data Mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, 2011.

[18]. Yadav, S.K., Bharadwaj, B. and Pal, S., 2012. 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.

[19]. Yadav, S.K. and Pal, S., 2012. Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal (WCSIT). (ISSN: 2221-0741), Vol. 2, No. 2, 51-56, 2012.

[20]. Prasada Rao, K. , M. V.P. Chandra Sekhara, and B. Ramesh. "Predicting Learning Behavior of Students using Classification Techniques." International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, April 2016.

[21]. Siddhi Parekh, Ameya Nadkarni, and Riya Mehta (2016) "Results and Placement Analysis and Prediction using Data Mining and Dashboard." International Journal of Computer Applications (0975 – 8887) Volume 137 – No.13, March 2016 In Proceedings of the 22nd International Conference on World Wide Web, pp. 413-418. ACM.