ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 6|| Jun – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Evaluation Of Various Feature Selection Algorithms In Educational Data Mining

N. Sai Sragvi Vibhushan1, Vikas.B2

[1] *UG Student, Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam, India*

[2] *Assistant Professor, Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam, India*

Correspondence Author: N. Sai Sragvi Vibhushan

ABSTRACT

Educational Data Mining is evolving as an important field of research that helps in predicting the performance of the students. Eventually, an educational institution can plan some strategies by the results obtained from an EDM process and improve the performance of the student. In EDM, prediction accuracy is the crucial concern. A feature selection algorithm removes the extraneous data and helps in increasing the accuracy of the classifier. Just a mere feature selection algorithm might not be completely helpful and hence to improve the performance of the classifier we perform an ensemble method on the classifier. An ensemble method produces different models and combines them to produce improvised results. We then evaluate various feature selection algorithms with their respective accuracies and then conclude the best possible feature selection algorithm for various classifiers on which the ensemble method is applied, for the data set.

KEYWORDS: Accuracy, Classifiers, Ensemble methods, Educational Data Mining, Feature Selection, Boosting, Prediction.

Date of Submission: 06-06-2018 Date of acceptance: 21-06-2018

I. INTRODUCTION

Education plays a key role in the progress of a nation [1]. An educational institution can use the Educational data mining and accordingly plan some strategies and help to improve the performance of students. Now-a-days, the usage of data mining by institutions being increasing drastically [2,3]. By applying the techniques on the student's data, interesting patterns and hidden knowledge of the student's performance can be generated [4].

In any data mining process, preprocessing is a crucial step. In this paper, we preprocess the data by applying feature selection algorithm to the data set. Feature selection is a process of selecting a subset of relevant features. There are various feature selection algorithms that can be used in an educational data mining task [5].

After the preprocessing, we now apply various classifiers to predict the performance of the students. Classification is a process where, different target classes are obtained by the division of data inputs [6]. The application of only classifiers might not give sufficient accuracy of prediction. So, we use a ensemble method named boosting to the classifiers. Boosting is an ensemble method that helps in increasing the accuracy of the classifier. By the application of an ensemble method to the classifiers increases the accuracy. When boosting is applied to the classifiers there is a change in the accuracy. Thus, in this paper we evaluated various classifiers by considering various measures.

Rest of the paper is organized as follows, Section I contains the introduction of Education Data Mining and Preprocessing of the data, Section II contain the related work of Performance analysis of various feature selection algorithms, section III explains the methodology with flow chart, Section IV describes results and discussion of the entire paper and Section V concludes research work with future directions.

II. RELATED WORK.

As an emerging field of research there has been a lot of work done in educational data mining. Some of the previous works include the analysis and comparison of performance of various feature selection algorithms that can be used in educational data mining.

Selecting optimal subset of features for student performance model, by H. M. Harb and M. A. Moustafa stated that the performance of a feature selection algorithm can be improved if we generate a hybrid selection algorithm by combining the filter and wrapper methods [6].

Modeling and Predicting Students Academic Performance Using Data Mining Techniques by A. Mueen, B. Zafar, and U. Manzoo , applied data mining techniques on a student's data and predicted whether the student will pass or fail in the semester [7].

A. Figueira, predicted the grades of the students by using the principal component analysis algorithm on student's dataset [8].

Performance Analysis of Feature Selection Algorithm for Educational Data Mining by Maryam Zaffar, Manzoor Ahmed Hashmi, K.S.Savita compared the performance of the combination of various feature selection algorithms with a set of classifiers and stated the best possible combination for the students data set [9].

Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods by Elaf Abu Amrich,Thair Hamtini and Ibrahim Aljarah, performed Information Gain attribute selection and then applied some classifiers along with the ensemble methods on the students dataset. The application of the ensemble methods helped in increasing the accuracy of the model. The accuracy of the model was higher when behavioural features were included [10].

A Review on Predicting Student's Performance Using Data Mining Techniques,by A. M. Shahiri and W. Husain, analysed the performance of data mining classifiers namely,Naïve Bayes, Decision tree, Neural Network,K-Nearest Neighbor,Support Vector Machine(SVM). They also stated that CGPA has been an important feature in predicting the performance and the measure of accuracy [11].

III. METHODOLOGY.

3.1 Data Collection.

The student's dataset used here is obtained from Kaggle.com. There are around 16 attributes and 480 instances in this dataset. The dataset includes students from different origins and is of two semesters. The dataset consists of 245 students of first semester and 235 students of second semester. The performance of the students is classified into L(low-level) which includes values from 0-69, M(middle-level) which includes values from 70-89, H(high-level) which includes values from 90-100.

3.2 Data Preprocessing.

In any data mining model, the initial step is the preprocessing of the data. The data goes through the following series of steps during preprocessing: Data Cleaning, Data Integration, Data Transformation, Data Reduction, Data Discretization.

3.2.1 Data Visualization.

Data Visualization aims to communicate the raw data clearly through graphical representation. The user can learn some interesting facts about the data through these representations. One can also discover a relationship, if there exists any, between the features of the dataset. We present the visualization of this dataset using WEKA tool.

WEKA short for Waikato Environment for Knowledge Analysis, is an open source software implemented in Java. It was developed by University of Waikato in New Zealand. We can implement various data mining techniques using this tool [12].

The features of the dataset can be classified into three categories namely demographic features such as gender and nationality, academic background features such as educational stage, grade Level and section and behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction. We now represent each of these features graphically.

The dataset consists of 305 males and 175 females and that information is represented in the Figure 1.

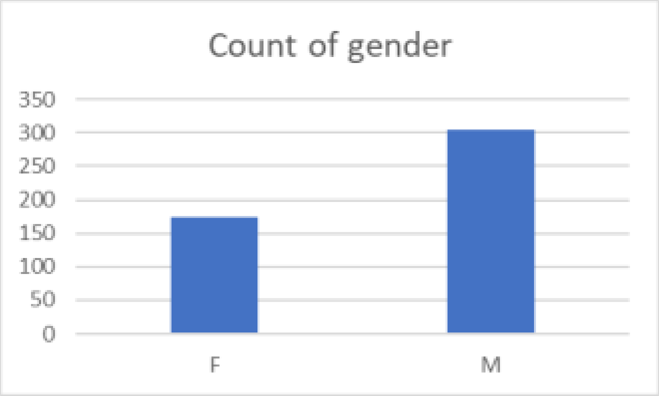

Figure 1. Gender Feature Visualization

The number of students from each origin is represented in a table.

Table 1. Stduents from different origin

Origin	Number of Students
Kuwait	170
Jordan	172
Palestine	28
Iraq	22
Lebanon	17
Tunis	12
Saudi Arabia	11
Egypt	9
Syria	7
USA	6
Iran	6
Libya	6
Morocco	4
Venezuela	1

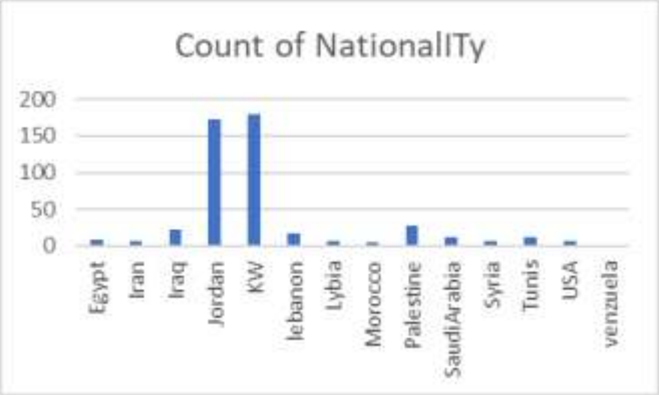

Figure 2. Nationality Feature Visualization

From the above figure we can depict a hidden impact on the students due to the diversities. Figure 3, gives visualization on the topics that have been chosen by the students.

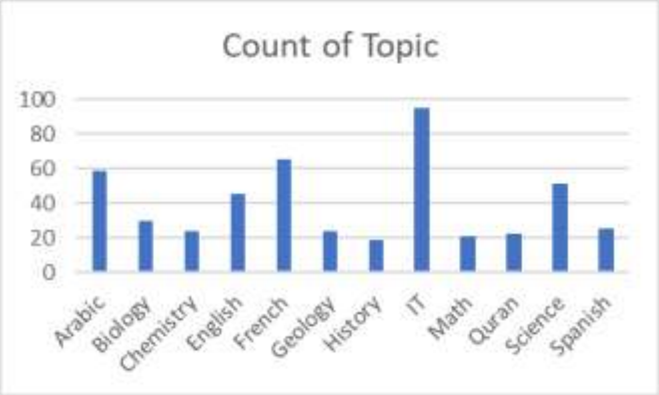

Figure 3. Topic Feature Visualization

From the above figure we can say that the performance of the student might depend on the topic chosen. Each student in the given dataset is followed either by their father or mothers.283 and 197 students are followed by their fathers and mother respectively. This data set has also an attendance feature. This is classified by the number of absence days of the student. This feature undoubtedly will have an impact on the performance of the student. This information is visualized in Figure 4.

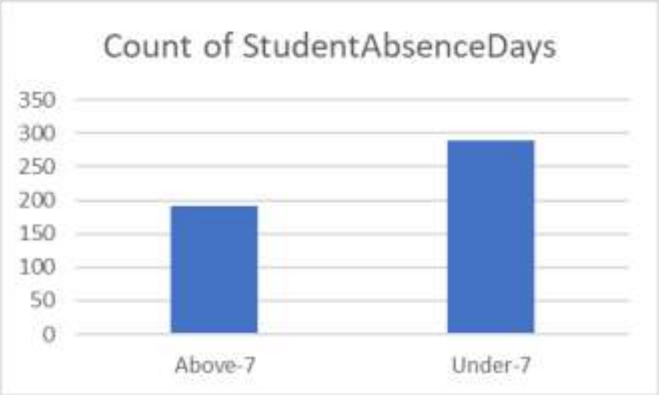

Figure 4. StudentAbsenceDays Feature Visualization

There is also a feature regarding a survey taken on the satisfaction of the parent on the school. 292 parents have mentioned that they are satisfied with the school, where as 188 parents have stated the opposite.

3.2.2 Feature Selection.

It is the selection of attributes that are most relevant to the predictive modeling problem. This helps in creating an accurate predictive model by choosing features that aid in giving better accuracy.

In this paper, we apply the feature selection algorithms namely Principal Component Analysis, ReliefAttributeEval and CfsSubsetEval. The Ranker Search method is chosen for the PCA and ReliefAttributeEval and the Best First search is chosen for the CfsSubsetEval.

After the preprocessing of the data, various classifiers were applied to predict the performance of the student. We have taken the Naïve Bayes, Multilayer Percepton , Sequential Minimal Optimization and J48 classifiers. We analyzed the performance of these classifiers by the accuracy it obtained.

Eventually, boosting is used to improve the performance of the classifiers.

The performance of the classifier is evaluated by the prediction accuracy, which is calculated using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

The above formula includes the following terms:

TP denotes the number of positive tuples that were correctly classified.

TN denotes the number of neagative tuples that were correctly clasified.

FP denotes the number of negative tuples that were incorrectly classified as positive.

FN denotes the number of positive tuples that were incorrectly classified as negative.

The values of TP, TN, FP and FN are taken from the confusion matrix that is obtained by the application of the classifiers.

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

Table 2 shows the representation of the confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive(TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

Table2. Confusion Matrix

All the preprocessing, classification and the application of boosting is done with the help of WEKA tool.

Figure 5 summarizes the steps that we have implemented in the methodology.

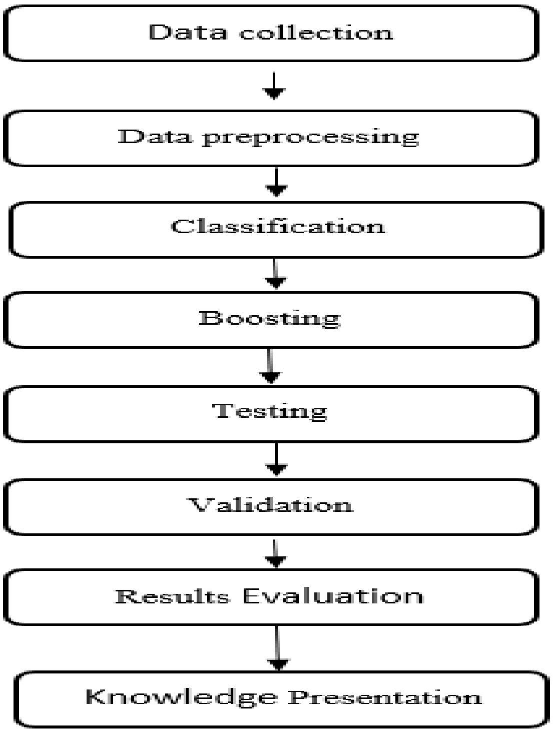

Figure 5. Steps implemented in methodology

IV. RESULTS AND DISCUSSIONS.

4.1 Principal Component Analysis.

PCA is a procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components. The main idea is to retain the variation present in the dataset, up to the maximum extent. Thus, the first principal component accounts for as much of the variability in the data as possible.

Table 3. Performance Evaluation of Principal Component Analysis

Classifiers	**Only**	**Boosting**
Naïve Bayes	57.38%	64.375%
Multilayer Perceptron	74%	77.0833%
SMO	72.2917%	75.3%
J48	65.416%	68.125%

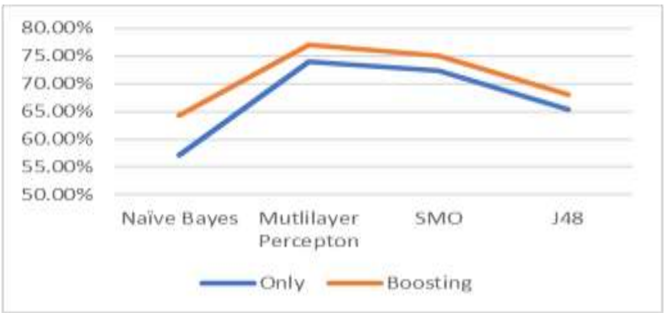

Figure 6. Graphical representation of the performance

The above table distinguishes the accuracy of Naïve Bayes, Multilayer Percepton , SMO, J48 classifier when applied without boosting and with boosting on the students data set. From the table we can infer that the best classifier for a PCA feature selection is the Multilayer percepton.

4.2 ReliefAttributeEval.

Relief uses a filter based approach for the feature selection. It selects the top scoring features on the basis of the score that is assigned to the feature.

Table 4. Performance evaluation of ReliefAttributeEval

Classifiers	Only	Boosting
Naïve Bayes	67.7083%	72.2917%
Multilayer Percepton	77.375%	79.375%
SMO	78.75%	80.20%
J48	75.625%	78.75%

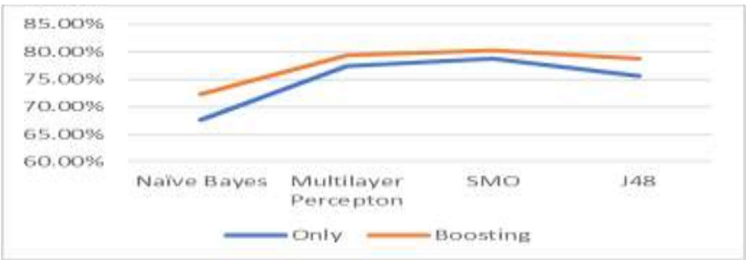

Figure 7. Graphical representation of the performance

The above table distinguishes the accuracy of Naïve Bayes , Multilayer Percepton , SMO,J48 classifier when applied without boosting and with boosting on the students data set. From the table we can infer that the best classifier for a ReliefAttributeEval feature selection is the SMO classifier.

4.3 CfsSubsetEval.

It evaluates a subset of attributes on the individual predictive ability of each feature along with the degree of redundancy between them. Preference is given to the subsets that are highly correlated with class.

Table 5. Performance evaluation of CfsSubsetEval

Classifiers	Only	Boosting
Naïve Bayes	69.38%	72.50%
Multilayer Percepton	74.38%	75.80%
SMO	77.04%	79.40%
J48	76.04%	78.2%

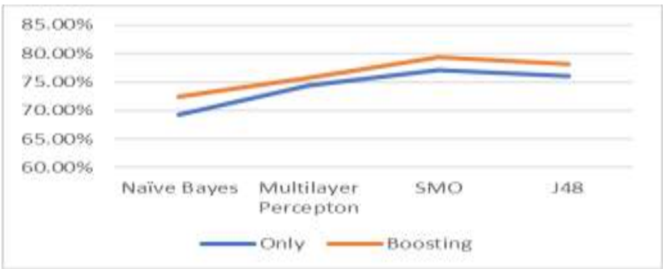

Figure 8. Graphical representation of the performance

The above table distinguishes the accuracy of Naïve Bayes, Multilayer Percepton , SMO,J48 classifier when applied without boosting and with boosting on the students data set. From the table we can disclose that the best classifier for ReliefAttributeEval feature selection is the SMO classifier.

V. CONCLUSION AND FUTURE SCOPE.

The academic performance of students plays a key role in their career development. In order to achieve better progress in the performance one needs some guidance and help. Therefore, predicting the performance beforehand, will help the educational institution plan strategies accordingly. In this study, we have designed a classification model that analyses the performance of various classifiers namely Naïve Bayes, Sequential Minimal Optimization, J48 and Multilayer Percepton with and without the ensemble method, Boosting. The evaluation is done on the basis of their respective prediction accuracy. The accuracy of 78.75% derived by the SMO classifier has been marginally better than other classifiers. When the same classifier is performed along with the boosting the accuracy has been improved to 80.20%. This has been obtained by using ReliefAttributeEval feature selection algorithm. This accuracy is slightly higher than the other combinations that were evaluated. The best performance of the ensemble method, Boosting has been derived with the Naïve Bayes classifier. We have noticed an increase of nearly 7% of accuracy from 57.38% to 64.375% when Boosting has been applied classifier along with PCA. This prediction helps the educational institutions to understand what factors influence the performance of a student.

In future, advanced ensemble methods can be applied on this dataset to increase the accuracy of the classifiers.

REFERENCES

[1]. Gaviria, A. Los quesuben y los quebajan: educacion y vilidad social en Colombia. Fedesarrollo, Alfaomega.(2002)

[2]. C. Romero and S. Ventura, Educational data mining: a review of the state of the art, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40:601-618, 2010.

[3]. A. M. Shahiri and W. Husain, A review on predicting student's performance using data mining techniques, Procedia Computer Science,72:414-422, 2015

[4]. E. Osmanbegović, M. Suljić, and H. Agić, Determining Dominant Factor For Students Performance Prediction By Using Data Mining Classification Algorithms, Tranzicija,16:147-158, 2015.

[5]. M. Ramaswami and R. Bhaskaran, A study on feature selection techniques in educational data mining, arXiv preprint arXiv:0912.3924, 2009.

[6]. Vikas B, Sipra Sarangi, Manaswini Chilla, K Santosh Bhargav, B S Anuhya. A Literature Review on The Rising Phenomenon PCOS. International Journal of Advances in Engineering & Technology,2(10), pages216-224,2017.

[7]. H. M. Harb and M. A. Moustafa, Selecting optimal subset of features for student performance model, Int J Comput Sci, pages5, 2012.

[8]. A. Mueen, B. Zafar, and U. Manzoor, Modeling and Predicting Students' Academic Performance Using Data Mining Techniques, International Journal of Modern Education and Computer Science, 8:36, 2016.

[9]. A. Figueira, Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analytics, in Advanced Learning Technologies (ICALT), IEEE 16th International Conference on, pages465-467,2016.

[10]. Elaf Abu Amrich,Thair Hamtini and Ibrahim Aljarah,Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods, International Journal of Database Theory and Application ,9:119-136,2016.

[11]. Maryam Zaffar, Manzoor Ahmed Hashmani, K.S.Savita, Performance Analysis of Feature Selection Algorithm for Educational Data Mining,2017 IEEE Conference on Big Data and Analytics(ICBDA).

[12]. A. M. Shahiri and W. Husain, A Review on Predicting Student's Performance Using Data Mining Techniques, Proceeding Computer Science,72:414-422, 2015.

[13]. M. Hall, E. Frrank, G.Holmes, B.Pfahringer, P. Reutemann, and I. H. Witten, The WEKA data mining software: an update, ACM SIGKDD exploration newsletter, 11:10-18,2009.

N. Sai Sragvi Vibhushan." Evaluation Of Various Feature Selection Algorithms In Educational Data Mining.” International Journal of Computational Engineering Research (IJCER), vol. 08, no. 06, 2018, pp. 19-25.