A ADVANCED JOURNAL IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING TRENDS IN BIOLOGICAL SCIENCES
Open Journal

Review

Using Ensemble Technique and Machine Learning Algorithms to Enhance Student Performance Prediction

Randhir Singh[1*]; Saurabh Pal[2]

[1]*Research Scholar, Department of Computer Applications, VBS Purvanchal University, Jaunpur, India*
[2]*Head, Dept. of Computer Applications, VBS Purvanchal University, Jaunpur, India*

[*]***Corresponding author***
Randhir Singh
Research Scholar, Department of Computer Applications, VBS Purvanchal University, Jaunpur, India

Article information
Received: October 30th, 2023; **Revised:** November 30th, 2023; **Accepted:** December 11th, 2023; **Published:** January 4th, 2024

Cite this article
Singh R, Pal S. Using ensemble technique and machine learning algorithms to enhance student performance prediction. 2024; 3(1).
doi: https://doi.org/10.70705/ppp.bioai.2024.v03.i01.pp1-4

ABSTRACT

Measuring The moral community relies on the students' work. Many areas have benefited from the expanded use of machine learning algorithms, including the prediction of diseases, student performance, agricultural output, and many more. The overarching goal of this research is to find ways to enhance the accuracy of student performance prediction by combining several machine learning algorithms into an ensemble method. We have used the student dataset, which includes 1000 occurrences and 22 characteristics, to assess students' performance. We used four machine learning algorithms—Decision Tree (DT), Naïve Bayesian (NB), K-Nearest Neighbors (KNN), and Extra Tree (ET)—in this study. Then, we created a model that uses Bagging and Boosting ensemble techniques to aggregate the findings of each base learner. To determine the most effective model, we examined the outcomes produced by the bagging and boosting ensemble methods. Several metrics, including sensitivity, specificity, accuracy, and f1-score, are used to evaluate the outcomes of all machine learning algorithms and ensemble approaches. When we examine the outcomes of bagging and bagging ensemble approaches, we discover that bagging produces the best results. Both the institution and the admissions officers may benefit from using the created model to determine which courses students are most likely to fail in.

Keywords
Educational data mining; Machine learning; K-Nearest neighbor classifier; Extra tree classifier; Ensemble technique.

INTRODUCTION

Clustering problems are a kind of academic performance assessment, where clusters represent students' grades (e.g., pass, second-class, third-class, and fail) and intelligence levels are measured. In order to choose the most worthy pupils and offer them a solid education that will help them reach their full potential, intelligence-based groups are crucial.

Life's purpose. Institutions of higher learning are increasingly making use of information technology tools to amass massive volumes of student data as the price of electronic equipment continues to fall. Currently, most affiliated universities gather information about students' registration and exams through various forms; these institutions also provide students with electronic mark sheets, certificates, and migration aid, but unfortunately, this data is often not used for anything beyond its initial collection. With the use of data mining methods, machine learning models may be built from these datasets. Educational Data Mining (EDM) is the process of applying data mining methods to these datasets. To better comprehend students' actions and capacities for learning, EDM incorporates research into novel instruments, models, and algorithms for processing massive amounts of data in search of actionable, previously unseen patterns. Accordingly, educational data mining allows for the discovery of novel solutions for the resolution of issues pertaining to educational domains.

Modern universities make use of ICT (Information and Communication Technology) assets, such as Management Information Systems, to collect and analyze massive amounts of data [1]. Educational data may be analyzed using perfect machine learning algorithms. In this research, we create a novel model for predicting students' academic success using machine learning classifiers.

Finding out what causes students at United College of Management in Prayagraj, UP, to have poor academic performance in their Bachelor of Computer Applications (BCA) programs is the main goal of this research. In order to improve the quality of the institute and

the performance of its students, the BCA department may use the established model to make informed choices about which students to focus on and how to best help them.

Here is the remaining paper. Section 2 follows with a description of prior work in the EDM sector that made use of machine learning methods. Details on the dataset, machine learning methods, ensemble strategies used to improve the performance of individual classifiers. The findings and comments are presented in Section 4. We reviewed the study's results and conclusions in portion 5, the last portion.

2. RELATED WORK

We reviewed the literature that was pertinent to the topic. With numerous variables retrieved from the university database, Ahamad and Elaraby [2] analyzed the six-year enrollment data of students in a particular academic program throughout (2005–2010). Final grades in the program were forecasted by the work.
Using the Naïve Bayes algorithm, Pandey and Pal [3] introduced a method for educational data mining that sorts students into two groups: those who do well and those who do not.

In a related work, Bhardwaj and Pal [4] examined several decision tree algorithms using a dataset of academic achievement in attempt to forecast how well students will do in school. Selecting an optimal decision tree algorithm and then providing a benchmark for each of them is the primary focus of the study. After evaluating the dataset using the accuracy and precision created during validation, it was found that the CART decision tree approach performed relatively better.

For the purpose of predicting students' mathematical ability, Livieris et al. [5] created an ANN classifier. They found that, when compared to other classifiers, the modified spectral Perry trained artificial neural network performed the best in classification.

In order to forecast which distance learning students will drop out, Kotsiantis et al. [6] investigated machine learning methods. One of the first efforts to use machine learning techniques in a scholarly setting, this project paved the way for educational data mining. Rather than using statistics on students' performance in class, their algorithm was given information on their demographics and a number of projects.

Using a combination of K-Means Clustering and Artificial Neural Network, Moucary et al. [7] developed a method to help kids learn and communicate in a new language while they are in school. The students' performance was first predicted using a Neural Network, and then they were fitted into a specific cluster that had been formed using the K-Means technique. When students were in the early phases of their academic careers, this clustering was a significant tool for teachers to use in identifying their skills.

In order to predict traffic flow, Hongsuk et al. [8] construct a model using a Deep Neural Network. When traffic was light and when it was heavy, a Traffic Performance Logistic regression was used. With a 99% degree of accuracy, the three-layer mold could predict the congestion.

The performance perdition of seventy-six students from Zaprešić's University of Applied Sciences Baltazar is examined by Zoric and Alisha [9]. Due to a lack of data from lower-grade kids (only 8%), the author was able to attain a prediction rate of 93.42% using a Neural Network.

In order to forecast how well students will do, Hassan et al. [10] utilized data from 1,170 pupils. Among the classifiers they tested—Decision Tree, SVC, KNN, Gradient Boosting, Random Forest, and Linear Discriminant Analysis—the greatest accuracy was 89.74%. Gamao and Gerardo [11] combined the MMFA with appropriate classification algorithms like NB and DT to create a model that can predict whether students would drop out based on their cumulative record. By modeling its behavior after that of fireflies and using mutation to find the best possible answer or model, the MMFA was able to traverse the search space. Consequently, academic administrators at Davao del Norte State College may use the study's accuracy to their advantage when formulating academic policies aimed at reducing student dropout rates by improving students' performance in the classroom and beyond. Researchers in the future may want to look at combining the MMFA with other applicable classification algorithms to see what it can do. Schools and teachers have a significant and difficult challenge in identifying the reasons why students drop out, as anticipated by Gil et. al. [12]. Consequently, they investigated if data mining procedures may help with this problem in all schools. The data mining classification approach was able to correctly forecast the student dropout indicators. In order to determine the signs of student dropout, the most popular data mining methods, based on C4.5 and Naive Bayes, were used. A ten-fold cross-validation method was used to train and evaluate these two distinct classification algorithms. Educators are notified to take the necessary steps to enhance student performance via personalized counseling and coaching. Various machine learning approaches for student performance evaluation were detailed by Singh and Pal [13]. We apply five ML techniques—PCA, SVM, LDA, RNC, and ET—to categorize students' predictions. With an accuracy of 94.86%, SVM outperforms all of the other methods. Among the methods tested, LDA had the second-highest accuracy at 93.21%. Among the existing studies on student performance prediction, they achieved the greatest accuracy. By treating the first-stage prediction as a feature instead of a distinct variable, the machine learning-based approach decreases generation mistakes and acquires more information.

3. METHODS

In this paper we have used three machine learning classifier algorithms, namely: Decision Tree (DT), Naïve Bayesian (NB), K-Nearest Neighbors (KNN) and Extra Tree (ET) for the purpose of testing our dataset which was taken from United Institute of Management, Prayagraj. A brief description of the classifiers used in this study is described below.

- Decision Tree (DT):

Decision tree is a type of supervised learning based predictive modeling tool. Decision tree is based on graphical representation of all possible solution on different conditions. A decision tree is generated from root following top-down approach that involves partitioning of data; entropy is used to calculate homogeneity of data. Category based data as well as numerical data both work with this model.

- Naïve Bayesian (NB):

Naive Bayes is a popular data classification technique. It is based on the probability theory concept and based on assumption that there is no dependency among predictors. In other words it assumed that the presence of a particular feature in a class is not related to each other.

- K-Nearest Neighbors (KNN):

K nearest neighbor algorithm is a classifier which creates different types of cases that is based on similarity measure It is a supervised machine learning algorithm applied for classification and regression problem. It is non-parametric because it does not have any assumptions about the distribution of data. In classification algorithm, learning depends on 'how similar' a data is from the other.

- Extra Tree (ET):

This method is an ensemble method which stands for Extremely Randomized Trees. This algorithm develops randomizing of tree for numeric input features. It often leads to increased accuracy when compared to the ordinary random forest.

3.1 Dataset Analysis

The data used in this study is of Bachelor of Computer Applications program, which has been collected from United Institute of Management, Prayagraj. The BCA course is divided in 3 years which consist of two semesters 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 the registration as well as examination form. The target and other variables discussed in this study are listed in table.

3.2 Data Preprocessing

The first step shown in methodology is data preprocessing. Data preprocessing includes (i) a method to select students' records and choosing important attributes and (ii) the students records are not clean and include inconsistent data. So apply different methods of data cleaning to clean such anomalies. The dataset for the study was collected from the United Institute of Management, PrayagRaj. The dataset had 1000 instances and 22 attributes. The student's dataset is pre-processed using equation

Ensemble Techniques

Ensembles techniques improve the accuracy of predication as compared to single classifier on the dataset. In this paper, we have applied two ensemble methods to improve the performance of classification algorithms. The two most popular ensemble techniques: Bagging classifier and Boosting (Adaboost classifier) are used to combine the results obtained by the four machine learning classifiers.

Bagging Classifier: Bagging method is applied to decrease the variance calculated by decision tree classifier. The objective of bagging ensemble method is dividing the dataset into various subsets for training selected at random with substitution. Now, these data subsets are trained using decision trees. Now, the average of the results obtained by each data subset is taken which gives better results as compared to single classifier.

Boosting Classifier: Boosting classifier is another important ensemble technique. It is applied to create a group of classifiers. In boosting method, classifiers are trained serially by classifiers fitting data and then analyzing errors. Decision trees are tainted successive to fit from the data and with the objective to get improved accuracy at each stage. Bagging method is used to converts weak classifiers to a good model.

RESULTS

Before applying machine learning classifiers the dataset is visualized using histogram and density map. The histogram and density maps of all attributes of student dataset represent the bar of frequency of different values and density map is a smooth continuous curve, which is formed by estimating the density from the data individually. In a density map, continuous curve is drawn at each individual data point and then all these curves are summed up to get a single smooth curve accordingly. We measured the density of each attributes on the basis of target variables classification and represent in Figure 2.

The analysis of dataset and implementation of classification in this paper has been done using Python code. The student dataset is divided into 80% as training set and 20% as test dataset using 10-fold cross validation.

To measure the performance of the predicted model accuracy, Recall, specificity, precision and F-1 scores are evaluated using the formula shown in table 2.

On comparing the results, we finds that the best accuracy achieved by Naïve Bayesian classifier as 86.83%, from the table it is clear that nearly all classifiers predict the accuracy in between 81% to 87%, which proves the selection of these classifiers are best for predicting the performance of students. K-nearest Neighbor classifiers also performs well as its accuracy achieved is 84.72%. The comparison of the four classifiers on the basis of accuracy is presented in figure 3 using box and whisker plot to illustrate the mean value of prediction.

To improve the results of the machine learning classifiers, we have

used Bagging and Boosting ensemble techniques. Ensemble techniques are applied to improve the accuracy of machine learning classifiers. After applying the Bagging and Boosting the results obtained by the two methods are shown in table 4.

From the table it is clear that Boosting is the better ensemble technique as compared with bagging because the accuracy of boosting classifier is 91.76% which is also approximately 5% more than the single machine learning classifier Naïve Bayesian. The comparisons of two ensemble techniques are shown in Figure 4.

4. CONCLUSION

The primary goal of this research is to develop a model that can enhance the accuracy of student performance predictions. Predicting how well a pupil will do in school now often makes use of machine learning and ensemble methodologies. To enhance the performance of individual machine learning classifiers, ensemble approaches are used. Basis learning methods used in this study are Decision Tree, Naïve Bayesian, K-nearest Neighbor, and Extra Tree, which are machine learning classifiers. Two ensemble methodologies are then employed. Improving the performance of single-base learners is possible with the use of bagging and boosting. The Naïve Bayesian classifier achieves the greatest accuracy at 86.83%, while the boosting ensemble approach achieves the best accuracy at 91.76%.

Finding the pupils who aren't doing well and focusing your efforts on helping them may be done using the data presented in this study. Because of this, the standard of higher education and might be good for universities.

REFERENCES

1. Learning Analytics in 2012: A Review and Future Challenges (Ferguson, R.). This is a technical report from the Knowledge Media Institute published in 2012. You can get it online at this URL: http://kmi. open. ac. uk/publications/techreport/kmi-12-01.
Two authors, Ahmed and Elaraby. Information mining: a categorization-based performance prediction based on student data. Articles 43–47 published in the 2014 volume 2, issue 2 of the World Journal of Computer Application and Technology.
2. Panday and Pal; S. Data mining: a classification-based predictor of high- or low-performing employees. 2nd International Journal of Computer Science and Information Technology, Issue 2, pages 686–690, 2011.

4. Data mining: A forecast for performance enhancement via categorization written by Bhardwaj B and Pal S. Volume 9, Issue 4, pages 136–140, 2012, International Journal of Computer Science and Information Security (IJCSIS).

5. A training approach for spectral conjugate gradient neural networks, by Livieris and Pintela. The 2012 volume 21, issue 1, pages 1250009-1-21 of the International Journal on Artificial Intelligence Tools. URL: https://doi.org/10.1142/S0218213011004757

6. Using machine learning approaches to predict students' success in distant learning (Kotsiantis S, Pierrakeas C, Pintelas P.). Volume 18, Issue 5, pages 411-426, Journal of Applied Artificial Intelligence, 2004.

Using data clustering and neural networks to improve student performance in foreign-language oriented higher education (Moucary C, Khair M, Zakhem W., 2007). The ACM Research Bulletin of Jordan, Volume 2, Issue 3, pages 3–10. 2011, pages 27–34.

8. A study conducted by Hongsuk Y, Jung H, and Bae S. used deep neural networks to forecast traffic flow. The 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), with proceedings spanning pages 328–331.

This may be accessed at this URL: https://doi.org/10.1109/BIGCOMP.2017.7881687.

Bilal Zorić worked on this in 2019. Students' Academic Performance Forecasting using Neural Networks. 2019. Zagreb: RENET—Society for Advancing Innovation and Research in the Economy, pp. 58-66.

10. The authors of this work are Hasan, Hasan R., Rabby, Islam, and Hossain.

A. S. (July 2019). Learning Algorithm for Predicting Students' Academic Performance. Included in the proceedings of the ICCCNT 10th International Conference on Computing, Communication, and Networking Technologies, pages 1–7. 2019, IEEE.

11. A Model for Student Dropout Prediction Using the Modified Mutated Firefly Algorithm (Gamao1, A.O. & Gerardo, B.B.). Publication date: 2019 in the International Journal of Advanced Trends in Computer Science and Engineering, volume 8, issue 6, pages 3461–3469. The link to the article is https://doi.org/10.30534/ijatcse/2019/122862019.

12. Utilizing Data Mining Techniques to Forecast Public School Dropout Indicators (Gil, J.S., Delima, A. J. P., & Vilchez, R. N.). Advances in Computer Science and Engineering: An International Journal, Volume 9, Issue 1, Pages 774–778 (2020).

The link to the article is https://doi.org/10.30534/ijatcse/2020/110912020.

13. Using Machine Learning Algorithms to Forecast Students' Academic Success (Singh, R. & Pal, S.). Paper published in 2020 in the International Journal of Advanced Science and Technology, volume 29, issue 5, pages 7249–7261.