Ensemble-Based Machine Learning Strategies for Predicting Student Dropout and Academic Outcomes.

Chironto Rudra Paul, Arnob Das

Department of Computer Science and Engineering

North East University Bangladesh

Email: chirontorudrapaul@gmail.com, arnobdas789@gmail.com

Abstract -

The issue of student attrition is serious to higher learning institutions in the sense that it lowers the academic performance and sustainability. Students who are likely to discontinue their studies at an early stage can be identified and therefore, institutions can establish intervention mechanisms in good time. The purpose of this paper is to present the ensemble-based machine learning model as a predictor of the student dropout and success in the academic performance based on a real-life dataset of higher education. The source used in data collection has been a combination of the different independent databases and in this case, the data collected relates to the entire undergraduate students who are undertaking the degree programs which involve agronomy, design, education, nursing, journalism, management, social services, and technology related courses. The data collected has enrolment-time variables, e.g. demographic, academic, and socio-economic variables together with academic performance variables, which will be measured after the first and second semesters. Among the supervised learning algorithms that were used and tested, there were the Logistic Regression, Decision tree, Support Vector machine, K-Nearest neighbors, random forests, and Stacking Classifier. The performance of Stacking Classifier creates an experimental outcome that shows the best with a result of 94.50. The results confirm the value of the ensemble learning approach and emphasize the importance of the data-driven decision support system in the academic environment.

Index Terms- Student Dropout, Academic outcome prediction, machine learning, ensemble model, and Educational Data Analytics.

I. INTRODUCTION

Student dropout is one of the most severe problems that a higher education facility needs to work with in the global context. The high dropout rates are not only degrading the academic performance of students, but also costing universities a huge sum of money and image. The conventional methods of following a dropout risk are largely reactive in nature and are not as scalable and effective as they are manual in nature. The developments in the accessibility of data and the computing processes have made machine learning in education analysis easier. Educational data mining, on its part, assists in the emergence of the hidden patterns of institutional data, and it is in good position to forecast successes or failures of students. To be able to be proactive and evidence-based in decision-making within the higher education system, the study will use both the machine learning and the ensemble modeling in the prediction of student dropout and academic performance.

II. DATASET DESCRIPTION

The data collected in this research was obtained in one of the institutions of higher learning and was assembled with records retrieved in different fragmented databases. It involves data of students who have undertaken other degree programs such as agronomy, design, education, nursing, journalism, management, social service and technology oriented courses. One can roughly divide the set of data into the following parts:

A. The enrollment-time attributes

Under this category, the academic path or demographic details and social economic status of the students that are available in the enrolment period are covered.

B. Educational Achievement Characteristics.

This group will incorporate measures pegged on the achievements of the students based on their academic grounds at the conclusion of the first and second semesters. The target attribute will be the final destination of students in their degree programs at the required time in which the degrees are supposed to be attained. The prediction problem is formulated to form a three-class classification problem that comprises of Dropout, Enrolled and Graduate classes. The said classes were also summarizable as binary in measuring the risk of the dropout to academic success, and other purposes of measuring the analysis. Zenodo publishes the metadata of the datasets [1].

III. METHODOLOGY

The proposed architecture relies on a machine learning workflow. To investigate the distribution of features, point to the disproportion of the classes and some potential correlation of the variables, the exploratory data analysis was conducted initially. Some of the data preprocessing processes included the handling of missing values, encoding of categorical features as well as the normalization of numerical values. The data was further split into training and testing set with the aim of ensuring that no biased test was taken. Individual supervised learning models were trained separately and then an ensemble Stacking Classifier was prepared to make collective prediction using a number of base learners. The measures of model effectiveness were assessed based on the accuracy and standard classification performance measures.

IV. MACHINE LEARNING MODELS

The learning procedures used in this study and overseen comprise the following: Logistic Regression Decision Tree Classifier Support Vector Machine (SVM) K-Nearest Neighbors (KNN) Random Forest Classifier Stacking (Ensemble Approach) Classifier. The Stacking Classifier is a mix of predictive outputs of two or more base models and it is this that enables it to be of higher generalization capacity and displays a reduced overfitting.

V. RESULTS AND ANALYSIS

Experiment study demonstrates that the use of ensemble-based learning increases predictive performance significantly. Of all the models that were tested, the Stacking Classifier produced the highest accuracy of 94.50. The future of the class distribution is estimated to be that approximately 63.1 percent of the students will show that there would be a dropout-risk group, and 36.9 percent would show that he/she would attain academic success. These results highlight the fact that the academic performance measures in the semester are valid measures of forecasting student performances. Besides, there are other demographic and socio-economic reasons which also bring certain value to the overall effectiveness of prediction of the model.

VI. DISCUSSION

The findings of the study are related to those of the earlier studies in the field of educational data mining since they dwell on the significance of early academic performance and background issues in the determination of student success. The ensemble modeling method is an effective methodology of modeling intricate interactions between the variables in the data, thus it is suitable to real world institutional setting. The accuracy that is obtained shows that such predictive models can be a dependable tool used by academic administrators in making decisions.

VII. CONCLUSION

The presented paper shows that machine learning predictors of ensemble may be useful to forecast the student dropout and academic performance on the basis of the real-life higher education data. The suggested Stacking Classifier is quite powerful in prediction because it is characterized by an accuracy of 94.50%. The findings support the data-oriented analytics to be included in the decision-making engines of the institutions to enhance the retention of the students. In the future, one can speak about behavioral indicators, longitudinal data, and deep learning structures as the potential ways of improving the accuracy of prediction in the workplace.

REFERENCES

[1] Zenodo Dataset, “Predict students’ dropout and academic success.” https://zenodo.org/records/5777340

[2] [6] C. R. Paul, “Student Dropout Prediction and Academic Success,” Kaggle, 2025.

[Online]. Available: https://www.kaggle.com/code/heyboss/student-dropout-prediction-academic-success

[3] V. Tinto, “Dropout from higher education: A theoretical synthesis of recent research,” *Review of Educational Research*, 1975.

https://doi.org/10.3102/00346543045001089

[4] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics*, 2010.

https://ieeexplore.ieee.org/document/5432695