Optimizing Multi-Class Classification in Educational Data with Ensemble Learning and Data Balancing Techniques

Mohammad F. Al-hammouri*, Ziad Akram Ali Hammouri†, Islam T. Almalkawi*, Ansam Lafee*
**Dept. of Computer Engineering*, *The Hashemite University*, Zarqa, Jordan
Email: {alhammouri, eslam.malkawi}@hu.edu.jo, lafiansam899@gmail.com
†*Dept. of Computer Science*, *Middle East University*, Amman, Jordan
Email: z.hammouri@meu.edu.jo

***Abstract*—The field of education is increasingly embracing AI tools to improve student outcomes. This work aims to reduce academic failure in higher education by employing machine learning techniques to identify at-risk students early in their educational journey, enabling the implementation of supportive strategies to assist them. This study examines a dataset from a higher education institution and utilizes it to develop a classification model for predicting students' academic performance. The problem is formulated as a multi-class classification task with three categories: Graduate, Enrolled, and Dropout, with a significant imbalance skewed toward the Graduate. To improve prediction accuracy toward the minority class, the data balancing technique SMOTE with Edited Nearest Neighbor (SMOTE-ENN) is applied. Three popular classification models—Random Forest, XGBOOST and CatBoost—are employed. The findings show that SMOTE-ENN significantly improves classification results. Moreover, XGBOOST demonstrated the highest accuracy (94.6%) in correctly identifying all classes, as evidenced by the confusion matrix evaluation, achieving the highest results compared to previous work in the literature. Implementing these models allows for accurate predictions of students' performance and helps reduce dropout rates.**

***Index Terms*—multiclass classification, machine learning, data balancing, student dropout, imbalanced datasets**

I. Introduction

One of the significant challenges facing higher education institutions globally is addressing students' diverse learning styles and academic performances. This diversity necessitates developing strategies to enhance student learning experiences and institutional efficiency. Institutions aim to proactively identify and support students at risk of academic failure or dropout by leveraging the extensive data they collect annually. This data encompasses students' educational paths, demographics, and socio-economic factors, creating a fertile ground for predictive analytics. The proactive anticipation of potential difficulties is crucial for implementing timely interventions. Predicting students' academic outcomes, including dropout risks and preferred majors, enables institutions to allocate resources effectively and provide targeted support, ultimately contributing to higher completion rates and better academic success. Understanding and utilizing relevant variables from university data can significantly improve the support provided to students, thereby enhancing overall educational outcomes and institutional performance.

The problem is formulated as a multiclass classification task (Graduate, enrolled, and Dropout), with a significant imbalance skewed towards one class. This imbalance is primarily due to more registered students than dropout students.

This work uses a dataset from the Polytechnic Institute of Portalegre (PIP) in Portugal [1] to build multiclass machine learning classification models to predict students who may be at risk of not completing their degrees on time or have difficulties in their academic path. The paper employs machine learning algorithms for student dropout multiclass classification, utilizing data-balancing techniques. Specifically, it uses SMOTE with Edited Nearest Neighbor (SMOTE-ENN) [2] to address the problem of an imbalanced dataset. Besides, the paper evaluates the performance of different ML models, including Random Forest (RF) [3], and boosting methods: XGBoost (Extreme Gradient Boosting) [4] and CatBoost (Categorical Boosting) [5].

The contributions of this article can be outlined as follows:

- Unlike other similar studies that focus on binary classification, our model considers three classes: Graduate (success), Dropout (failure), and Enrolled (relative success). Therefore, data balancing is required due to the unbalanced nature of these classes.
- The effectiveness of SMOTE-ENN in handling imbalanced multiclass datasets is emphasized and evaluated.
- Our model shows a significant improvement in the classification results for this multiclass classification task using XGBoost and CatBoost compared to previous work in the literature.

The paper is organized as follows: In Section II, we review the related work from the literature. Section III discusses the methodology applied in the paper, including dataset description, analysis, and machine learning model description. Section IV presents the experimental results and evaluation. Finally, Section V covers the findings and conclusion.

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

II. Related Work

The problem of predicting dropout students is an issue of interest for many researchers and higher education institutions. They focus on addressing student dropout using different datasets and machine learning techniques, which help them to offer timely interventions [6]–[8]. We review a few recent works similar to our work presented in this paper.

Mónica and Daniel [6] presented a study on early prediction models for student performance in higher education. They analyzed a dataset (3623 records and 25 independent variables) from the Polytechnic Institute of Portalegre (PIP), Portugal. The authors balanced the data using synthetic oversampling and classification models like standard machine learning and boosting algorithms; they utilized various machine-learning techniques to identify at-risk students early on. The best accuracy they achieved using standard machine learning was 72%.

In their study, Kumar [9] explored binary classification algorithms to predict student dropout behavior in universities. The author applied classification models utilizing conventional methods on a historical student dataset for ten academic years. The authors used the association matrix to select the most important features, and then they evaluated eight models using various performance measures. The Random Forest was the best model, with an accuracy of 90.41% and an AUC score of 93.8% followed by XGBoost which had an accuracy of 89.54% and an AUC of 93.1%.

Realinho et al. [10] developed a dataset containing 4424 records and 35 attributes. He enriched the database with several attributes to predict the dropout students enrolled in the institute. He did a comprehensive analysis of factors influencing student dropout and academic success. The author highlighted the significance of academic performance, attendance, and socioeconomic background as best predictors.

Mduma [11] focused on data balancing techniques for predicting student dropout using machine learning in two different datasets. The first dataset was Uwezo data learning at the country level in Tanzania, and the second one was collected in 2016 to assess student dropout rates in India. The study addressed the challenge of imbalanced datasets in educational data, proposing methods to improve the accuracy of predictive models. The research concludes that the SMOTE ENN balancing technique provides a good solution for achieving greater performance. On the other hand, the logistic regression model was the best model to correctly classify the largest number of dropout students (57348 for the Uwezo dataset and 13430 for the India dataset) using the confusion matrix as the evaluation matrix.

One study by Llauró [12] aimed to reduce the dropout rate in the first year of study toward a degree. They identified and compared the main variables affecting early university dropout rates across different knowledge areas and institutions.

A study by Nie and Dehrashid [13] evaluated student failure in higher education using a novel approach combining adaptive neuro-fuzzy inference system (ANFIS) and Harris Hawk's Optimizer (HHO) algorithms. Their strategy achieved 0.7565 AUC and 0.71543 MSE. The study used a dataset of 4424 records and 14 variables. Seventy percent of the data is used in the training phase, and the remaining thirty are used for testing.

Attiya and Shams [14] showed various data mining techniques and machine learning used to predict student retention in higher education. Their literature review examined 10 studies between 2020 and 2022 and identified the most important features and algorithms that mainly predict student retention.

Nema and Palwe [15] investigated using machine learning algorithms to predict student's academic success. Their study suggests using the Voting Classifier model which was trained using the Logistic Regression, Decision Tree and Random Forest models as base classifiers, the performance of the Voting Classifier model had an accuracy of 89.66% after training it on a dataset of 4424 rows or entries and 35 columns.

Bonifro [16] explored student dropout prediction using different machine-learning techniques. Their study analyzes a dataset of 15,000 students enrolled in several courses from eleven schools. The study emphasizes the potential of machine learning in identifying at-risk students, providing timely interventions and improving their academic careers by building new predictive systems. The study described in [17] addresses a binary classification problem involving dropout and graduate classes. They achieve high accuracy using Random Forest; however, the dataset used is limited to only six attributes.

Table I provides an overall summary of the related work, highlighting the dataset used in the study and its size, classification type, ML algorithms, and accuracy.

III. Methodology

This section describes the dataset, outlines the methods to address data imbalance issues, and explains the methodology for building and evaluating the classification models. The methodology employed in this work is summarized in Fig. 1.

A. Dataset

In order to build a classifier model which addresses the student dropout problem and predicts academic success, we used a public dataset collected in 2021 from the Polytechnic Institute of Portalegre (PIP) in Portugal [1]. A higher education institution created the dataset containing information about students' enrollment in various undergraduate degrees such as education, design, nursing, agronomy, journalism, social service, management, and technologies. This dataset has information about students at the time of enrollment before they begin their studies, such as academic background, socioeconomic data, demographic details, and student performance at the end of the first two semesters.

The PIP dataset comprises 35 features, 4424 records, and no missing values. The target value for each record is one of three categories: graduate (2208 records), Dropout (1421 records), or Enrolled (794 records), as shown in Fig. 2. Fig. 3 presents all features and their correlations (sorted by absolute value while keeping the sign) with the target variable — *Student Status*.

TABLE I
CLASSIFICATION METHODS USED IN THE LITERATURE FOR STUDENT DROPOUT PREDICTION

Paper	Year	Dataset	Sample Size	Classification	Algorithms	Accuracy
[6]	2021	PIP dataset	3623	Multi-class	XGBoost	73%
[9]	2024	PIP dataset	4424	Binary	Random Forest	90.4%
[13]	2024	PIP dataset	4424	Multi-class	HHO-ANFIS	AUC = 0.76
[15]	2023	PIP dataset	4424	Binary	RF	91.7%
[11]	2023	Uwezo, India	61,340	Binary	Logistic Regression	93.4%
[16]	2020	Pseudo-anonymized data (8 ATTRs)	15,000	Binary	CC+RF	87%
[18]	2019	Budapest University (BME)	10,196	Binary	ANN	85.8%
[17]	2020	BPPD (6 ATTRs)	44,406	Binary	RF	95%
[19]	2022	(SUSIS) Dataset	230K	Binary	LSTM	88%

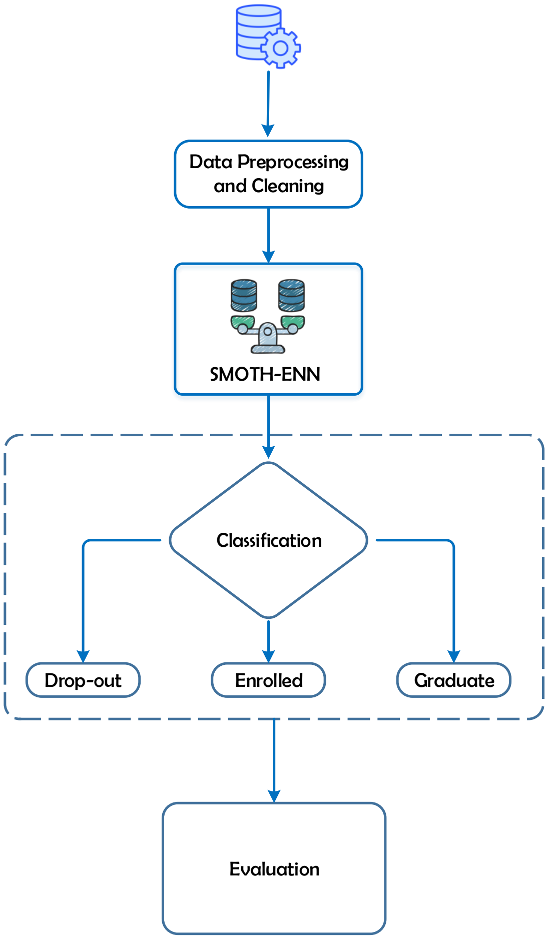

Fig. 1. Overview of Student Dropout Prediction Design

B. Data Sampling and Pre-processing

Data sampling and balancing techniques are commonly used to address imbalanced datasets, particularly in multiclass scenarios. Difference methods used for data balancing include SMOTE [20], SMOTE-ENN [2], and others.

SMOTE generates synthetic samples, increasing the instances of the minority classes. It creates new instances that combine existing ones, which helps balance the class distribution. SMOTE-ENN is a hybrid balancing method that combines SMOTE and Editing Nearest Neighbor (ENN). After SMOTE, ENN removes noisy and misclassified instances from the dataset. SMOTE and ENN help balance the class distribution and improve the dataset's quality. We used this approach in this paper.

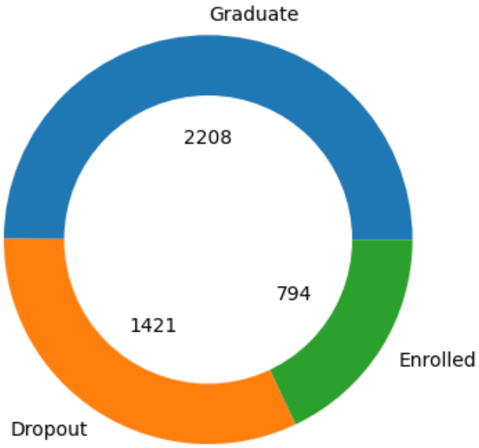

Fig. 2. Distribution of students' records across the three categories

Fig. 4 shows the distribution of classes for the *Student Status* feature after balancing the data using the SMOTE-ENN technique. The minority class **Enrolled** increased significantly due to SMOTE; the new 1615 instances indicate the effectiveness of SMOTE in generating synthetic samples. The majority class **Graduate** is dropped significantly from 2208 to 876 because ENN removes many misclassified instances, which is typical for the majority class with the highest noise potential. Finally, the **Dropout** class is slightly decreased from 1421 into 1329, remaining relatively close to its original count.

In general, SMOTE-ENN aims to rebalance the classes by increasing the minority classes and then cleaning the dataset to remove the misclassified instances, resulting in the observed changes in class distributions.

C. Classification Models

In this work, we focus on multiclass classification to predict student dropout. Three supervised learning algorithms, Ran-

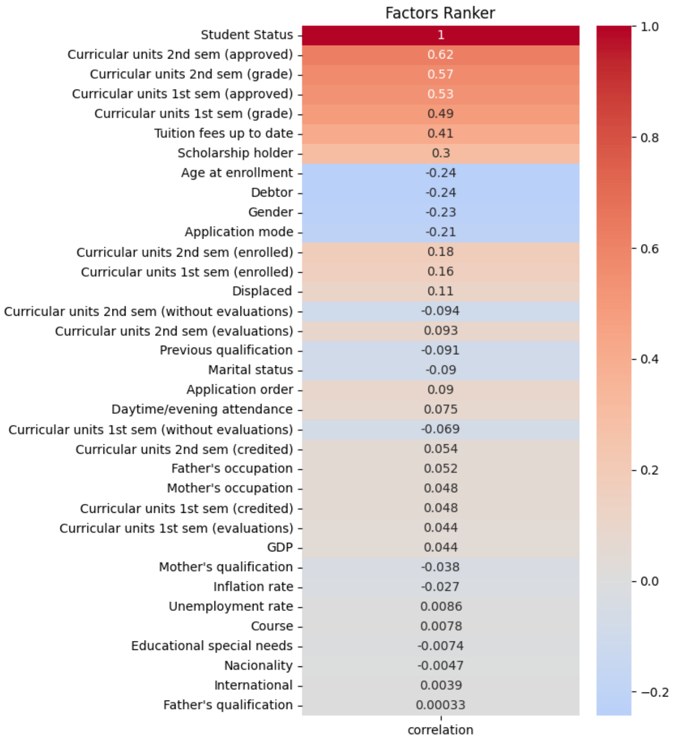

Fig. 3. The features of the PIP dataset along with their correlations to the target value

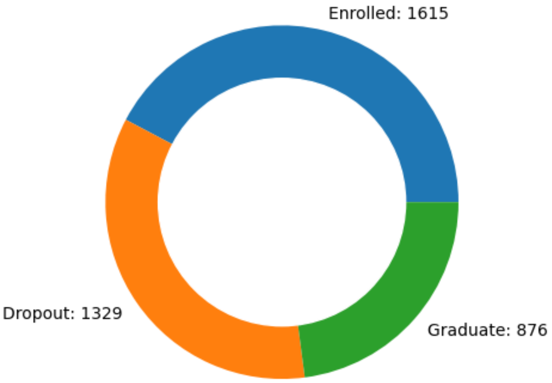

Fig. 4. Distribution of students' records After applying SMOTE-ENN

dom Forest (RF), XGBoost, and CatBoost, were selected for their high accuracy in this task.

- Random Forest excels in handling imbalanced datasets through its ensemble approach, which combines several decision trees to improve classification accuracy. It is considered one of the top-performing algorithms, outperforming 178 classifiers across 17 different families [21]. Its ability to handle many features and its robustness against overfitting make it well-suited for the complexities of dropout prediction across various classes.
- XGBoost is an ensemble-supervised ML algorithm and gradient-boosting library for classification and regression problems. It has built-in mechanisms to address the class imbalance, such as adjusting the scale of gradients, which enhances the model's ability to predict minority classes effectively.
- CatBoost considered another algorithm for gradient boosting over decision trees. It is is designed to work with categorical features and imbalanced data efficiently. Its use of ordered boosting and symmetric trees helps mitigate issues related to class imbalance, this provides more accurate predictions for less frequent classes.

These algorithms were chosen for their ability to manage imbalanced classes and their strong performance in multiclass classification tasks, making them ideal for predicting student dropout in the educational dataset.

IV. RESULTS AND EVALUATION

In this section, we provide a detailed explanation of model development and testing results for each model. We trained three classifier models using the provided dataset: Random Forest, XGBoost, and CatBoost. These models are used for multiclass classification to predict student States—specifically, whether students will *graduate*, *dropout*, or remain *enrolled*. Therefore, the testing focuses on evaluating the performance of each class prediction. We used *recall*, *precision*, and *f1-measures* as evaluation metrics for assessing the performance of the classification model on each class, and overall accuracy for evaluating the classifier in general.

The data described in section III is used for training and testing, with a split ratio of 75% for training and 25% for testing.

Fig. 5 presents the accuracy values for the three algorithms. The figure shows the accuracy for each classifier applied to the original imbalanced data and the balanced data after applying SMOTE-ENN. Two main observations can be made: first, the balanced data using SMOTE-ENN enhanced the accuracy values for all models; for example, XGBoost's accuracy increased from 77.9% to 94.6 %. Second, XGBoost provides the highest accuracy among all the classifiers, surpassing the results found in the literature for this multi-class classification problem, as demonstrated in Table I.

Figs. 6 to 8 show the confusion matrices, which provide a detailed breakdown of the performance of our classification models: Random Forest, XGBoost, and CatBoost, respectively. Each matrix allows us to visualize the classifier's ability to distinguish between the three classes: Dropout, Enrolled, and Graduate. This provides a comprehensive view of the model's performance. For instance, the XGBoost model (see Fig. 7) correctly classified 304 out of 321 dropout cases, resulting in high precision and recall for the dropout class, as shown in Table II. However, 15 instances were misclassified as enrolled and two as graduates, indicating where the model's predictions could be improved.

TABLE II
PRECISION, RECALL, AND F1-SCORE FOR XGBOOST, RANDOM FOREST, AND CATBOOST MODELS

Model	**XGBoost**			**Random Forest**			**CatBoost**		
	f1-score	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**	**precision**	**recall**
Dropout	0.953	0.959	0.947	0.938	0.940	0.935	0.944	0.947	0.941
Enrolled	0.943	0.930	0.956	0.920	0.906	0.933	0.931	0.925	0.938
Graduate	0.945	0.959	0.930	0.915	0.936	0.895	0.934	0.942	0.926

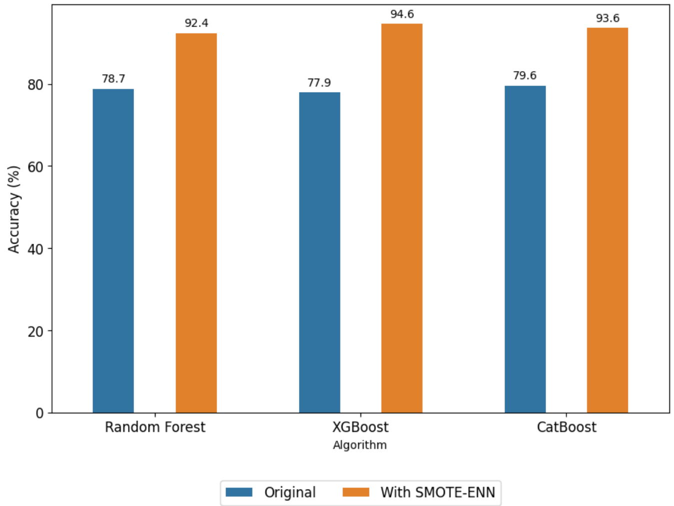

Fig. 5. Accuracy Comparison of Algorithms with and without SMOTE-ENN

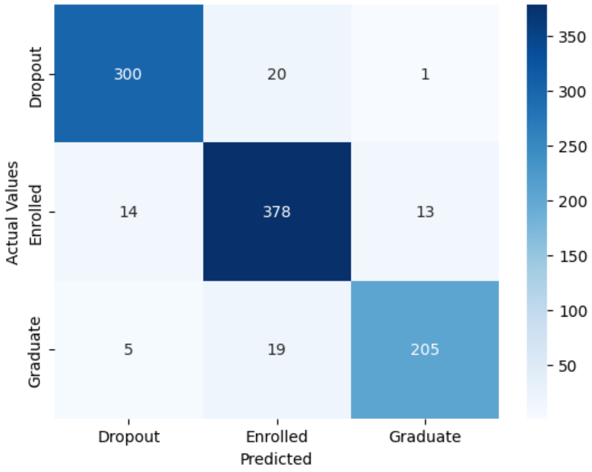

Fig. 6. Confusion Matrix for the Random Forest Model

Table II shows a detailed comparison of the multiclass classifier results for XGBoost, Random Forest, and CatBoost, showing each class's precision, Recall, and F1-score values.

Overall, the boosting algorithm XGBoost shows the best performance results across all metrics and for three classes, closely followed by CatBoost, both models outperform the Random Forest. The misclassified instances highlight areas for improvement, particularly in minimizing confusion between similar classes like enrolled and graduate. Additionally, refinement, such as classifier-specific feature selection, could further enhance predictive performance.

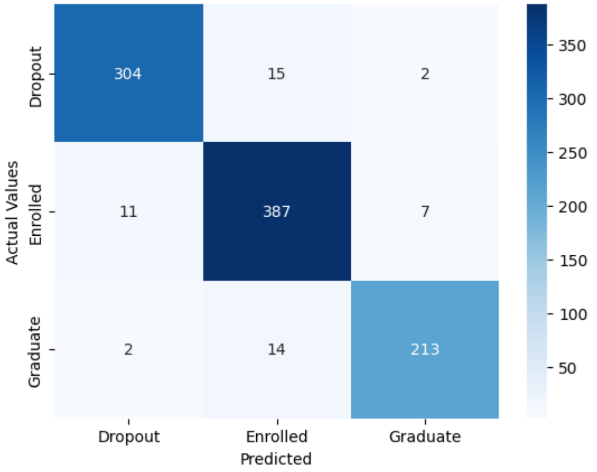

Fig. 7. Confusion Matrix for the XGBoost Model

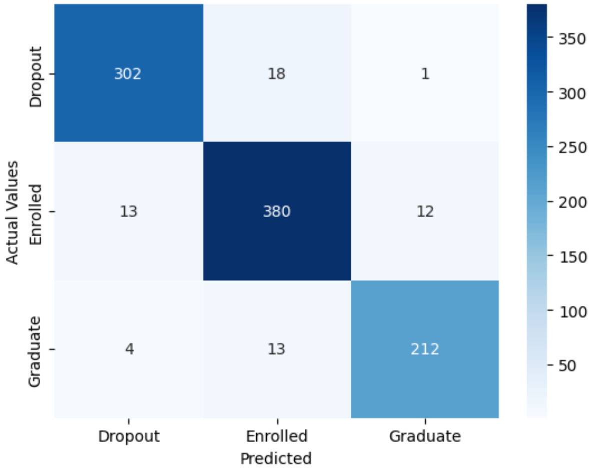

Fig. 8. Confusion Matrix for the CatBoost Model

V. CONCLUSION

This paper uses a dataset from the Polytechnic Institute of Portalegre (PIP) to build a classification model to predict student academic performance. We address an imbalanced multiclass classification problem with three categories for students: graduate, dropout, and enrolled. To address the class imbalance, we apply SMOTE-ENN hybrid sampling technique, which increases the minority classes using SMOTE and removes misclassified instances from the majority class using ENN. Three machine learning models—Random Forest, XGBoost, and CatBoost—are trained and used for this classification task. A comprehensive performance evaluation uses various metrics, revealing that the boosting algorithm XGBoost achieves the highest accuracy at 94.6%, followed by CatBoost. This result represents a significant improvement over previous approaches in this multiclass classification context. This research is important for predicting student academic performance and contributes to efforts to reduce dropout rates.

References

[1] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predict students' dropout and academic success (1.0) [data set]," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5777340

[2] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early prediction of student's performance in higher education: A case study," in *Trends and Applications in Information Systems and Technologies: Volume 1*. Springer, 2021, pp. 166–175.

[7] N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," *Data Science Journal*, 2019.

[8] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[9] K. Gupta, K. Gupta, P. Dwivedi, and M. Chaudhry, "Binary classification of students' dropout behaviour in universities using machine learning algorithms," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2024, pp. 709–714.

[10] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022.

[11] N. Mduma, "Data balancing techniques for predicting student dropout using machine learning," *Data*, vol. 8, no. 3, p. 49, 2023.

[12] A. Llauró, D. Fonseca, S. Romero, M. Aláez, J. T. Lucas, and M. M. Felipe, "Identification and comparison of the main variables affecting early university dropout rates according to knowledge area and institution," *Heliyon*, vol. 9, no. 6, 2023.

[13] J. Nie and H. A. Dehrashid, "Evaluation of student failure in higher education by an innovative strategy of fuzzy system combined optimization algorithms and ai," *Heliyon*, vol. 10, no. 7, 2024.

[14] W. M. Attiya and M. B. Shams, "Predicting student retention in higher education using data mining techniques: A literature review," in *2023 International Conference On Cyber Management And Engineering (CyMaEn)*. IEEE, 2023, pp. 171–177.

[15] T. Nema and S. Palwe, "Predicting students' academic success using machine learning algorithms," in *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, 2023, pp. 1–7.

[16] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student dropout prediction," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I*. Springer, 2020, pp. 129–140.

[17] N. S. Sani, A. F. M. Nafuri, Z. A. Othman, M. Z. A. Nazri, and K. N. Mohamad, "Drop-out prediction in higher education among b40 students," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020.

[18] B. Kiss, M. Nagy, R. Molontay, and B. Csabay, "Predicting dropout using high school and first-semester academic achievement measures," in *2019 17th international conference on emerging eLearning technologies and applications (ICETA)*, 2019, pp. 383–389.

[19] H. S. Brdesee, W. Alsaggaf, N. Aljohani, and S.-U. Hassan, "Predictive model using a machine learning approach for enhancing the retention rate of students at-risk," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, pp. 1–21, 2022.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[21] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.