

Prediction of Student Dropout Risk Using Multi Factor ML Combining Attendance, Grade and Engagement

Santhanalakshmi S

and Senthilkumar G

Department of Computer Applications, Periyar Maniammai Institute of Science and Technology, Thanjavur Vallam-613403, Tamil Nadu, India, gsrandhra@gmail.com and Assistant Professor -SG/CA, Periyar Maniammai Institute of Science & Technology, Tamilnadu, India, senthil@pmu.edu

Abstract.

The problem of student dropout is one of the most significant problems that affect the performance and effectiveness of organizations. In the current research paper, three machine learning ensemble models are considered for prediction of student dropout depending on discussions, semester, number of absences, and the survey of the parents. For this purpose, the accuracy, precision, recall, F1-score, and confusion matrix were used for comparing these machine learning models. According to the results of experiments, the random forest model performed better than XGBoost and CatBoost models.

Keywords. Student Dropout Prediction, Machine Learning, Random Forest, XGBoost, CatBoost, Educational Data Mining

1. INTRODUCTION

Machine learning optimization for Moodle logs data enhances the precision of predicting student dropouts and makes early interventions possible [1]. Effective machine learning models can be used to predict student dropouts in the field of e-learning, thereby helping in early detection of at-risk students [2]. Machine learning techniques provide accurate predictions about student dropout using information about their academic behavior [3]. Machine learning is an efficient method for predicting student performance [4].

2. LITERATURE REVIEW

Cannistrà et al. (2021) developed methods based on multilevel machine learning and statistical models to predict university student dropout and show increased ability to identify students at risk [5]. Mduma (2023) used data balancing algorithms in order to increase machine learning models' accuracy in predicting student dropout in imbalanced dataset environments [6]. Vaarma and Li (2024) examined machine learning models used to predict student dropout in the case of Finnish higher education institutions and their effectiveness in detecting risks in advance [7]. Niyogisubizo et al. (2022) presented a two-level stacked ensemble machine learning method that greatly increased the accuracy of university student dropout prediction [8].

3. METHODOLOGY

Parent Responding to Survey: The selected features were selected according to their significance in the field of student retention research, while the Class feature refers to the dropping out of students. **Data Preprocessing:** Consistency checking, encoding of the categorical features into numbers (Discussion and Semester), and dividing of data into 80% training data and 20% testing data in accordance with class proportion was carried out. There was no need for feature scaling because tree ensembles are insensitive to the feature magnitude. The target variable (Class) is the dropout risk category allocated to each individual. In this research, Class 0 is the dropout risk category of students having low dropout risks, Class 1 is the dropout risk category of students having moderate dropout risks, and Class 2 is the dropout risk category of students having high dropout risks. This classification of classes is used uniformly for training, testing, and evaluation purposes, allowing for better understanding of the confusion matrix and performance of each class. **Development of Models:** The models have been built using classifiers namely, Random Forest, XGBoost and CatBoost with the same default parameter values. Random Forest classifier has been chosen for its stability and to model non-linear relationship, XGBoost has been used for its regularization properties and good classifying power, and CatBoost because of its fast processing of categorical data with ordered boosting.

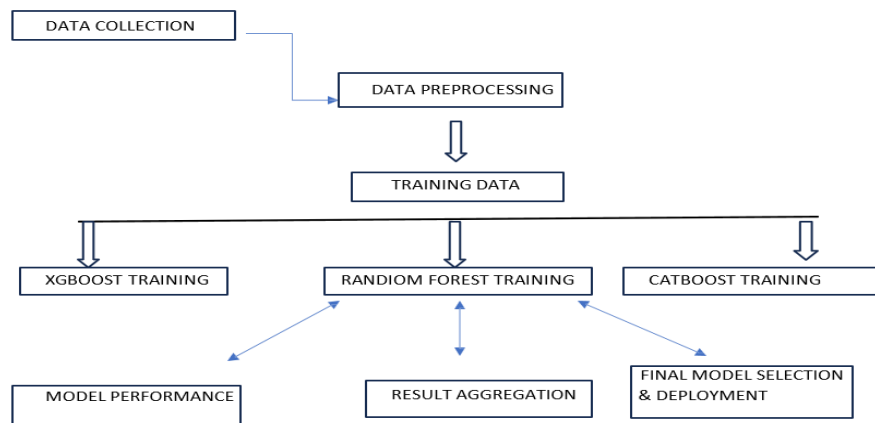


Fig. 1 illustrates the complete workflow from data collection and preprocessing to model training, performance evaluation, result aggregation, and final model selection.

After training, all the models were tested on the test set. The performance metrics considered in this paper are accuracy, precision, recall, F1-score, and confusion matrices. These metrics are chosen because they can comprehensively explain the effectiveness of the models, especially for multi-class classification problems. Accuracy gave the overall success rate, while precision and recall showed class- dependent performance. The F1-score represented a balance between these two measures; hence, it was informative for classes with uneven distributions. Confusion matrices were analyzed for visualizing model behavior in distinguishing among three classes of dropouts. Finally, feature importance

analysis was done using the Random Forest model, which shows the respective contributions of the selected features toward the task of prediction. analysis also helps in underlining those behavioral and academic attributes most strongly influence the likelihood of dropout. Visualizations from the analysis were created, together with model performance graphs, for inclusion in the results section to further help interpret the findings.

4. RESULTS

Test dataset analysis was performed using Random Forest, XGBoost, and CatBoost models. Among all, Random Forest model gave the highest accuracy in terms of F1 score. CatBoost gave results similar to Random Forest, whereas XGBoost gave lower results than other two. Although Random Forest correctly classified most data points, but recall value for class 0 was low, implying that there was higher misclassification rate for Class 0 than Class 1 and Class 2. Figure 4.1 provides the confusion matrices for the model, highlighting such behavior. However, Random Forest was still the most stable and performant model, with a top F1-score.

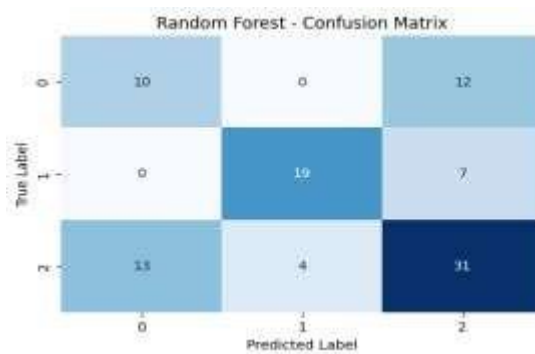


Figure 4.1: Confusion Matrix for Random Forest

XGBoost yielded a mediocre performance; it performed reasonably well on Class 1, but failed to distinguish Class 2 from the other categories. This was one of the reasons for the lower overall performance in terms of overall accuracy and F1-score. Figure 4.2 demonstrates misclassification patterns, particularly the overlap involving Class 2.

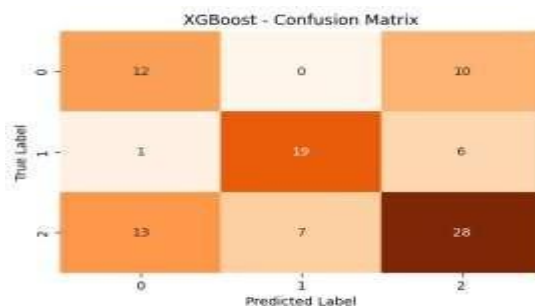


Figure 4.2: Confusion Matrix for XGBoost

CatBoost achieved a performance similar to that of Random Forest. As the ordered boosting approach was used by CatBoost, it made dealing with categorical features very easy, resulting in high recall on Class 1.

A final cross-model comparison showed both Random Forest and CatBoost achieved the same accuracy, but Random Forest still kept a slight advantage in F1-score, which allowed it to be regarded as the most reliable model in general. XGBoost consistently scored lower than both other models. The performance of the algorithms Random Forest, XGBoost, and CatBoost when compared based on the evaluation metrics is illustrated in Figure 4.4.

Table 1: Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.6250	0.6333	0.6250	0.6282
XGBoost	0.6146	0.6219	0.6146	0.6168
CatBoost	0.6250	0.6323	0.6250	0.6273

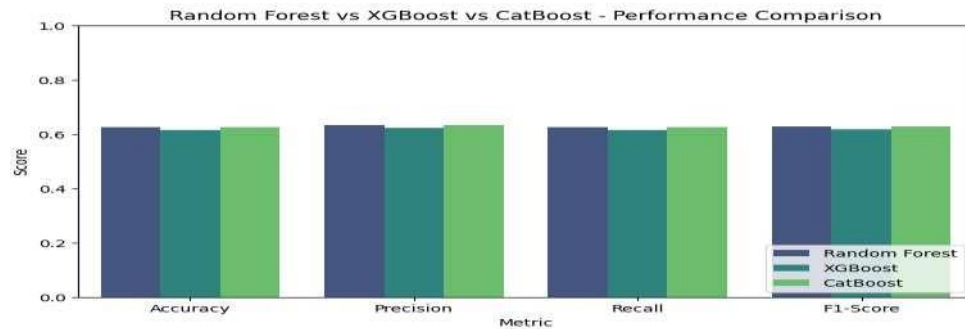


Figure 4.4: Model Performance Comparison Graph

5. DISCUSSION

The performance comparison between Random Forest, XGBoost, and CatBoost on the test data revealed significant variations among the models. The Random Forest outperformed the other two models by demonstrating the highest performance level in terms of F1 score. Despite its poor recall in Class 0, the model was able to classify Classes 1 and 2 without errors making it more stable (Fig. 4.1). XGBoost was able to demonstrate acceptable performance by correctly identifying Class 1 but had problems with Class 2 recognition. Therefore, the precision and F1 scores were low (Fig. 4.2). CatBoost had a good performance comparable to Random Forest by handling the categorical variables and high recall for Class 1. Even though both Random Forest and CatBoost had the same accuracy, the former showed better F1 score. While Random Forest had the best overall performance amongst the considered algorithms, more validation is needed to ensure that the results are generalizable. Future studies need to focus on using larger datasets with varying

characteristics gathered from several educational institutions and environments. Hyperparameter optimization could be done to improve the performance of Random Forest, XGBoost, and CatBoost algorithms. Newer deep learning techniques such as LSTM and GRU could be considered for the study since they take care of temporal trends associated with student absenteeism and engagement. Explainable Artificial Intelligence (XAI) methods could also be introduced to make interpretations of the factors involved in predicting individual dropout risks.

6. CONCLUSION

This research evaluated the performances of Random Forest, XGBoost, and CatBoost in predicting student dropouts with respect to both behavior and academic attributes. Among the models, Random Forest performed well on all measures of evaluation, thanks to its capability of modeling non-linear relationships and preventing overfitting. In particular, predictors such as student absence days, discussions, semester, and parent's interaction became prominent, showing the significance of engagement-related attributes. The findings suggest the adoption of Random Forest in developing early warning systems for detecting at-risk students and increasing their retention rates. However, due to several limitations including feature selection, hyperparameter tuning, and class imbalance, this research opens up possibilities for further investigation.

7. FUTURE WORK

Some possible directions for future research in the area of the proposed research may include application of the models to different and bigger datasets from other educational establishments. Different hyperparameters might be tuned to improve the performance of the ensemble machine learning models. Moreover, new methods of deep learning, such as LSTM, GRU and Transformer, might be evaluated in comparison with the suggested machine learning models in order to understand the ability of the models to detect patterns in students' behavior. The use of techniques of explainable artificial intelligence, such as SHAP and LIME, will increase the explainability of the dropout risk predictions. Furthermore, implementation of the optimized model into the early warning dashboard is recommended.

8. REFERENCES

- [1] M. Rebelo Marcolino et al., "Student dropout prediction through machine learning optimization: Insights from Moodle log data," *Scientific Reports*, vol. 15, no. 1, 2025.
- [2] M. Tan and P. Shao, "Prediction of student dropout in e-learning program through the use of machine learning method," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 10, no. 1, p. 11, 2015.
- [3] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, 2020.
- [4] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

- [5] M. Cannistrà et al., “Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques,” *Studies in Higher Education*, vol. 47, no. 9, pp. 1935–1956, 2021.
- [6] N. Mduma, “Data balancing techniques for predicting student dropout using machine learning,” *Data*, vol. 8, no. 3, p. 49, 2023.
- [7] M. Vaarma and H. Li, “Predicting student dropouts with machine learning: An empirical study in Finnish higher education,” *Technology in Society*, vol. 76, p. 102474, 2024.
- [8] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,” *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

Biographies



Santhanalakshmi currently pursuing the bachelor's degree in Computer Applications in Periyar Maniammai Institute of Science and Technology.