Early Student Dropout Prediction Using Pre-Acceptance Variables with Gaussian Naive Bayes and XGBoost

Saladin Zhalifunnas Ahfar
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
saladin.ahfar@binus.ac.id

Michael Hendrilie
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
michael.hendrilie@binus.ac.id

Valentino Allen Prasetyo
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
valentino.prasetyo@binus.ac.id

Rifqi Alfinnur Charisma
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
rifqi.charisma001@binus.ac.id

Henry Lucky
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
henry.lucky@binus.ac.id

***Abstract*—Student dropout prediction is an important task in educational data mining because early identification of at-risk students can support more timely institutional intervention. This study investigates whether dropout risk can be predicted using only pre-acceptance variables by comparing Gaussian Naive Bayes and XGBoost. The original dataset contained 4,424 student records with three target categories: Dropout, Graduate, and Enrolled. Since this study focuses on binary early dropout risk identification, 794 Enrolled records were removed, resulting in 3,630 records consisting of 2,209 Graduate and 1,421 Dropout records. Two feature configurations were evaluated: a Background-Only Setting and a Background-Application Setting. The models were trained using the same preprocessing, 80:20 train-test split, SMOTE handling, and evaluation metrics. The best overall performance was achieved by XGBoost under the Background-Application Setting, with an accuracy of 0.7328, precision of 0.6510, recall of 0.6831, and F1-score of 0.6667. These results show that pre-acceptance variables can provide useful early dropout risk signals, although the performance remains moderate. GitHub: https://github.com/LecyLecy/student-dropout-preacceptance**

Keywords—student dropout prediction, pre-acceptance variables, educational data mining, XGBoost, Gaussian Naive Bayes, higher education.

I. Introduction

Student dropout remains a major challenge in higher education because late identification of at-risk students limits the opportunity for timely intervention. In recent years, educational data mining and machine learning have been widely used to predict student outcomes such as graduation, dropout, academic success, and retention in order to support earlier institutional decision-making [1]–[4]. However, many existing studies still rely on semester-based academic performance variables, class activity records, or post-enrollment information, even though such information is not always available at the earliest stage of admission and student monitoring [2], [3], [5]. As a result, the practical value of these models for very early warning systems may still be limited.

Recent studies show that XGBoost has strong potential for dropout and academic success prediction, while Naive Bayes remains useful as a simple and interpretable baseline [6], [7]. Since educational prediction also needs practical value and interpretability [8], comparing these two models under a restricted early-stage feature setting remains relevant.

This study uses the Predict Students' Dropout and Academic Success dataset, whose original target consists of Dropout, Graduate, and Enrolled categories [9]. To align the task with binary early dropout risk identification, Enrolled records are excluded, while Graduate is encoded as 0 and Dropout is encoded as 1. Although dropout prediction has also been widely studied in online learning environments [10], this study focuses on a stricter pre-acceptance setting in higher education. Two feature configurations are evaluated: the Background-Only Setting and the Background-Application Setting. Through this design, the study aims to examine whether meaningful dropout prediction can already be achieved at a very early stage and whether XGBoost provides more balanced performance than Gaussian Naive Bayes under the two pre-acceptance settings.

II. Literature Review

A. Student Dropout Prediction in Higher Education

Student dropout prediction has become an important topic in educational data mining because institutions need data-driven methods to identify at-risk students before dropout occurs. Previous studies have used student background, secondary school records, first-year academic information, and pre-enrollment data to support earlier identification of dropout risk [11], [12]. These studies show that early prediction is useful for intervention planning, but the practical value of a model

depends strongly on when its input variables become available.

Student outcome prediction has also been studied through different algorithms and datasets. Prior reviews and applications show that model effectiveness depends on feature design, evaluation metrics, institutional context, and algorithm selection [13]–[15]. Therefore, model comparison remains important when the feature setting is restricted to early-stage variables.

B. Machine Learning Models for Educational Prediction

Advanced tree-based and ensemble methods are commonly used in educational prediction because they can capture non-linear relationships among variables. XGBoost is relevant for this study because it performs well on structured data, has been applied to imbalanced classification, and has shown strong performance in learner performance prediction and broader classification settings [16]–[19]. These characteristics make it suitable for datasets that combine demographic, socioeconomic, macroeconomic, and application-related variables.

Simpler models such as Naive Bayes remain useful because they are efficient, interpretable, and easy to implement. Previous educational prediction studies show that Naive Bayes can still provide useful baseline performance in learning outcome prediction [20]. Therefore, comparing Gaussian Naive Bayes with XGBoost allows this study to contrast a lightweight probabilistic model with a stronger boosting-based model under the same experimental setting.

C. Dropout Prediction with XGBoost and Related Approaches

Recent dropout prediction studies have increasingly used XGBoost and ensemble-based methods. Random Forest and XGBoost have been applied to identify students at risk, while other studies combine XGBoost with genetic algorithms, fuzzy-based modeling, or explainable machine learning pipelines in higher education and online learning contexts [21]–[25]. These studies confirm the relevance of XGBoost for dropout prediction, but many of them use broader feature settings, online activity data, or post-enrollment variables. This highlights the need for clear experimental boundaries when the objective is early prediction.

D. Research Gap and Study Positioning

Based on the reviewed literature, the main gap is that many dropout prediction studies still use mixed feature settings that include behavioral, academic, post-enrollment, or semester-based variables. Although these variables can improve performance, they reduce the usefulness of the model for very early warning systems because they are not available before students begin their studies. Therefore, a controlled comparison between XGBoost and Gaussian Naive Bayes under a strict pre-acceptance setting is still needed.

A closely related baseline study used the same Predict Students' Dropout and Academic Success dataset with XGBoost and achieved strong performance by using a broader feature set, including demographic, socio-economic, and academic variables such as admission grade and semester-based curricular unit records [6]. However, several of these variables are post-acceptance or post-enrollment variables because they are only available after admission or after students have started their academic activities. In contrast, this study intentionally excludes those variables and uses only pre-acceptance variables to preserve the early-warning purpose of the model. Therefore, this study positions itself as a controlled comparative analysis of Gaussian Naive Bayes and XGBoost for early student dropout prediction using only pre-acceptance variables. The study does not aim to propose a new algorithm or maximize performance using all available variables. Instead, it evaluates whether useful dropout risk signals can already be identified before semester-based academic and post-acceptance variables are available. By comparing the Background-Only Setting with the Background-Application Setting, this study also examines whether application-related variables improve early prediction without violating the pre-acceptance boundary.

III. METHODS

This section explains the research design used to evaluate early student dropout prediction using pre-acceptance variables. The overall workflow includes dataset collection, target transformation, feature setting definition, preprocessing, model development, evaluation, and interpretation of the best-performing model.

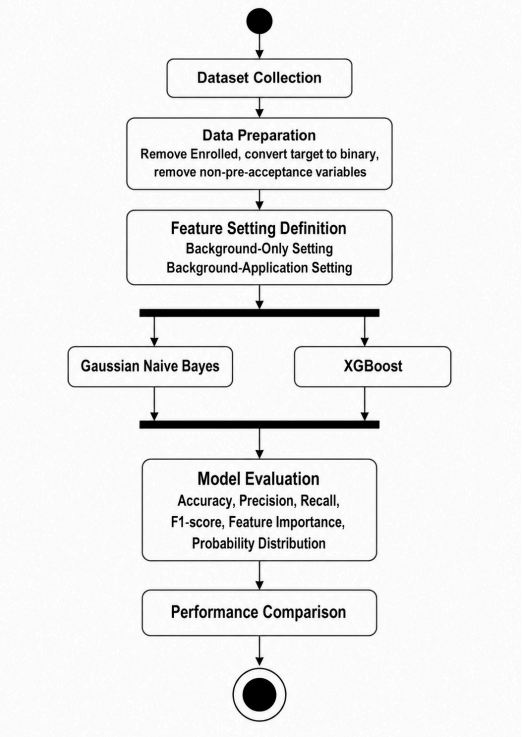


Fig. 1. Research workflow of the proposed pre-acceptance student dropout prediction framework.

A. Dataset and Early Target Construction

As shown in Fig. 1, the workflow summarizes the main stages of the experiment. This study uses the publicly available Predict Students' Dropout and Academic Success dataset as the main source of analysis [9]. The dataset is originally associated with the UCI Machine Learning Repository and was accessed in this study through Kaggle. The dataset link used in this study is https://www.kaggle.com/datasets/thedevastator/higher-edu

cation-predictors-of-student-retention. The dataset contains demographic, family, application, macroeconomic, academic, and final target information for undergraduate students. Since this study focuses on early dropout prediction, the original multiclass target is transformed into a binary task by removing Enrolled records and encoding Graduate = 0 and Dropout = 1. After this transformation, the final dataset contains 3,630 records, consisting of 2,209 Graduate records and 1,421 Dropout records, as summarized in Table I. Semester-based academic variables and post-acceptance administrative variables are excluded so that the model uses only information available before student acceptance is finalized.

TABLE I. DATASET TRANSFORMATION AND TARGET DISTRIBUTION

Description	Count
Initial dataset	4,424
Removed Enrolled records	794
Final binary dataset	3,630
Graduate records	2,209
Dropout records	1,421

The final binary dataset consists of 60.85% Graduate records and 39.15% Dropout records. Since the class distribution is not fully balanced, SMOTE is applied only to the training data to reduce bias toward the majority class without changing the original test distribution.

B. Pre-Acceptance Variable Settings

To make the comparison more informative, two feature settings are defined. The Background-Only Setting represents a more restrictive early-stage perspective, while the Background-Application Setting extends it by including application-related variables that are still available during the admission stage. The Background-Only Setting uses background-only variables, consisting of Marital status, Nacionality, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation, Displaced, Educational special needs, Gender, Age at enrollment, International, Unemployment rate, Inflation rate, and GDP. The Background-Application Setting uses background plus application-related variables. In addition to all variables in Background-Only Setting, this setting includes Application mode, Application order, Course, Daytime/evening attendance, and Previous qualification. Post-acceptance and semester-based academic variables are excluded, including Debtor, Tuition fees up to date, Scholarship holder, and all first- and second-semester curricular unit variables. These variables are removed because they are only available after admission or after students begin their academic activities.

C. Data Preparation and Model Development

After target transformation and variable filtering, the dataset is divided into training and testing sets using an 80:20 split ratio. The same split strategy is applied consistently across both feature settings to preserve fairness in model comparison. Because the selected variables include both categorical and numerical attributes, preprocessing is conducted before model training. Categorical variables are transformed using encoding, while numerical variables are standardized where necessary. To address possible class imbalance in the binary target distribution, Synthetic Minority Over-sampling Technique (SMOTE) is applied only to the training data. SMOTE generates synthetic minority samples by interpolating between a minority sample and its nearest minority-class neighbor, as shown in (1):

$$x_{new} = x_i + \delta(x_{nn} - x_i),\ 0 \le \delta \le 1 \qquad (1)$$

where x_i is a minority-class sample, x_{nn} is one of its nearest minority-class neighbors, and δ is a random value between 0 and 1. This is done to reduce prediction bias toward the majority class and to improve the model's ability to identify dropout cases more effectively. Two classification models are used in this study. Gaussian Naive Bayes is selected as the baseline model because it is simple, efficient, and interpretable. XGBoost is selected as the stronger comparison model because it is widely recognized for its predictive performance in structured tabular data and its ability to capture more complex feature relationships. Both models are trained separately under Background-Only Setting and Background-Application Setting so that their performance can be compared not only across algorithms, but also across the two early-stage variable configurations.

D. Evaluation Strategy

Model performance is evaluated using accuracy, precision, recall, and F1-score. These evaluation measures are chosen because they provide a more complete assessment than accuracy alone, particularly when class imbalance may affect model behavior. The evaluation is carried out for both Gaussian Naive Bayes and XGBoost under each feature setting. This produces four experimental conditions in total: Gaussian Naive Bayes with Background-Only Setting, XGBoost with Background-Only Setting, Gaussian Naive Bayes with Background-Application Setting, and XGBoost with Background-Application Setting. After the best-performing scenario is identified, feature importance analysis and prediction probability distribution are used to interpret how the selected model uses pre-acceptance variables and why its performance remains moderate. The final comparison is then used to determine whether background-only variables are sufficient for meaningful early dropout prediction, whether application-related variables improve predictive performance, and whether XGBoost provides more balanced performance than Gaussian Naive Bayes under the strict pre-acceptance setting.

IV. Results and Discussion

A. Experimental Performance Results

The experiment was conducted under four scenarios. The Background-Only Setting uses background-only variables, while the Background-Application Setting extends it by adding application-related variables. Each setting was evaluated using Gaussian Naive Bayes and XGBoost under the same train-test split, preprocessing procedure, class imbalance handling, and evaluation metrics.

TABLE II. MODEL PERFORMANCE COMPARISON ACROSS FEATURE SETTINGS

Feature Setting	Model	Accuracy	Precision	Recall	F1-score
Background-Only Setting	Gaussian Naive Bayes	0.3953	0.3901	**0.9683**	0.5561
	XGBoost	0.7011	0.6227	0.5986	0.6104
Background-Application Setting	Gaussian Naive Bayes	0.3994	0.3895	0.9437	0.5514
	XGBoost	**0.7328**	**0.6510**	0.6831	**0.6667**

The results show that XGBoost outperformed Gaussian Naive Bayes in both feature settings based on accuracy, precision, and F1-score. The best overall result was achieved by XGBoost under the Background-Application Setting, with an accuracy of 0.7328, precision of 0.6510, recall of 0.6831, and F1-score of 0.6667. This indicates that adding application-related variables improved the model's ability to distinguish between Graduate and Dropout cases. Gaussian Naive Bayes achieved very high recall in both settings, but its low precision shows that the model produced many false positives. Since XGBoost under the Background-Application Setting achieved the strongest overall balance, it was selected as the best-performing model in this study. This result can be explained by XGBoost's ability to capture non-linear relationships and by the additional early-stage information provided by application-related variables such as application mode, application order, course, attendance mode, and previous qualification. Therefore, XGBoost under the Background-Application Setting provides the best balance between predictive performance, minority class detection, and realistic early-stage feature availability.

B. Feature Importance Analysis

After identifying XGBoost under the Background-Application Setting as the best-performing scenario, feature importance analysis was used to examine which variables contributed most to the model. As shown in Fig. 2, the model relies on a combination of background and application-related variables rather than a single dominant predictor. This supports the result in Table II, where the Background-Application Setting improved the performance of XGBoost compared with the Background-Only Setting.

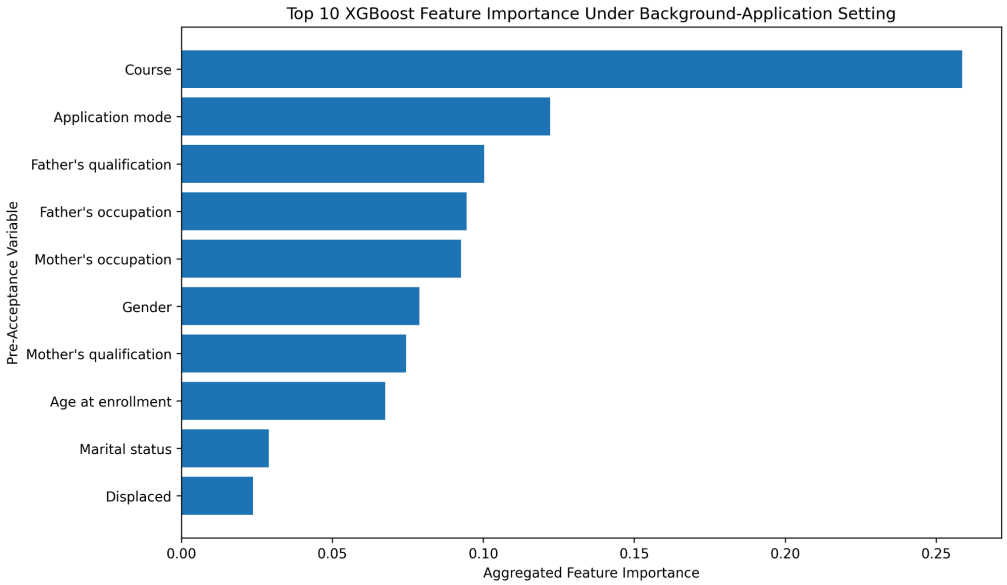


Fig. 2. Top 10 XGBoost Feature Importance Under Background-Application Setting

The presence of application-related variables, especially course choice and application pathway, among the important predictors indicates that application-stage information can add useful early-stage signals for dropout risk prediction. This explains why the Background-Application Setting produced better performance than the Background-Only Setting.

C. Prediction Probability Distribution

To further interpret the best-performing model, the predicted dropout probability distribution of XGBoost under Background-Application Setting was analyzed. As shown in Fig. 3, the predicted probabilities of Graduate and Dropout cases are not perfectly separated, which explains why the model performance remains moderate even though it is the strongest scenario in this study.

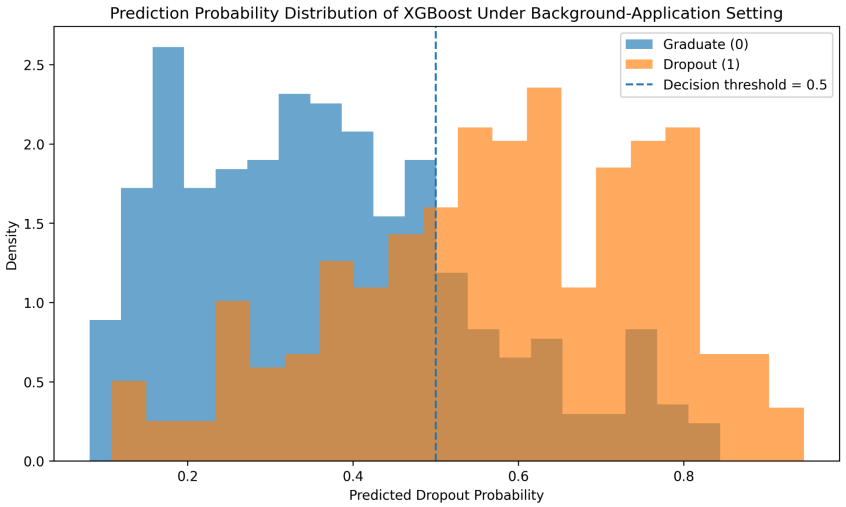


Fig. 3. Prediction Probability Distribution of XGBoost Under Background-Application Setting

The distribution shows that some Graduate and Dropout cases overlap around the middle probability range. This overlap indicates that pre-acceptance variables provide useful early signals, but they cannot fully separate both classes because semester-based academic performance and post-acceptance variables are intentionally excluded. Therefore, the model is more suitable as an early warning support tool rather than a final decision-making system.

D. Overall Discussion

Overall, pre-acceptance variables provide useful early signals for dropout risk identification, although performance remains moderate because strong post-acceptance and semester-based academic variables are intentionally excluded. XGBoost produced more balanced performance than Gaussian Naive Bayes, while Gaussian Naive Bayes achieved high recall but low precision. The improvement from the Background-Only Setting to the Background-Application Setting shows that application-related variables add meaningful information beyond background characteristics. Therefore, the proposed approach is valuable not because it maximizes accuracy, but because it tests whether dropout risk can be identified before academic performance records emerge.

V. Conclusion

This study evaluated early student dropout prediction using only pre-acceptance variables by comparing Gaussian Naive Bayes and XGBoost under the Background-Only and Background-Application settings. The results show that XGBoost under the Background-Application Setting achieved the best overall performance, with an accuracy of 0.7328, precision of 0.6510, recall of 0.6831, and F1-score of 0.6667. The main contribution is not to outperform post-acceptance prediction, but to show that useful dropout risk signals can still be identified earlier, although performance remains moderate because stronger post-acceptance and semester-based academic variables are intentionally excluded.

Future work should validate the framework using real admission data from other institutions, examine whether early risk signals remain consistent after semester records become available, and include fairness and ethical evaluation before deployment. A pilot implementation with academic advisors can also be conducted to test whether the model is useful as an early warning support tool.

Author Contribution

Michael Hendrilie: Software, Data Curation, Validation. **Saladin Zhalifunnas Ahfar**: Conceptualization, Methodology, Formal Analysis, Writing-Original Draft. **Valentino Allen Prasetyo**: Investigation, Visualization, Writing-Review and Editing. **Rifqi Alfinnur Charisma**: Supervision, Project Administration. **Henry Lucky**: Supervision, Project Administration.

References

[1] A. Ridwan, T. Sutikno, I. Riyadi, and W. C. Wahyudin, "On-Time Student Graduation Prediction Modeling: A Comparative Analysis of Naive Bayes Algorithm and Other Data Mining Classifications," JOINCS (Journal of Informatics, Network, and Computer Science), vol. 8, no. 2, pp. 128–135, 2025, doi: 10.21070/joincs.v8i2.1679.

[2] S. Kim, "Student Dropout Prediction for University with High Precision and Recall," Applied Sciences, vol. 13, no. 10, Art. no. 6275, 2023, doi: 10.3390/app13106275.

[3] Z. Song et al., "All-Year Dropout Prediction Modeling and Analysis for University Students," Applied Sciences, vol. 13, no. 2, Art. no. 1143, 2023, doi: 10.3390/app13021143.

[4] A. R. Sinaga, "Variabel Non Akademik Untuk Memprediksi Prestasi Siswa Dengan Data Mining Menggunakan Metoda Naïve Bayes," Nuansa Informatika, vol. 17, no. 2, 2023.

[5] M. J. Shayegan and R. Akhtari, "A Stacking Machine Learning Model for Student Performance Prediction Based on Class Activities in E-Learning," Computer Systems Science and Engineering, vol. 48, no. 5, pp. 1251–1272, 2024, doi: 10.32604/csse.2024.052587.

[6] A. Ridwan, A. M. Priyatno, and L. Ningsih, "Predict Students' Dropout and Academic Success with XGBoost," Journal of Education and Computer Applications, vol. 1, no. 2, pp. 1–8, 2024, doi: 10.69693/jeca.v1i2.13.

[7] T. Destanto and H. W. Nugroho, "Comparison of Data Mining for Classifying Student Graduation Levels Using Naive Bayes, Decision Tree, and Random Forest Methods (Case Study of the Undergraduate Program at Mitra Indonesia University)," CCIT Journal, vol. 18, no. 1, pp. 48–55, 2025, doi: 10.33050/ccit.v18i1.3409.

[8] K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond Performance: Explaining and Ensuring Fairness in Student

Academic Performance Prediction with Machine Learning," Applied Sciences, vol. 15, no. 15, Art. no. 8409, 2025, doi: 10.3390/app15158409.
[9] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," Data, vol. 7, no. 11, Art. no. 146, 2022, doi: 10.3390/data7110146.
[10] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, "A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses," ACM Computing Surveys, vol. 53, no. 3, Art. no. 57, pp. 1–34, 2020, doi: 10.1145/3388792.
[11] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student Dropout Prediction," in Artificial Intelligence in Education, I. I. Bittencourt et al., Eds. Cham, Switzerland: Springer, 2020, pp. 129–140, doi: 10.1007/978-3-030-52237-7_11.
[12] B. Carballo-Mendívil, A. Arellano-González, N. J. Ríos-Vázquez, and M. d. P. Lizardi-Duarte, "Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data," Applied Sciences, vol. 15, no. 16, Art. no. 9202, 2025, doi: 10.3390/app15169202.
[13] E. F. Agyemang, J. A. Mensah, O.-A. Ampomah, L. Agyekum, J. Akuoko-Frimpong, A. Quansah, and O. M. Akinlosotu, "Predicting Students' Academic Performance Via Machine Learning Algorithms: An Empirical Review and Practical Application," Computer Engineering and Intelligent Systems, vol. 15, no. 1, 2024, doi: 10.7176/CEIS/15-1-09.
[14] S. Begum and S. S. Padmannavar, "Student Performance Analysis using Bayesian Optimized Random Forest Classifier and KNN," International Journal of Engineering Trends and Technology, vol. 71, no. 5, pp. 132–140, 2023, doi: 10.14445/22315381/IJETT-V71I5P213.
[15] A. A. Alharbi and J. Allohibi, "A new hybrid classification algorithm for predicting student performance," AIMS Mathematics, vol. 9, no. 7, pp. 18308–18323, 2024, doi: 10.3934/math.2024893.
[16] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," International Journal of Distributed Sensor Networks, vol. 18, no. 6, 2022, doi: 10.1177/15501329221106935.
[17] S. Hakkal and A. A. Lahcen, "XGBoost To Enhance Learner Performance Prediction," Computers and Education: Artificial Intelligence, vol. 7, p. 100254, 2024.
[18] Z. Ergul Aydin and Z. Kamisli Ozturk, "Performance Analysis of XGBoost Classifier with Missing Data," in Proceedings of the 1st International Conference on Computing and Machine Intelligence (ICMI 2021), 2021.
[19] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and Random Forest Algorithms: An In-Depth Analysis," Pakistan Journal of Scientific Research, vol. 3, no. 1, pp. 26–31, 2023.
[20] L. Baridah and R. A. Putri, "Predicting Student Learning Outcomes in Vocational Computer and Network Engineering Using Naïve Bayes," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 8, no. 3, pp. 667–676, 2025, doi: 10.24014/ijaidm.v8i3.38333.
[21] L. G. R. Putra, D. D. Prasetya, and Mayadi, "Student Dropout Prediction Using Random Forest and XGBoost Method," INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.
[22] H. E. Aouarib, S. E. Henouda, and F. Z. Laallam, "Predicting Student Dropout in MOOCs Using Genetic Algorithms and XGBoost," Journal of Information Systems Engineering and Management, vol. 10, no. 57s, 2025.
[23] W. Romsaiyud, P. Nurarak, T. Phiasai, M. Chadakaew, N. Chuenarom, P. Aksorn, and A. Thammakij, "Predictive Modeling of Student Dropout Using Intuitionistic Fuzzy Sets and XGBoost in Open University," in Proceedings of the 2024 The 7th International Conference on Machine Learning and Machine Intelligence (MLMI 2024), Osaka, Japan, Aug. 2024, pp. 104–110, doi: 10.1145/3696271.3696288.
[24] A. Bisri, Supardi, Y. Heryatun, Hunainah, and A. Navira, "Educational data mining model using support vector machine for student academic performance evaluation," Journal of Education and Learning, vol. 19, no. 1, pp. 478–486, 2025, doi: 10.11591/edulearn.v19i1.21609.
[25] A. Bettahi, F.-Z. Belouadha, and H. Harroud, "A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education," Algorithms, vol. 18, no. 10, Art. no. 662, 2025, doi: 10.3390/a18100662.