International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Enhancing Academic Success Prediction: An Ensemble Approach

Dr. Ashok M V[1], Dr. Safira Begum[2]

[1,2]Dept. of Computer Applications, HKBKDC, Bangalore, Karnataka, India

Abstract

The process of extracting meaning and knowledge from large volumes of data is termed as Data mining. It also refers to the way of inferring information from databases, which can be used in diverse fields including educational domain. Educational data mining plays a key role in finding ways to discover knowledge from data in the education sector. Educational Data Mining has evolved as a significant element in prediction of students' academic performance. The most important goal of the paper is to analyze and evaluate the engineering students' performance by applying stacking, an ensemble method in orange tool. Ensemble Stacking combines several base classifier models in order to create one optimal prediction model. Engineering students' dataset was used to build predictive model using traditional classifiers SVM, Logistic Regression, Naïve Bayes and then stacking technique was implemented. The results showed that the proposed stacking technique obtains a high performance, which has a superior result compared to the other base classifiers techniques. Therefore, conclusion could be reached that the stacking performance is better than that of different algorithms.

Keywords: educational data mining, classification, ensemble method, stacking.

Introduction:

There has been a rapid increase in the amount of data generated and stored in educational databases, which has created a dire need for analyzing such datasets and infer meaning from it[1]. Educational data mining focuses on developing techniques for discovering knowledge from extensive data related to learning activities in educational environment. Exploring these huge datasets, by using multiple data mining techniques, makes it is possible to find unique patterns that provide insights in studying, predicting and improving the students' academic performance[2].

Educational institutions are looking at educational data mining to predict student performance and behavior to enhance curriculum design, assessments and accordingly plan interventions in order to support and guide the students more effectively. In data mining, there are a number of methods applied to educational datasets such as classification, clustering, and association rules. These methods help in processing the educational data to predict performance of students, grades or even the risk of dropping out of courses.

Classification is a type of supervised learning technique that classifies the data items into specific predefined class labels with an aim to predict the future output based on the training datasets available[3].Decision trees, Naïve Bayes classifiers, and artificial neural networks are some of the commonly used techniques to predict and classify various factors of student performance and such techniques are often called educational data mining techniques[4].Accuracy levels and Confusion matrices

along with the execution time are some of the important factors impacting the results of the classification models [5].

Studies have shown that prediction from a compound model provides enhanced results as compared to single model prediction. The research in the field of ensemble methods has became popular from the last decade. Ensemble modeling combines the set of classifiers to create a single composite model which gives better accuracy[6].

An ensemble contains base learners and base learners are usually created from training data. These learning algorithm can be K-nearest neighbors(KNN), decision tree, neural network or other kinds of learning algorithms[7]. Two types of ensembles are used. In homogeneous ensemble method, members having single base learning algorithm is taken, Bagging and boosting are the examples for homogeneous ensemble. In heterogeneous ensemble method, different base learning algorithms such as KNN, Support Vector Machine (SVM), Decision Trees and others are taken. Stacking is an example for heterogeneous ensemble,

The purpose of this research study is to predict the engineering students performance in the later semesters using ensemble stacking method by considering 1st to 5th semester academic performance. In first phase, the prediction model was built using traditional classifiers namely, SVM, Logistic Regression and Naïve Bayes. In second Phase, stacking technique was implemented and finally, the usefulness of ensemble method over the individual classifiers was evaluated.This technique involves two stages, the first one generates a group of base level classifiers and the second stage generates a meta-level classifier that associates the outputs of all the base level classifiers considered.

Assessment:

Student's academic performance is commonly measured by conducting examinations or continuous assessment.

1. Summative Assessment:

Summative Assessment is very important in the education process. Summative assessments are used to measure what students have learnt at the end of a course, by conducting mid-term exam, unit tests, final exam, to promote students and to provide certification for completion, so that student can enter certain occupations, or to pursue further education [8].

2. Formative Assessment:

Formative assessment occurs frequently that is short term. Instructors use ongoing feedback method in order to improve their teaching methodology and to improve students' learning capability, portfolios, group projects, quiz, homework methods can be used [9].

.

Methodology:

In this study, demographic and academic data is gathered from undergraduate engineering students studying in different regions of Karnataka. Orange tool was used for processing the dataset and for prediction. The framework that has been used in building a predictive model is shown in the Figure 1.

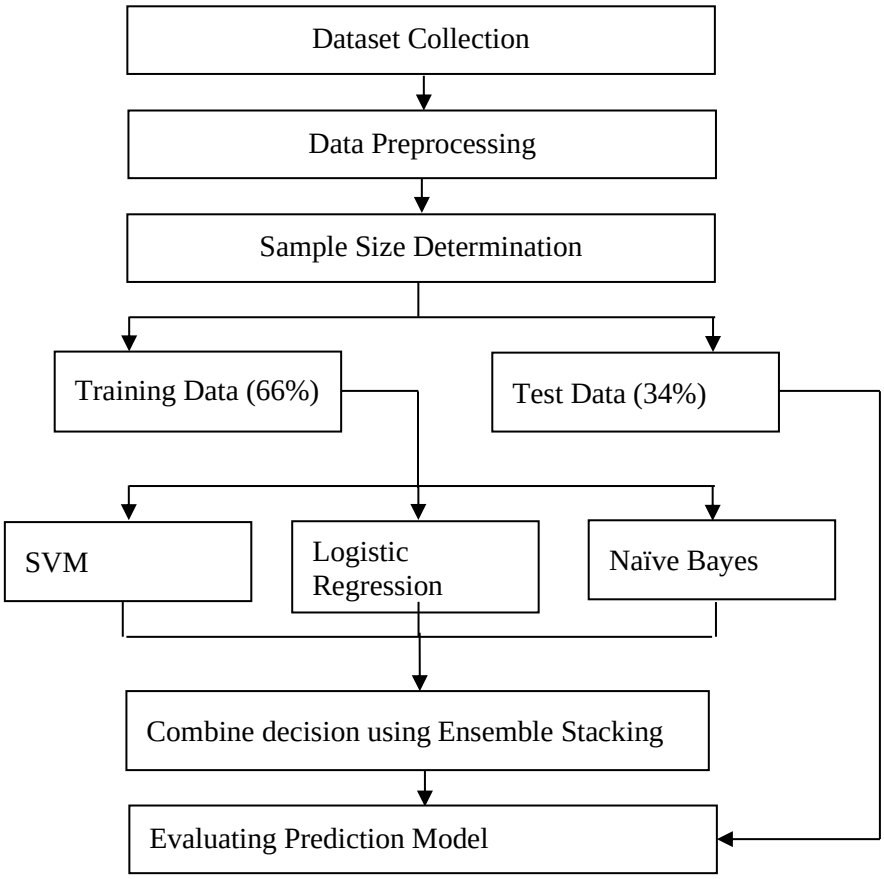

Figure1. Illustration of the framework

A. Data Collection:

Original dataset was collected from students enrolled in Computer Science Engineering, Mechanical Engineering and Civil Engineering and Electronics & Communication Engineering courses in various institutes across Karnataka. The dataset records represent the students' demographics, and information about the students' performance. The dataset consists of a total of 3000 records and attributes such as Student-ID, Student Name, Campus, Gender, GPA, Semester-1 marks, Semester-2 marks, Semester-3 marks, Semester-4 marks and Semester-5 marks as shown in Table-1.:

B. Data Pre-processing

The real-world dataset is prone to miss values, noisy instances, outliers, and so on. Therefore it becomes necessary to clean the dataset which can lead to optimum performance. So, before working on the data and applying the different analytics, the dataset is pre-processed and prepared as follows: Some attributes considered as irrelevant attributes because they only include private information about the students (i.e. Student ID and Student Name). These attributes were excluded and removed as they don't give any knowledge and don't have any importance. Dataset with missing values were removed. Numeric attributes were discretized into categorical as shown in Table-II.

C. Feature Selection:

The predictive model accuracy is mainly affected by attributes used as predictors in the model. Attribute selection (also called feature selection) is important in the process, where irrelevant attributes are identified and removed. In the above dataset wrapper feature selection method was used, the branch, location attribute does not have direct relation with the performance of the student, so it was removed.

TABLE I: FEATURES USED IN THE STUDY

Feature	Data type
College Name	Text
Student's Name	Text
Location	Text-Categorical
Gender	Text-Categorical
Branch	Text-Categorical
Semester Grade Point Average (GPA)	Number continuous
Semester-N* Marks	Integer

* N= 1, 2, 3, 4, 5.

TABLE II: DATA DISCRETIZATION

GPA	9-10	8-9	7-8	6-7	< 6
Normalized Value	Very Good	Good	Above average	Average	Poor

D. Sample Size Determination:

After pre-processing, the entire data of 3000 records was then split into training data and testing data with a ratio of 66:34 respectively.

Result analysis

The results from the three traditional base classifiers namely; SVM, Logistic Regression, Naïve Bayes and Ensemble Stacking using Orange tool are discussed below;

A. SVM

Support Vector Machine algorithm was taken as one of the base classifier and the results of the classification shows the least accuracy of 80%, and F1 score is 79%.

B. Logistic Regression

Logistic Regression algorithm was taken as one of the base classifier and the results of the classification reveal the accuracy as 89%, and F1 score as 89%.

C. Naïve Bayes

Naïve Bayes algorithm was taken as one of the base classifier and the results of the classification are analysed, which reveal that the accuracy is 81%, and F1 score is 81%.

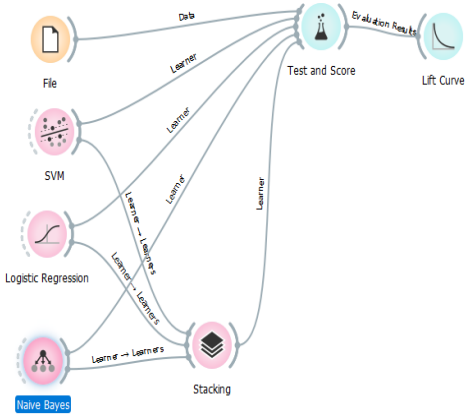

Fig. 2. Data connection for stacking in orange tool

D. Stacking

Ensemble Stacking is used to aggregate all the three classifiers into a single model and the following results were observed; the accuracy is 92%, and F1 score is 92% which was the highest in comparison to all the other considered models.

E. Performance comparison between the applied classifiers

It is observed that,

- Logistic Regression registered highest classification accuracy, when compared between the three base classifier models.
- The classification accuracy is above 80% in all of the three traditional classifier models including ensemble stacking.
- The results of classification under 5 fold cross-validation reveals that the ensemble stacking classifiers performs better in comparison with individual base classifier models i.e., SVM, Logistic Regression, Naïve Bayes as shown in Table-III.

F. Lift Curve

Of the four algorithms tested, Ensemble Stacking outperforms the SVM, Logistic Regression and Naive Bayes classifiers. The Lift curve in Figure 3, shows that by selecting the first 30% of students as ranked by the model, will get three times more positive instances than by selecting a random sample with 30% of students.

TABLE III. EVALUATION RESULTS

Model	AUC	CA	F1	Precision	Recall
Stacking	0.992	0.925	0.925	0.925	0.925
SVM	0.956	0.800	0.799	0.799	0.800
Logistic Regression	0.986	0.891	0.891	0.891	0.891
Naïve Bayes	0.952	0.815	0.815	0.816	0.815

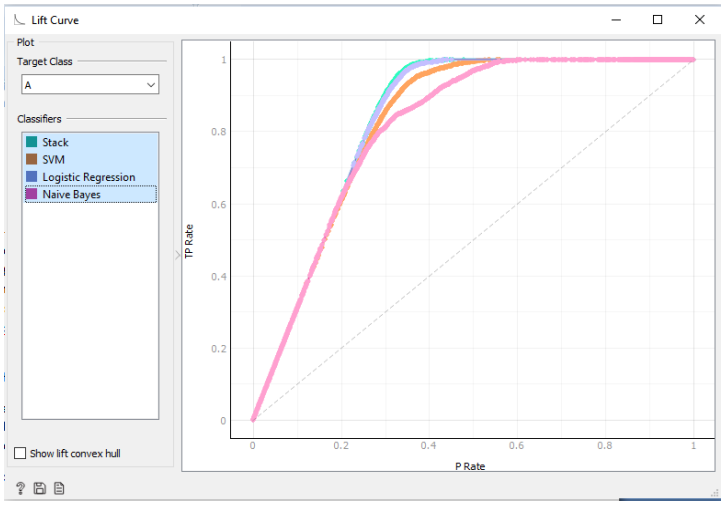

Fig. 3 Lift Curve

Conclusion

The results indicate that the proposed stacking technique achieves high performance, surpassing that of other base classifier techniques. Consequently, it can be concluded that the stacking method outperforms various algorithms.

References:

1. Jothi Lakshmi, S., & Thangaraj, "Recommender System for Student Performance Using EDM", Asian Journal of Computer Science and Technology, ISSN: 2249-0701 Vol.7 No.3, pp. 53-57, 2018.
2. Nagendra, K. V., K.Sreenivas, & P.Radhika. (2018). Student performance prediction using different classification algorithms. International Journal Of Current Engineering And Scientific Research (IJCESR) , Volume-5, Issue-4.
3. Adebayo, A. O., & Chaubey, M. S., "Data Mining Classification Techniques on the Analysis of Student's Performance", GSJ, Volume 7, Issue 4, 2019.
4. SaaMostafa, A. A., Al-Emran, M., & Shaalan, K. "Factors Affecting Students' Performance In Higher Education: A Systematic Review Of Predictive Data Mining Techniques. Technology, Knowledge And Learning", DOI: 10.1007/s10758-019-09408-7, 2019.
5. Mythili, B. A, "A Study on Students Academic Performance Analysis Using Classification and Prediction Using Data Mining Techniques in Arakkonam Higher secondary School", IPASJ International Journal of Computer Science (IIJCS), 6(9), 2018.
6. D.Gopika, & B. Azhagusundari, "An Analysis on Ensemble Methods in Classification Tasks". International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, 2014.
7. Singh, R., & Pal, S. "Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance", International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, No. 3, 2020.
8. Deshmukh, V., "Student Performance Evaluation Using Data Mining Techniques for Engineering Education", Advances in Science, Technology and Engineering Systems Journal , Vol. 3, No. 6, 259-264, 2018.
9. Gogri, M. H., Shaikh, S. A., & Iyengar, V. V., "Evaluation of Students Performance based on Formative Assessment using Data Mining", International Journal of Computer Applications (0975 – 8887) , Volume 67– No.2, 2013.