

# A Majority Vote Based Ensemble Classifier for Predicting Students Academic Performance

A.Salini <sup>1</sup>, U.Jeyapriya <sup>2</sup>

<sup>1</sup>PG Scholar, Department of Computer Science,  
Stella Maris College, Chennai,

<sup>2</sup>Asst.Professor, Department of Computer Science,  
Stella Maris College, Chennai,

salinianand14@gmail.com <sup>1</sup>, jeyapriya75@gmail.com<sup>2</sup>

March 22, 2018

## Abstract

Education is the backbone for the development and success of every countrys economy and industry. It is important to extract the hidden knowledge from the educational data. The knowledge extracted from these educational data helps to improve the training methods and the learning process which in turn enhances the performance of the students and the overall educational outcomes. In this paper, we propose a new students performance prediction model using features associated with students behaviour. These type of features are based on the learners interaction with e-learning management system. The performance of students predictive model is evaluated by set of classifiers, namely; Support Vector Machine, Logistic Regression and Random Forest and have found that Majority Voting ensemble method help us to arrive at a better performance accuracy.

**Key Words:**Educational Data Mining (EDM), Logistic Regression, Majority Voting Ensemble, Random Forest, Support Vector Machine (SVM).

## 1 INTRODUCTION

Due to the widespread increase in the development of online learning courses or web-based education a large amount of educational data is being generated. The main challenge lies in transforming these educational data into useful knowledge. By making use of data mining techniques like clustering and classification, these problems can be addressed. Educational Data Mining (EDM) refers to the techniques or methods used to extract significant information from large quantities of educational data. EDM is used in a variety of fields like predicting the performance of the students and identifying the students at-risk, raising the graduation rates in universities, assessing the performance of the educational institutions and comparing it with the performance of other institutions to find its strength and weaknesses, maximizing campus resources, and analyzing the students performance based on their demographic and interactive behaviors [8].

Different classification methods have been used for predicting students performance, such as Decision Tree, Rule Induction, Nave Bayes and function-based (e.g., Neural Network, Logistic regression, Support Vector Machine). This study compares different classification algorithms for predicting students performance and majority voting ensemble method is used to improve the performance by combining the prediction of different classifiers.

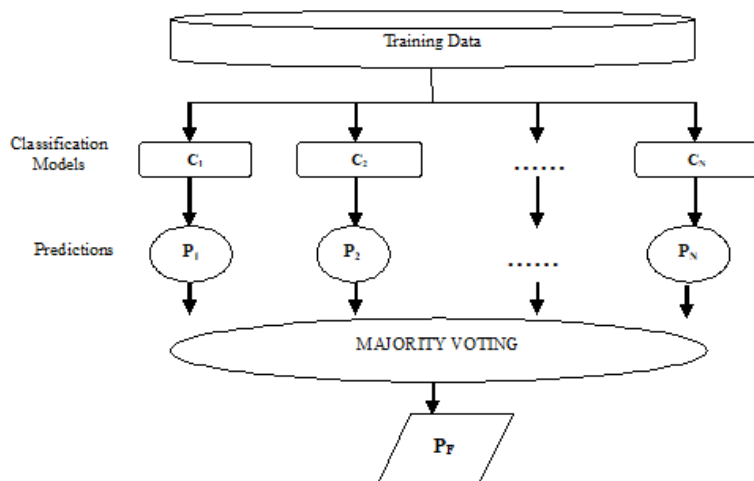


Fig.1. Majority Voting Model

This paper is organized into 5 sections. Section 1 contains the introduction to Educational Data Mining (EDM). The related works done in EDM is discussed in section 2. Section 3 presents the methodology and the various machine learning algorithms used in the proposed work. Section 4 contains the experimental results and in section 5 it is concluded.

## 2 RELATED WORK

S. Agarwal, et al.[3] has found that the performance of the students in the test and qualifying examination are the key factors that affects their placement. In this research work, a comparative study was done on the performance of the different kernel types of the Support Vector Machine algorithm and it was found that the Radial Basis Kernel achieved a higher accuracy than the polynomial, sigmoid and the linear kernel.

Pandey & Sharma, [4] has developed a classification model for predicting the grades of the undergraduate students using different classification algorithms like NB tree, J48 and Reptree. In this research paper, it was found that J48 achieves a higher accuracy when compared to the other classification techniques.

M.S. Mythili , et al. [5] has suggested some of the possible factors affecting the performance of students in exams like attendance, parent education, etc. In this paper, the performance of Random forest, J48, IBI and decision table algorithms are compared. The results obtained prove that Random forest algorithm is more accurate and the time taken to develop the model is very less when compared to the other classifiers.

L.Pandeeswari, et al. [6] has proposed a comparative study on the performance of the nave Bayes algorithm with an integrated model by combining the k-means clustering algorithm with the nave Bayes classification technique. It was found that the integrated model performs better with a dataset having missing values.

Elaf Abu Amrieh, et al.[7] has predicted the performance of the students based on their behavioral features like discussion groups visited resources, raised hand and viewing announcements using classification techniques like ANN, Nave Bayes and Decision tree .In this research work ensemble methods like Boosting, bagging and random forest are used for improving the performance of these classifiers.

### 3 METHODOLOGY

The steps involved in the methodology are graphically represented below in fig.2.

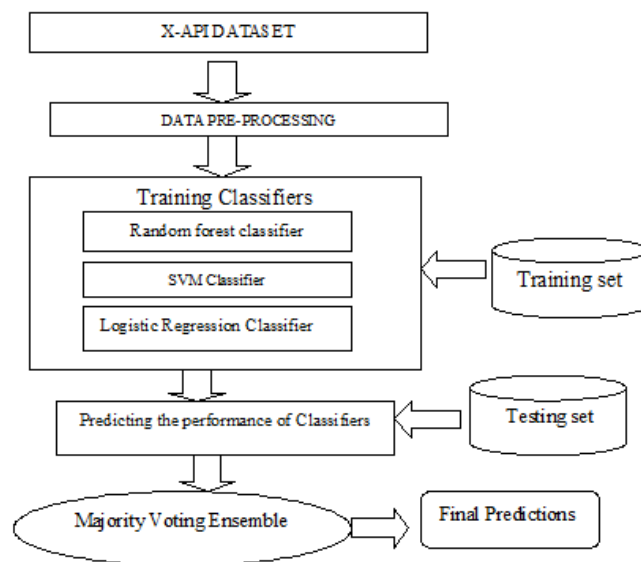


Fig 2.

In this paper, we use a Majority voting based ensemble method for predicting the students performance. Ensemble method is a learning approach that combines multiple models to solve a problem. In contrast to traditional learning approaches which train data by one learning model, ensemble methods try to train data using a set of models, and then combine them to take a vote on their results. The predictions made by ensembles are more accurate than predictions made by a single model [7]. The aim of such approach is to provide an accurate evaluation for the features that may have an impact on students academic success. In this paper, the performance of several classification techniques like SVM, Random forest and Logistic regression are compared. A majority vote based classification algorithm is used for improving the performance by combining the results of the above three classification techniques.

### 3.1 EDUCATIONAL DATASET

In this research work, the students academic performance dataset X-API was taken from kaggle.com. This dataset contains 16 features of 500 students. These features can be classified into four

categories as shown in table 1.

| FEATURES               | CATEGORY                                                                                |
|------------------------|-----------------------------------------------------------------------------------------|
| DEMOGRAPHICAL FEATURES | NATIONALITY<br>GENDER<br>PLACE OF BIRTH<br>PARENT RESPONSIBLE FOR STUDENT               |
| ACADEMIC FEATURES      | EDUCATIONAL STAGES<br>STAGE<br>GRADE LEVELS<br>SECTION ID<br>STUDENT ABSENCE DAYS       |
| PARENTAL FEATURES      | PARENT ANSWERING SURVEY<br>PARENT SCHOOL<br>SATISFACTION                                |
| BEHAVIORAL FEATURES    | DISCUSSION GROUPS<br>VISITED RESOURCES<br>RAISED HAND ON CLASS<br>VIEWING ANNOUNCEMENTS |

TABLE 1

## 3.2 Data Preprocessing

The data set is standardized and used for the proposed work. The standardization of data is done through 0-1 normalisation. Data Preprocessing involves a) Data Cleaning and b) Feature Selection.

### 3.2.1 Data Cleaning

Data cleaning is one of the important preprocessing tasks which help to eliminate irrelevant data items and missing values. In this research work the dataset contains 4% of missing values; therefore

the records with missing values are detached from the dataset. Finally, the dataset is reduced to 480 observations after performing data cleaning.

### **3.2.2 Feature Selection**

In data preprocessing feature selection is an important step, the purpose of this method is to identify a set of features which can efficiently describe the input data, reduce the dimensions of the feature space and remove the redundant and unsuitable data. Feature selection helps to improve the data quality and the performance of the machine learning algorithms [9]. In this research work, the Interactional features were considered for research in the field of Learning Analytics and Educational Data Mining as interactional features help in identifying the interest and behavior of the students.

## **3.3 Classification Algorithms**

### **3.3.1 Random Forest Algorithm (RF):**

Random forest is a machine learning algorithm that reports the aggregated result of different classifiers. Random forest algorithm is used for both regression and classification. In Random forest, several classification trees are created and each tree is trained in a bootstrap sample of the original training data. Every tree in the RF will cast a vote for some input  $x$ . Then the output of the RF will be determined based on the majority voting of the trees. High dimensional data can be handled by the Random forest classifiers [2]. Random forest is an ensemble classifier technique that includes many decision trees. The output of random forest algorithm is the class that is the mode of the classes output by individual trees. Random forest over fits with few datasets.

### **3.3.2 Support Vector Machine (SVM):**

A support vector machine is a supervised machine learning method with associated learning. It is used to analyze and recognize patterns in data. SVM is mainly used for both regression and classification. SVM is a binary classifier and each training data is assigned

to either of the two classes. A new model is constructed from the SVM algorithm which is then used to categorize a new instance into either one class or another [3]. Since SVM supports Kernel functions, it can learn various combinations of the given features with less computational complexity. In SVM the problem of overfitting is eliminated.

### 3.3.3 Logistic Regression:

In Logistic regression the dependent variable is dichotomous. The independent variables in this model are predictors of the dependent variable and can be measured on a nominal, ordinal, interval or ratio scale. In logistic regression the association between the independent and the dependent variable is nonlinear [1].

## 3.4 Majority Voting:

In Majority voting, several classifiers generate predictions for each instance in the testing data. The final prediction for each instance is the one which receives the maximum votes and it more than half of the votes. The pseudocode of the Majority Voting used in the proposed work is as follows:

| PSEUDOCODE 1:<br>Majority Vote based Ensemble Classifier                                                                                                                                                                                                                                                                                        |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Step 1: Apply 3 classifiers (Random Forest, SVM and Logistic Regression) on the training data.<br>Step 2: Compare the performance of the 3 classifiers.<br>Step 3: Performing Majority Voting for every observation.<br>Step 4: Compare the performance of the Majority Voting with the Random Forest, SVM and Logistic Regression classifiers. |

Majority Vote Based Ensemble Classifier method increases the accuracy by combining the advantages of each individual classifier.

## 4 EXPERIMENTAL RESULTS

In this research work, we are predicting the performance of the students based on their behavioral and parental features using a

majority vote based ensemble method. The dataset after cleaning contains 480 records. 70% of the data is used for training the classifiers and 30% for testing. The performance of SVM, Random forest and logistic regression is compared and it was found that Random Forest acquires a higher accuracy when compared to SVM and Logistic Regression.

| EVALUATION MEASURES | RANDOM FOREST | SVM  | LOGISTIC Regression | MAJORITY VOTING |
|---------------------|---------------|------|---------------------|-----------------|
| PRECISION           | 81.7          | 81.3 | 78.8                | 82.3            |
| RECALL              | 78.9          | 78.7 | 76.9                | 80.5            |
| F-MEASURE           | 79.3          | 79.2 | 77.2                | 81.01           |
| ACCURACY            | 78.8          | 78.6 | 76.4                | 80.90           |

TABLE 2

The above results show that the majority voting based ensemble method achieves a higher accuracy rate when compared to the individual performance of SVM, Random Forest and the logistic regression classification technique. This ensemble method increases the accuracy rate by combining the predictions of several individual classifiers and selecting the prediction having the maximum vote as the final output prediction for each instance.

## 5 CONCLUSION

In this paper, we have proposed a majority voting based ensemble for the prediction of student performance. It was found that majority voting based ensemble method achieves a higher accuracy when compared to the SVM, Random forest and the logistic regression techniques. As the accuracy of the prediction increases, it will be of immense help to improve the performance of students. Hence, this work will be of great help in predicting the student performance which in turn will help the institutions and academicians to analyse the reasons for the performance and improve them.

## References

- [1] E.Yesilnacara,T.Topal, Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey), Engineering Geology, 2005 Elsevier.
- [2] Mohammed Khalilia , Sounak Chakraborty, Mihail Popescu, Predicting disease risks from highly imbalanced data using random forest , BMC Medical Informatics and Decision Making 2011. //random forest
- [3] Agarwal.S, Pandey.G. N & Tiwari, M. D, " Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.
- [4] Pandey, M & Sharma, V. K, " A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction", International Journal of Computer Applications (0975 8887) Volume 61 No.13, January 2013.
- [5] Mythili, M. S & Shanavas, A. R. M, "An Analysis of students performance using classification, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. III (Jan. 2014), PP 63-69 .
- [6] L.Pandeeswari , K.Rajeswari, "K-Means Clustering and Naive Bayes Classifier For Categorization Of Diabetes Patients ", IJISSET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015.
- [7] Elaf Abu Amrieh, Thair Hamtini and Ibrahim Aljarah, Mining Educational Data to Predict Students academic Performance using Ensemble Methods, International Journal of Database Theory and Application Vol.9, No.8 (2016),September 2016.
- [8] Abdulmohsen Algarni, Data Mining in Education, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016.

- [9] Sanyam Bhararal, Sai Sabithal, Abhay Bansall, Application of learning analytics using clustering data mining for students disposition analysis, springer Educ Inf Technol 2017.
- [10] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.