Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques

Deepti Aggarwal, Sonu Mittal, Vikram Bali

Abstract - In today's competitive world of educational organizations, the universities and colleges are using various data mining tools and techniques to improve the students' performance. Now a days, when the number of drop out students is increasing every year, if we get to know the probability of a student whether he/she will be able to cope up easily with the course, it is possible to take some preventive actions beforehand. In other words, if we get to know that a student will clear his papers in the course or he will have reappear in papers, a teacher/parent can focus more on such students. The data set of students has been taken from the UCI Machine Learning repository where a sample of 131 students have been provided with twenty-two attributes. The results of six classification algorithms have been compared in order to predict the most appropriate model for classifying whether a student will have a reappear in a course or not.

Index Terms: Classification, Multi-Layer Perceptron, Prediction, Random Forest

I. INTRODUCTION

EDM (Educational Data Mining) is an evolving area of research that deals with studying and developing of various methods to analyze data that originates in the field of education. EDM uses various computational methods to analyze academic data to solve educational queries. Data mining, also called as knowledge discovery, is extracting the interesting patterns from large data sets automatically. It can be used to train the learning models and to predict and evaluate by discovering useful learning information from the historical data.

Classification is an important task in data mining where student's performance can be predicted using a particular algorithm. Various classification algorithms can be used for the prediction like Naïve Bayes, Logistic Regression, Random Forest, J48, CART, Multi-Layer Perceptron etc. In this paper, six algorithms, "Naive Bayes", "Logistic Regression", "Support Vector Machine", "Multi-Layer Perceptron", "J48" and "Random Forest" have been compared based on the classification accuracy and other comparison metrics.

Revised Manuscript Received on July 5, 2019

Deepti Aggarwal, Research Scholar, School of Computer and System Sciences, Jaipur National University, Jaipur, India,

Sonu Mittal, Associate Professor, School of Computer and System Sciences, Jaipur National University, Jaipur, India,

Vikram Bali, Professor & Head, Department of Computer Science and Engineering, JSS Academy of Technical Education, Noida, India,

The data set has been taken from the UCI machine Learning repository which includes the data of three different colleges of Assam, India. The data contains not only the demographic details, but also the academic details of a student. The parameters that are available at the time of admission are considered and the post admission parameters like internal assessment percentage, end semester percentage were removed to develop the model. The data set consists of 131 instances with 19 features. The highly influential features are extracted using Naïve Bayes classifier. The class variable is taken to be 'arr' that tells whether a student has reappear/s in the course or not.

II. LITERATURE REVIEW

Hussain et al. (2018) used J48, PART, Random Forest and Bayes classifiers to predict students' end semester grades on a data set of 300 students from different colleges and found that Random Forest classification algorithm gives the best results based on accuracy and classifier errors [1].

Bali et al. (2018) proposed the use of optimization technique for Optimal Component Selection and 'Rock Predication by using Artificial Neural Networks' [2][9][11].

Mittal et al. (2018) designed a model for predicting diabetes using Naïve Bayes, K-NN and Support Vector Machine (SVM) and concluded that the SVM classifier outperforms the rest of the two classifiers [3].

Mittal et al. (2018) merged the SMOTE and decision tree classifiers and got a very high accuracy on the resultant model for predicting diabetes prognosis [4].

Evandro B.Costa et al. (2017) compared the effectiveness of educational data mining techniques on the data with pre-processing and without pre-processing. The authors also applied fine tuning on algorithms to check if it increases the effectiveness of EDM techniques [5].

Almarabeh, H.(2017) used Bayesian Network, J48, Naive Bayes, ID3, J48 and Neural Network classifiers to analyse and evaluate students' performance grades and found that Bayesian Network classifier has the highest accuracy among all classifiers . They concluded that the performance of the students of a university can be best classified using Bayesian Network classification methods. The dependency among random variables is depicted by using directed acyclic graphs, where the nodes in the graph represent the random variables. The dependency of random variables is depicted when a connection exists between a node and an arc [6].

Govindaswamy and Velmurugan (2017) used classification and clustering algorithms such as C4.5, Expectation Maximization, k-nearest neighbour, k-means and Naïve Bayes for predicting the performance of students. Their results show that C4.5 classifier gives the best results compared to other classification algorithms. They also found that clustering algorithms give better performance as compared to classification algorithms [7].

Anuradha and Velmurugan (2015) evaluated various classifiers for predicting the performance of students in their research. They used the classification algorithms such as J48, OneRip, JRip, Naïve Bayes classifiers and Bayesian Net classification algorithms on the data set of students from three private colleges in Tamil Nadu, India. Their results illustrate that the prediction rates of these algorithms vary from 61-75% [8].

Patil et al. (2013) compared the results of decision tree classifier and Naïve Bayes algorithm and found that 'decision tree' gives better results than 'Naïve Bayes' algorithm. "The benefit of 'decision tree' is that it is easy to understand and interpret. The decision tree gives good performance with bot numerical and categorical variables. Thus, it is one of the very powerful and widely used classifiers. WEKA uses J48 classifier to implement the C4.5 decision tree [10].

Dekker et al. (2009) worked on Random Forest classifier and found that Random forest classifier reduce bias, variance and overfitting. Hence, it is very accurate as well as robust. It merges various decision trees together to give a better prediction model [12].

Romero and Ventura (2007) worked on the recommendation agents that watches the student activities and suggests some actions that will be beneficial for the students [13].

After studying the various researches, that have been done for doing predictive analysis through different educational data mining techniques, the authors found the following six classification algorithms: 'Naïve Bayes', 'Logistic regression', 'Multilayer Perceptron', 'Support Vector Machine', 'J48' and 'Random Forest' to be the most promising classifiers to build a model for predicting whether a students will have a reappear in a course or not.

A. Classifier Evaluation and Comparison Metrics

The classifiers are evaluated by a confusion matrix which is a combination of four outcomes. In binary classification, the output is either positive or negative. The four different classifications are:

- True Positives (TP)- Accurate positive prediction
- False Positives (FP)- Wrong positive prediction
- True Negatives (TN)- Accurate negative prediction
- False Negatives (FN)- Wrong negative prediction

The effectiveness metrics for classifiers used in the research are:

➢ Precision (P)

Precision = $\frac{TP}{TP+FP}$, number of true positives classifications divided by the sum of true positives and false positive classifications

➢ Recall (R)

Recall = $\frac{TP}{TP+FN}$ i.e number of true positives classifications divided by the sum of true positive and false negative classifications

➢ F1-Score

F1-Score is the harmonic mean of precision and recall.

$$\text{F1-Score} = \frac{2*P*R}{(P+R)}$$

➢ Accuracy

Accuracy is measured by dividing the number of correctly classified instances by the total number of instances.

➢Mean Absolute Error (MAE)

MAE measures the average magnitude of errors in a set of predictions. It is the summation of the differences between predicted and actual observation divided by the total number of test samples.

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

➢ Root Mean Square Error (RMSE)

It is the square root of the summation of the squared differences between predicted and actual observations, divided by the number of total test samples.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}$$

III. DATA PRE-PROCESSING

The dataset is a collection of students' details collected from three colleges of Assam sate of India from UCI Machine Learning Repository (Table 1). The student performance data set consisted of 22 attributes, from which three attributes(internal assessment percentage, end semester percentage and attendance) were removed beforehand because the data required for the prediction is the data available at the time of admission of a student The data set considered in this paper consists of 131 attributes with 18 attributes and one class variable.

The various attributes contained in the data set are:

1. Gender (ge)
2. Caste (cst)
3. Matric Percentage (tnp)
4. XII Percentage (twp)
5. Reappear/back paper (arr)
6. Marital status (ms)
7. Living status (ls)
8. Admission category (as)
9. Family income (fmi)
10. Family size (fs)
11. Father's qualification (fq)

12. Mother's qualification (mq)
13. Father's occupation (fo)
14. Mother's occupation (mo)
15. Number of friends (nf)
16. Study hours (sh)
17. Type of school attended (ss)
18. Medium (me)
19. Travel time between college and home (tt)

The description of the data set used is shown in Table I along with the possible attribute labels.

Table I: Description of data set

Attribute	Labels	Number of Instances Under Each Label
ge	(M, F)	72,59
cst	(Gen,SC,ST,OBC,MOBC) Gen : General SC : Schedule Caste ST : Schedule Tribe OBC : Other Backward Class MOBC : Minorities and other backward class	44,20,4,57,6
tnp	(Best, Very Good, Good, Pass, Fail)	9,38,59,25,0
twp	(Best, Very Good, Good, Pass, Fail)	5,44,65,17,0
arr	(Yes, No)	53,78
ms	(Married, Unmarried)	0,131
ls	(Town, Village)	39,92
as	(Free, Paid)	55,76
fmi	(Very-High, High, AM, Medium, Low) Very-High: fmi >= 30000 High: 20000 <= fmi < 30000 Above Medium : 10000 <= fmi < 20000 Medium : 5000 <= fmi < 10000 Low : fmi < 5000	6,15,27,63,20
fs	(Large, Average, Small) Large : fs > 12 Average : 6 <= fs < 12 Small : fs < 6	2,40,89
fq	(IL, UM, 10, 12 , Graduate, Post Graduate) IL : Illiterate UM : Under Matric	20,40,23,22,20, 6
mq	(IL, UM, 10, 12 , Graduate, Post Graduate) IL : Illiterate UM : Under Matric	27,52,25,17,7,3
fo	(Service, Business, Retired, Farmer, Others)	38,34,3,27,29
mo	(Service, Business, Retired, Farmer, Others)	12,1,1,115,2
nf	(Large, Average, Small) Large : nf > 12 Average : 6 <= nf < 12 Small : nf < 6	58,43,30
sh	(Good, Average, Poor) Good : sh >= 6 hours Average : sh >= 4 hours Poor : sh < 2 hours	27,59,45
ss	(Goverment, Private)	91,40
me	(English, Assamese, Hindi, Bengali)	62,60,7,2
tt	(Large, Average, Small) Large : tt >= 2 hours Average : tt >=1hour Small : tt < 1 hour	10,43,78

In the data set, out of 131 instances, 53 instances are positive (marked as blue) and 78 instances are negative (marked as red) for the class variable 'arr'. In other words, out of 131 students, 53 students have reappear and 78 students don't have reappear in their course. The nineteen features listed above can be visualized with respect to class feature(arr) as shown in figure 1, e.g. the first graph (topmost left corner) shows the graph of attribute 'gender' w.r.t. the class variable 'arr'. The first bar in the graph tells that there are 72 male students (Fig. 1) and the number of male students who have reappear (marked as blue) is less than the number of male students who don't have reappear (marked as red). The second bar in the same graph tells that there are 59 female students (Fig. 1) and the number of female students who have reappear (marked as blue) is less than the number of female students who don't have reappear (marked as red). In the same manner, the remaining graphs are a visualization of all the remaining 18 features w.r.t. the class variable 'arr'.

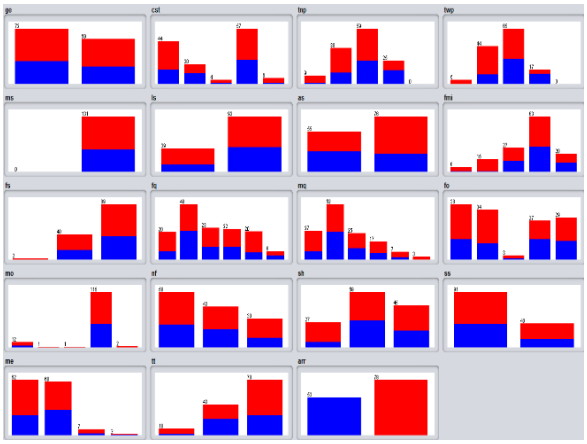

Fig. 1: Visualization of all features w.r.t. class variable (arr)

IV.FEATURE EXTRACTION

Feature extraction is done using several methods like wrapper method, info gain method etc., which creates a subset of features/attributes from the overall set of features/attributes. The features are selected using a particular classification algorithm like Naïve Bayes or Random forest etc. The selected subset of features are kept along with the class variable and the remaining features are removed from the data set. The model has been trained using this subset of features by using an appropriate search method like depth first search or breadth first search or any other.

In this paper, the selection of features has been done using Naïve Bayes classification algorithm, where the best first search has been used to identify the subset of features. The Naïve Bayes algorithm tells that six features are of more importance as compared to the remaining 15 features. The Naïve Bayes classification tells that the features that gives the best results in predicting the class variable(arr) are:

1. Gender (ge)
2. Caste (cst)
3. Percentage in XII (twp)
4. Family income (fmi)
5. Mother's occupation (mo)
6. Number of friends (nf)

The six features listed above can be visualized with respect to class feature(arr) as shown in fig. 2 to fig. 7, where positive instances are marked blue and negative instances are marked red. Fig. 2 shows that there are 72 male students and the number of male students who have reappear(marked as blue) is less than the number of male students who don't have reappear(marked as red). The second bar in the same graph tells that there are 59 female students and the number of female students who have reappear (marked as blue) is less than the number of female students who don't have reappear (marked as red).

Name: ge
Missing: 0 (0%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	M	72	72.0
2	F	59	59.0

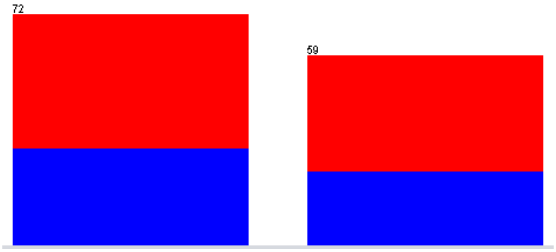

Fig. 2: Visualization of attribute 'ge' w.r.t. class variable(arr)

Fig. 3 shows the number of students in different caste categories e.g. number of students in general category are 44. The first bar in the graph shows that the number of students with reappear in general category is less than the number of students who don't have reappear in course. The second bar shows that out of 20 students under ST category, students with reappear are more as compared to students without reappear. Similarly, the remaining bars visualize the rest of the three labels (SC, OBC and MOBC) w.r.t. the class variable 'arr'.

Name: cst
Missing: 0 (0%) Distinct: 5 Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	G	44	44.0
2	ST	20	20.0
3	SC	4	4.0
4	OBC	57	57.0
5	MOBC	6	6.0

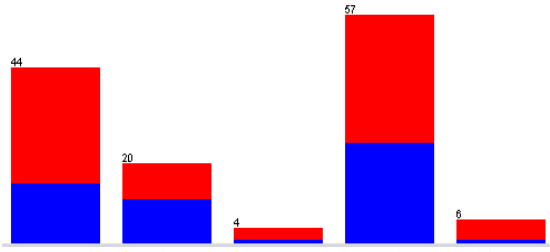

Fig. 3: Visualization of attribute 'cst' w.r.t. class variable(arr)

Fig. 4 shows the number of students with class XII %age under five different labels (Best, Very good, Good, Pass, Fail). The first bar in the graph shows that the students with label 'best' do not have any reappear in course. The second bar shows that out of 44 students under Very Good (VG) category, students with reappear are less as compared to students without reappear. Similarly, the remaining bars visualize the rest of the three labels (Good, Pass and Fail)

to class feature(arr) as shown in fig. 2 to fig. 7, where positive instances are marked blue and negative instances are marked red. Fig. 2 shows that there are 72 male students and the number of male students who have reappear(marked as blue) is less than the number of male students who don't have reappear(marked as red). The second bar in the same graph tells that there are 59 female students and the number of female students who have reappear (marked as blue) is less than the number of female students who don't have reappear (marked as red).

Name: ge
Missing: 0 (0%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	M	72	72.0
2	F	59	59.0

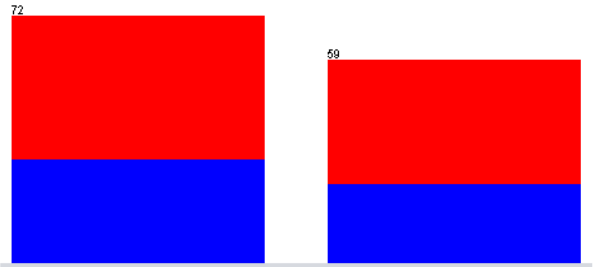

Fig. 2: Visualization of attribute 'ge' w.r.t. class variable(arr)

Fig. 3 shows the number of students in different caste categories e.g. number of students in general category are 44. The first bar in the graph shows that the number of students with reappear in general category is less than the number of students who don't have reappear in course.

The second bar shows that out of 20 students under ST category, students with reappear are more as compared to students without reappear. Similarly, the remaining bars visualize the rest of the three labels (SC, OBC and MOBC) w.r.t. the class variable 'arr'.

Name: cst
Missing: 0 (0%) Distinct: 5 Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	G	44	44.0
2	ST	20	20.0
3	SC	4	4.0
4	OBC	57	57.0
5	MOBC	6	6.0

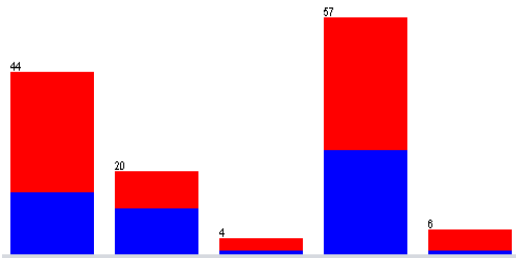

Fig. 3: Visualization of attribute 'cst' w.r.t. class variable(arr)

Fig. 4 shows the number of students with class XII %age under five different labels (Best, Very good, Good, Pass, Fail). The first bar in the graph shows that the students with label 'best' do not have any reappear in course. The second bar shows that out of 44 students under Very Good (VG) category, students with reappear are less as compared to students without reappear. Similarly, the remaining bars visualize the rest of the three labels (Good, Pass and Fail) w.r.t. the class variable 'arr'

Name: twp Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	Best	5	5.0
2	Vg	44	44.0
3	Good	65	65.0
4	Pass	17	17.0
5	Fail	0	0.0

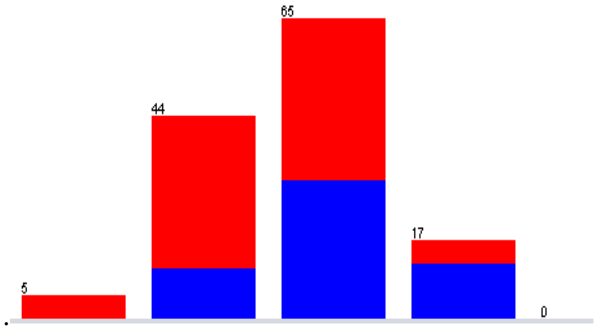

Fig. 4: Visualization of attribute 'twp' w.r.t. class variable(arr)

Fig. 5 shows the number of students with different family monthly income e.g. number of students with very high family monthly income is 6. The first and second bars in the graph shows that the number of students with very high and high family monthly income are more likely to have no reappear. Rest of the bars show that students with above medium, medium and low family monthly income are almost equal in both the cases.

Name: fmi Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	Vh	6	6.0
2	High	15	15.0
3	Am	27	27.0
4	Medium	63	63.0
5	Low	20	20.0

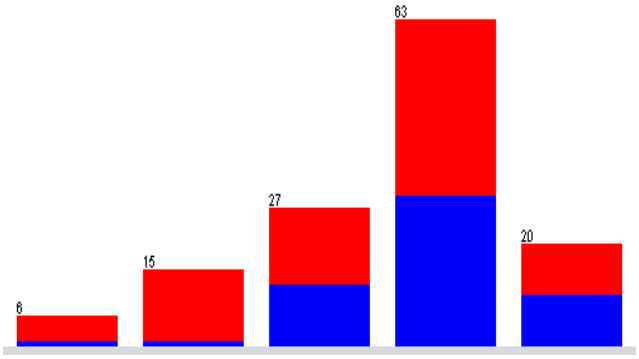

Fig. 5: Visualization of attribute 'fmi' w.r.t. class variable(arr)

Fig. 6 shows that the number of students, whose mothers are service class, are more likely not to have reappear in course. Students whose mothers are in business or retired have one instance each and do not have reappear in course. Similarly, rest of the two bars visualize the remaining two labels w.r.t. the class variable 'arr'.

Name: mo Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 2 (2%)

No.	Label	Count	Weight
1	Service	12	12.0
2	Business	1	1.0
3	Retired	1	1.0
4	Housewife	115	115.0
5	Others	2	2.0

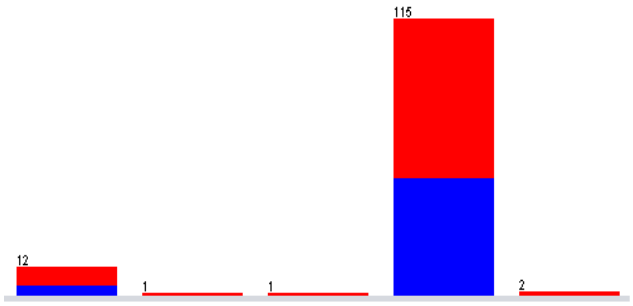

Fig. 6: Visualization of attribute 'mo' w.r.t. class variable(arr)

The first bar in Fig. 7 shows that students who have large number of friends are 58 and out of 58, more students will not have reappear in course. In the same way, rest of the bars shows the visualization of labels, average and small corresponding to the class variable 'arr'.

Name: nf Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	Large	58	58.0
2	Average	43	43.0
3	Small	30	30.0

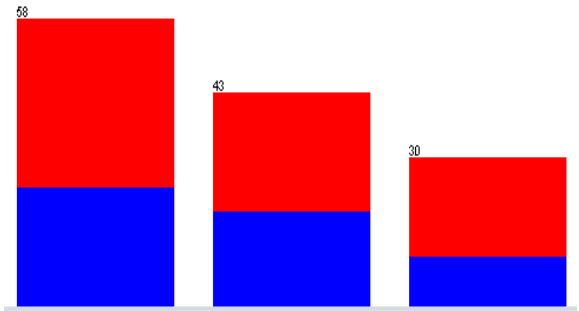

Fig. 7: Visualization of attribute 'nf' w.r.t. the class variable(arr)

The class variable arr (whether a student will have reappear in a course or not) is predicted based on these six features using the six different classification algorithms : "Naïve Bayes", "Logistic Regression", "Multilayer Perceptron", "Support Vector Machine", "J48" and "Random Forest decision tree".

V. RESULTS

The models were developed using different classification algorithms using the features extracted using Naïve Bayes classification algorithm. Six different classification algorithms are used in this paper to predict the performance of student. The six different algorithms are: Naïve Bayes, Multilayer Perceptron, Logistic regression, Support Vector Machine, J48 and Random Forest. The Multi-Layer Perceptron model is designed using one hidden layer. The comparison of the various models according to different classification algorithms are summarized in Table II. The comparison indicates that the highest classification accuracy is given by the models designed using Multi-Layer Perceptron and Random Forest classification algorithms which gives an accuracy of 92.3%. The Relative absolute Error is minimum when using Multi-Layer Perceptron (MLP) which is coming out to be 22.4%. The comparison chart (Fig. 8) clearly indicates that the best results are obtained with Multi-Layer Perceptron when using Naïve Bayes algorithm for selecting the features.

Table II: Comparison of classifiers

Classification algorithm	Classification Accuracy	Mean absolute Error	Root Mean Square Error
Naïve Bayes	75.5%	0.3872	0.4301
Logistic Regression	73.2%	0.3481	0.4177
Multi-Layer perceptron	92.3%	0.1046	0.2229
Support Vector Machine	70.99%	0.2901	0.5386
J48	74.0%	0.3515	0.4192
Random Forest	92.3%	0.1896	0.2543

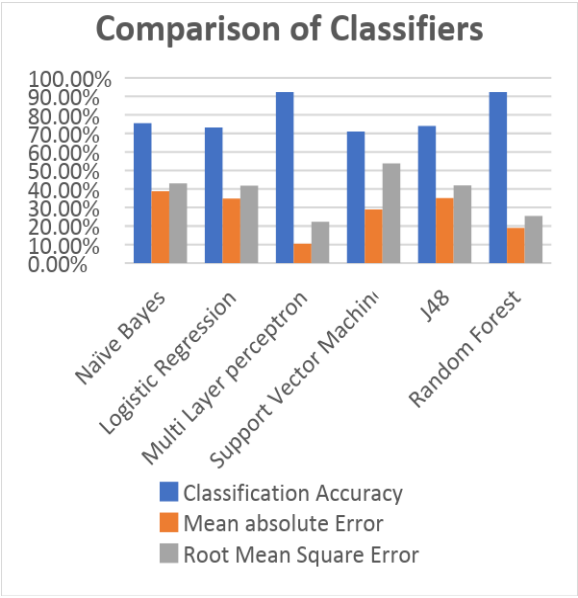

Fig. 8: Bar Chart for comparison of classifiers

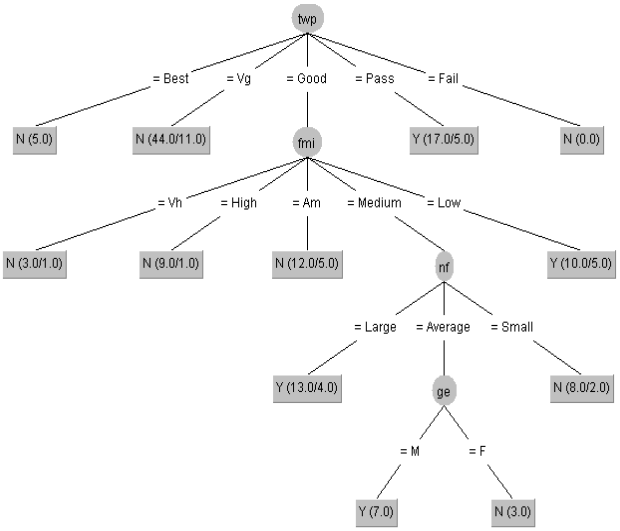

Fig. 9: Decision Tree using J48 classifier

The decision tree using J48 decision tree classifier is shown in Figure 9. When using Naïve Bayes classification for feature extraction, the precision, recall, F-measure and Area under the curve for the two most appropriate classifiers, Multi-Layer perceptron and Random Forest are shown in the next two tables. The detailed accuracy by class using Multi-layer Perceptron is shown in the Table III and the detailed accuracy by class using Random Forest is shown in the Table IV.

Table III: Detailed accuracy using MLP

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.849	0.026	0.957	0.849	0.9	0.975	Y
0.974	0.151	0.905	0.974	0.938	0.975	N

Table IV: Detailed accuracy using Random Forest

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.887	0.051	0.922	0.887	0.904	0.980	Y
0.949	0.113	0.925	0.949	0.937	0.980	N

The results show that the most appropriate classification model is designed using Multi-Layer perceptron and Random Forest with a classification accuracy of 92.3% in both the classifiers. Whereas the area under ROC curve is 97.5% in Multi-Layer Perceptron (Fig. 10) and that in Random Forest is 98% (Fig. 11).

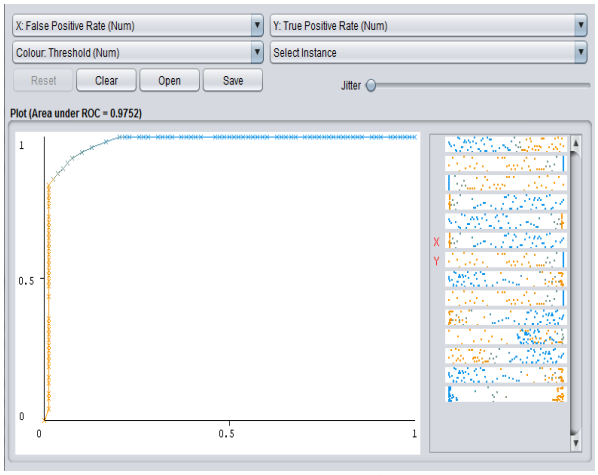

Fig. 10: Area under ROC for MLP

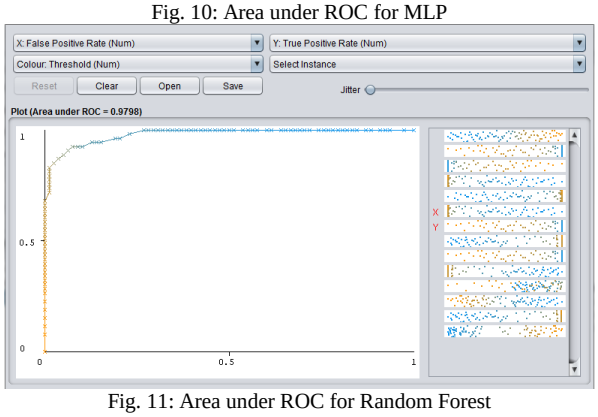

Fig. 11: Area under ROC for Random Forest

The Precision-Recall curve for multi-layer perceptron is shown in Fig. 12 and Precision-Recall curve for Random Forest is shown in Fig. 13. Comparing the F-measure and area under ROC curve of all the six classifiers, the Multi-layer Perceptron and Random Forest classifiers gives the most optimal result among all the classification algorithms.

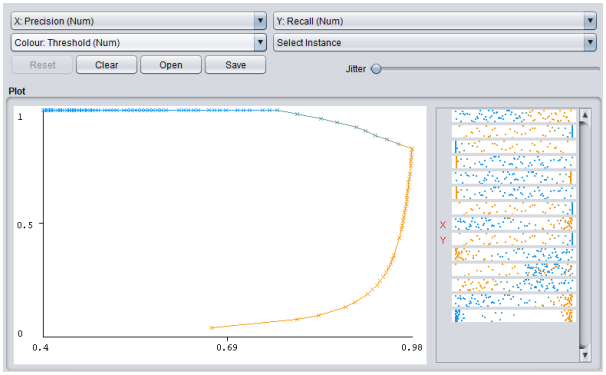

Fig. 12: Precision - Recall curve for MLP

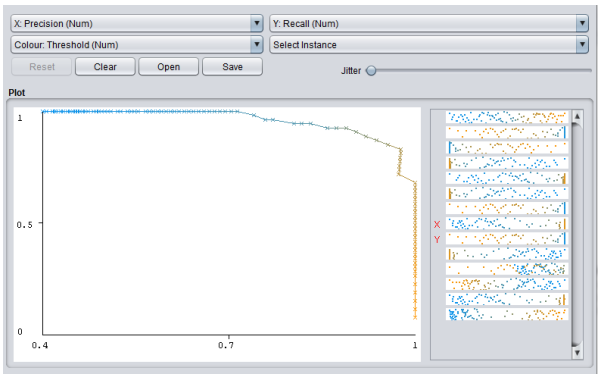

Fig.13: Precision - Recall curve for Random Forest

VI. CONCLUSION

The authors compared the outcomes of six different classifiers to identify those students who have high probability to have reappear in some of the courses, by doing predictive analysis. The investigation in this paper differs from previous related works in respect that the authors compared the accuracy of six EDM techniques to identify those students, who are likely to have reappear in course at an early stage, so that action can be taken to minimize the rate of failure at the end of course. Specifically, the six EDM techniques ("Naïve Bayes", "Logistic Regression", "Multilayer Perceptron", "Support Vector Machine", "J48 decision tree" and "Random Forest") were compared to conduct the study. These six techniques were evaluated on the data set collected from three different colleges in Assam state of India. In addition, pre-processing tasks was also performed during the realization of the experiment to select the most influential features using Naïve Bayes classification algorithm. The research results allow us to conclude that the analyzed classification algorithms are quite effective for early identification of those students who can have reappear in course, which can then be useful for parents or teachers to take a corrective action at an early stage. The prediction of student's performance can be done to a good extent by using the six features, gender, caste, percentage in XII, family income, mother's occupation and number of friends.

The two classifiers, Multi-layer perceptron and Random Forest prove to be the most appropriate classifiers for predicting student's performance. For future work, this analysis can be further taken forward by using data sets from different universities and also applying other data pre-processing techniques.

ACKNOWLEDGMENT

The authors are highly grateful to the Principal, Management and Department of Computer Science and Engineering of JSS Academy of Technical Education, Noida, Uttar Pradesh to provide complete support in carrying out the research work and writing this research paper.

REFERENCES

1. Hussain S, Dahan N.A, Ba-Alwi F.M, Ribata N., " Educational Data Mining and Analysis of Students Academic Performance Using WEKA", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 9, Issue 2, pp. 447-459, 2018
2. Kumar N., Mishra B., and Bali V. , "A Novel Approach for Blastit-Induced Fly Rock Predication Based on Particle Swarm Optimization and Artificial Neural Network", B.Tiwari et al. (Eds.), Proceedings of International Conference on Recent Advancement in Computers and Communication, Lecture Notes in Networks and Systems 34, Springer Nature Singapore Pvt. Ltd, Chapter 3, Book Id: 448040_1_En, ISBN: 978-981-10-8197-2, 2018
3. Shuja Mirza, Sonu Mittal, Majid Zaman, "Design and implementation of predictive model for prognosis of diabetes using data mining techniques", International Journal of Advanced Research in Computer Science. Vol. 9, Issue 2, pp-393-398, 2018
4. Shuja Mirza, Sonu Mittal, Majid Zaman, "Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree", International Journal of Applied Engineering Research,Vol. 13, Issue 11, pp. 9277-9282,2018

5. Evandro B.Costa, Baldoino Fonseca, Fabrísia Ferreirade and Araújo, Joilson Regod , "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", Computers in Human Behavior, ELSEVIER, Vol. 73, pp. 247-256,2017
6. Almarabeh, H., "Analysis of Students' Performance by Using Different Data Mining Classifiers" I.J. Modern Education and Computer Science, Vol. 9, Issue 8, pp. 9-15, 2017
7. K. Govindaswamy and T. Velmurugan, "A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance Prediction", International Journal of Control Theory and Applications, Vol. 10, Issue 23, 2017
8. Anuradha, C. and 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, Issue 15, pp. 1-12,2015
9. Jha P.C., Bali V., Narula S. and Kalra M., "Optimal Component Selection Based on Cohesion and Coupling for Component based Software System under Build-or-Buy Scheme", Journal of computational sciences, ELSEVIER, Vol. 5, pp 233-242, ISSN: 1877-7503. (SCI Impact Factor: 1.748, SCOPUS: 0.481), 2014
10. Patil, T. and S.S. Sherekar, "Performance Analysis of Naïve Bayes and J4.8 Classification Algorithm for data classification", International Journal of Computer Science and Applications, Vol. 6, Issue 2, pp. 256-261, 2013
11. Jha P.C. and Bali V. (2012), "Goal Programming Approach for Selection of COTS Components in Designing a Fault Tolerant Modular Software System under Consensus Recovery Block Scheme", International Journal of Computer and Communication Technology (IJCCT), Vol. 3, No. 1, pp. 1-8, ISSN (Online) 2231-0371, (Print) 0975-7449, 2012
12. Dekker, G.W., M. Pechenizkiy, and J.M. Vleeshouwers, "Predicting students drop out: A case study", EDM '09-Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, pp. 41-50, 2009
13. C. Romero and S. Ventura, " Educational data mining: A survey from 1995 to 2005", ELSEVIER, Expert Systems with Applications, Vol. 33, Issue 1, pp. 135–146, 2007

AUTHORS PROFILE

Deepti Aggarwal has received her B. Tech. (CSE) from MDU, Rohtak, M.Tech. (CSE) from Rajasthan Vidyapeeth, Udaipur and pursuing Ph.D. from Jaipur National University, Jaipur. She has overall teaching experience of more than 18 years and currently working as Assistant Professor at JSS Academy of Technical Education, Noida since 2005. She has 4 papers in international journals and conference. She has written a book on 'Computer Organisation'. Her area of interest is machine learning, data mining, operating system and compiler design.

Dr. Vikram Bali has received his B.Tech (CSE) from REC, Kurukshetra, M.E. (CSE) from NITTTR, Chandigarh and Ph.D from Banasthali Vidyapith, Rajasthan. He has more than 18 years of rich academic experience. He is a Professor & Head of Department (CSE) at JSS Academy of Technical Education, Noida. He is life time member of Indian Society for Technical Education (ISTE), Computer Society of India (CSI) and Institution of Engineers (IE). He has contributed 21 Research papers in International Journal and 7 Research papers in National Conferences/ proceedings and Edited Books. He has also attended Faculty Enablement programme organised by Infosys and NASSCOM. He has been the member of board of studies of different Indian Universities and member of organizing committee for various national and international seminar/conferences. He has written books on Fundamental of "Cyber Security and Laws", "Software Engineering" and "Operating System". He is reviewer to many International Journals of repute like Inderscience and IGI Global. His research interest includes Software Engineering, Cyber Security, Automata Theory, CBSS and ERP.

Dr. Sonu Mittal received his Ph.D. from SGV University, Jaipur. Prior to that, he received his Master's Degree (M. Tech.) from IGNOU. He is currently working as Associate Professor, Department of Computer Science and Engineering at Jaipur National University, Jaipur since 2008. He has more than 14 years of research and teaching experience. He has more than 20 publications in national/international journals and conferences. His area of interest include Machine Learning, Software Engineering and Computer Networks.