EdinBurg Peer Reviewed Journals and Books Publishers
Journal of Information and Technology
Vol. 5||Issue 4||pp 31-40||June||2025
Email: info@edinburgjournals.org||ISSN: 3080-9576

EdinBurg
Peer Reviewed Journals & Books Publishers

Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis

Uwimana Olive[1*], Musabe Jean Bosco [2] & Nyesheja Muhire Enan[3]
[1,3]Faculty of Computing and Information Sciences, University of Lay Adventists of Kigali
[2] School of Science & Technology, Kigali Independent University
Corresponding Emails: uwimanaolive79@gmail.com; bosulus@gmail.com; nyenani@gmail.com

Accepted: 09 May 2025 || Published: 06 June 2025

Abstract

Student dropout in higher education remains a global challenge, particularly in developing regions where early interventions are hindered by reliance on traditional indicators like GPA or attendance. This study addresses the issue by proposing a predictive model using ensemble machine learning techniques, integrating Logistic Regression, Random Forest, and AdaBoost. These models were combined using soft and hard voting classifiers to enhance prediction accuracy and reliability. The dataset, comprising 4,424 student records, includes demographic, academic, and socio-economic features. Results showed that the soft voting ensemble achieved the highest accuracy (80.56%) and AUC (91%), outperforming individual classifiers. Feature importance analysis revealed academic performance, tuition status, and parental background as key predictors of dropout. The model not only identifies at-risk students with high precision but also offers actionable insights for early intervention. This approach equips higher learning institutions with data-driven strategies to improve retention and student success outcomes.

Keywords: *Student dropout, Grade Point Average, Logistic Regression, Random Forest, Hard voting, Soft voting*

How to Cite: Olive, U., Bosco, M. J., & Enan, N. M. (2025). Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis. *Journal of Information and Technology, 5*(4), 31-40.

1. Introduction

Higher education plays a crucial role in both individual and national development, yet many institutions face challenges related to student retention and graduation (UNESCO, Higher education global data report, 2022). While access to higher education has expanded globally, student dropout rates remain high, posing concerns for workforce planning and social equity (Bank, 2022). Dropout often results from a mix of academic, personal, and institutional issues, such as poor performance, financial struggles, and inadequate support systems. Unfortunately, traditional approaches tend to be reactive and fail to identify at-risk students early enough (Glick et al., 2020). Advancements in artificial intelligence and machine learning offer new strategies for early intervention. Predictive models, especially those using ensemble learning methods like Random Forest, XGBoost, and Gradient Boosting, can analyze complex data to identify students at risk before traditional warning signs appear. These models not only improve prediction accuracy but also allow institutions to understand which factors matter most, through feature importance analysis (Islam et al., 2024). This makes it possible to act early with personalized support.

For example, if a model detects that late tuition payments or poor early academic performance predict dropout, the institution can respond with financial aid or academic counseling (Kemper et al., 2020). This proactive use of data helps target the underlying reasons for disengagement, improving the chances of student success. Despite these benefits, many predictive models are developed using data from high-income countries and may not fit the realities of developing regions. This gap highlights the need to adapt and apply such tools in diverse educational settings. Student dropout stems from a wide range of factors including academic difficulty, economic hardship, health issues, and weak institutional support (Bäulke et al., 2022; Lorenzo-Quiles et al., 2023; Neupane, 2024). Many universities still rely on narrow metrics like GPA or attendance, which are often too late to guide effective intervention (UNESCO, 2023). While machine learning holds great promise, its use remains limited, and few studies focus on the benefits of ensemble models with interpretability features. This research, therefore, aims to develop a predictive model using ensemble learning techniques that not only identifies students at risk of dropping out but also highlights the most influential factors behind their disengagement. The goal is to support timely, targeted, and interpretable interventions that enhance student outcomes across higher learning institutions.

2. Literature Review

Recent studies have explored the use of machine learning models for predicting student dropout, often achieving only moderate accuracy. Research by Assegie et al. (2024) and Gonzalez-Nucamendi et al. (2023) highlighted challenges such as imbalanced datasets and insufficient feature diversity, which affected model performance. Ensemble models like stacking and blending have shown promise, as seen in the work of Islam et al. (2024), yet these approaches still face limitations in predictive power. (Bako et al. (2023 also emphasized the importance of incorporating socio-economic and early academic indicators to improve model accuracy. Overall, while ensemble methods like Random Forests and XGBoost offer potential, further improvement is needed in data handling and model tuning.

Existing machine learning models also have specific drawbacks. Logistic Regression, though interpretable, assumes linear relationships and struggles with high-dimensional or complex data. Random Forests, despite being robust, are computationally heavy and less transparent, with a known bias toward categorical variables. AdaBoost, while effective, is sensitive to noisy data and outliers, and its sequential nature limits scalability for large datasets (Xiao et al., 2018). These limitations suggest that no single model is universally effective and highlight the need for ongoing experimentation.

Despite advancements, key gaps remain in current dropout prediction research. Many studies focus on data from single institutions, limiting the generalizability of their findings. There is also a need for deeper investigation into stacked ensemble models, which have seen success in other fields but are underutilized in educational contexts. Additionally, few studies explore how predictive insights can be translated into timely interventions. Future research should not only aim to improve accuracy but also examine how models can support early warnings and help institutions reduce dropout through proactive strategies.

3. Methodology

3.1 Study Area and Data Sources

Data were sourced from Kaggle and are available in CSV format. The dataset includes 35 columns with both categorical and numerical data. Important features include student demographics (e.g., gender, nationality), academic performance (e.g., curricular units credited), and socio-economic indicators (e.g., parental occupation). This data is used to develop machine learning models for predicting student dropout. The dataset includes 4,424 records and 35 fields, each containing demographic, academic, and socio-economic attributes. Key features include marital status, previous qualifications, parental education and occupation, academic performance indicators, and socio-economic background. The goal is to analyze these factors and predict which students are at risk of dropping out.

3.2 Development technologies

The dropout prediction models were developed using Python, chosen for its versatility in data science and machine learning. Key libraries such as Pandas, Numpy, and Scikit-learn were used for data manipulation, model implementation, and algorithm training. The ensemble approach, using the Voting Classifier, combines predictions from RFC, LR, and ABC to improve model performance and robustness.

3.3 Model Performance Evaluation

Model performance evaluation uses key metrics like the confusion matrix to compare true and predicted results (Chauhan, 2020). Metrics such as accuracy, F1 score, precision, and recall help measure prediction quality (Aditya, 2018). MCC offers a balanced view even with imbalanced data (Chicco et al., 2021), while AUC shows how well the model separates classes (Terra, 2024). Feature importance analysis, using models like Random Forest and AdaBoost, highlights factors like tuition fees and academic performance that influence student dropout.

4. Results and Discussion

4.1 Introduction

The study evaluates machine learning models using metrics like accuracy, precision, recall, F1-score, and MCC, following the previously described methodology and presenting results with tables and graphs. Experiments were carried out on an HP laptop with a 13th Gen Intel Core i7 processor, 16 GB RAM, and Windows 11 Pro. Random Forest, Logistic Regression, and AdaBoost models were used to predict student dropout, and their performance was compared.

4.2 Comparison of Random Forest, Logistic Regression, and AdaBoost Classifiers

Three machine learning models: Logistic Regression, AdaBoost, and Random Forest were evaluated using a dataset of 3,539 student records to predict student dropout, graduation, or continued enrollment. In terms of accuracy, Random Forest performed the best with 79.89%, followed by Logistic Regression at 77.74% and AdaBoost at 75.48%. For the Area Under the Curve (AUC), which measures class separation, Logistic Regression achieved the highest value at 89%, with AdaBoost at 86.73% and Random Forest at 80%. When assessing performance using the Matthews Correlation Coefficient (MCC), Random Forest again led with 70.31% for dropout, 44.43% for graduates, and 74.28% for enrolled students, outperforming both Logistic Regression and AdaBoost in these categories. Although Logistic Regression had the best AUC, indicating strong discriminative power, Random Forest showed the most balanced performance across all metrics, making it the most reliable model for predicting student outcomes. The results suggest that Random Forest is a more robust and accurate model for predicting student dropout as shown in Figures 1,2, and 3.

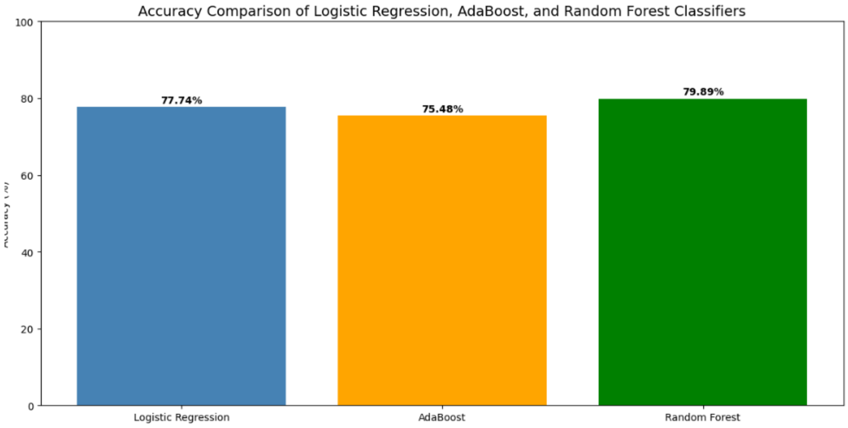

Figure 1: Comparison of individual models based on the accuracy

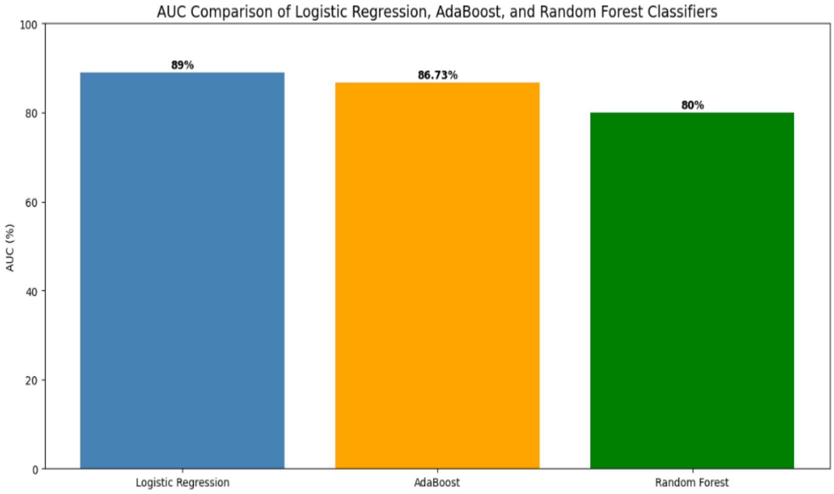

Figure 2: Comparison of individual models based on Area Under the Curve - Receiver Operating Characteristic

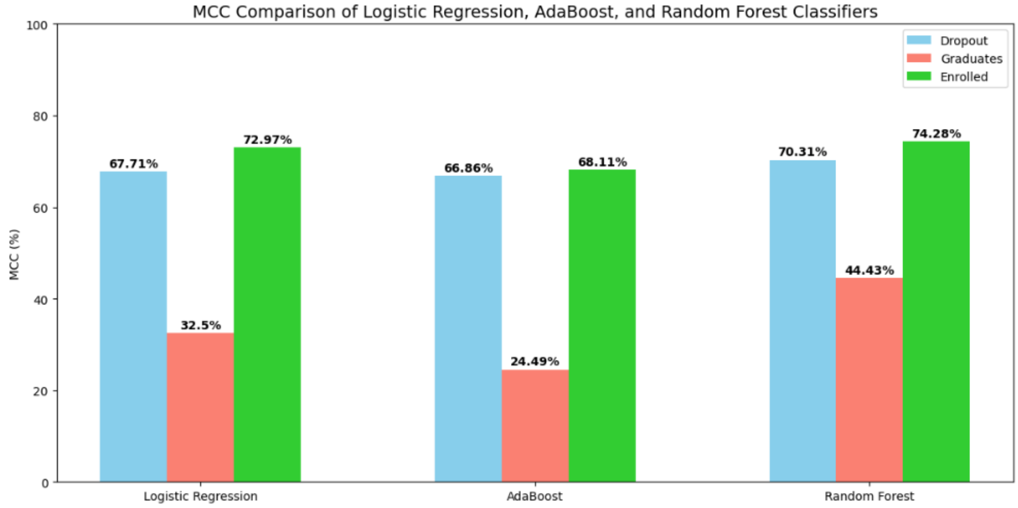

Figure 3: Comparison of individual models based on Matthews Correlation Coefficient (MCC)

4.3 Comparison of Soft and Hard Voting Ensemble Classifiers: Accuracy and MCC for Dropout, Graduate, and Enrolled.

An ensemble model combining Logistic Regression, Random Forest, and AdaBoost with hard voting was applied to predict student dropout based on socio-economic, demographic, and academic factors. Trained on 3,539 student records, the model achieved 79.77% accuracy and varying Matthews Correlation Coefficients (MCC) for each class: 71.48% for dropouts, 40.82% for graduates, and 73.89% for enrolled students. The model performed well in identifying dropouts and enrolled students but struggled with graduates. When compared to Soft Voting, which achieved 80.56% accuracy and higher MCCs (72.13% for dropouts, 44.57% for graduates, and 75.12% for enrolled), Soft Voting showed superior performance, especially in handling multi-class predictions, likely due to its ability to combine classifier probabilities instead of relying on majority voting.

Table 1: Comparison of Soft and Hard Voting Ensemble Classifiers: Accuracy and MCC for Dropout, Graduate, and Enrolled

Model	Accuracy (%)	AUC (%)	MCC (Dropout) (%)	MCC (Graduate) (%)	MCC (Enrolled) (%)
Random Forest	79.89	80	70.31	44.43	74.28
Logistic Regression	77.74	89	67.71	32.50	72.97
AdaBoost	75.48	86.73	66.86	24.49	68.11
Hard Voting	79.77	—	71.48	40.82	73.89
Soft Voting	***80.56***	***91***	***72.13***	***44.57***	***75.12***

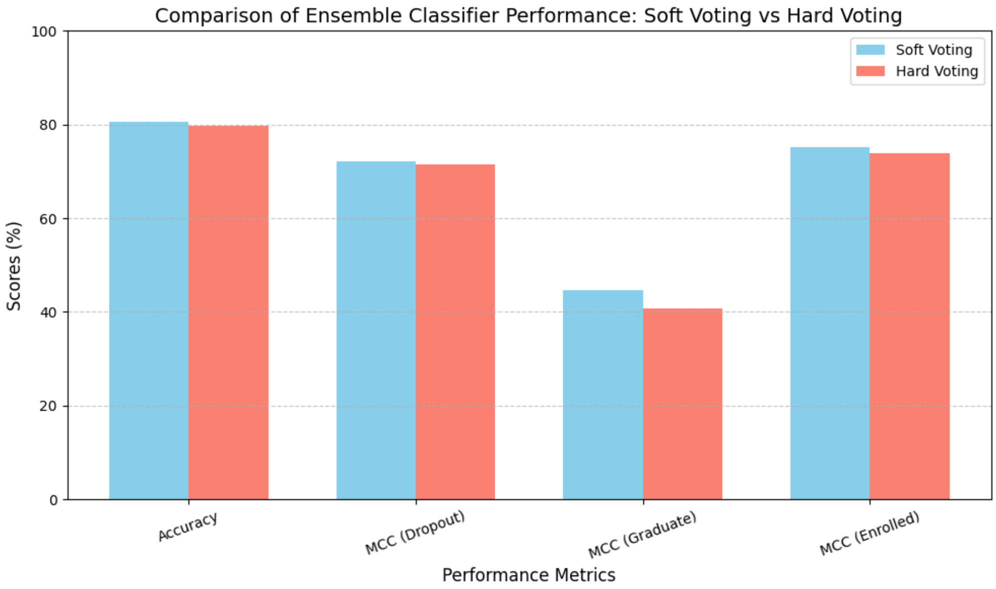

Figure 4: Comparison of Soft and Hard Voting Ensemble Classifiers: Accuracy and MCC for Dropout, Graduate, and Enrolled.

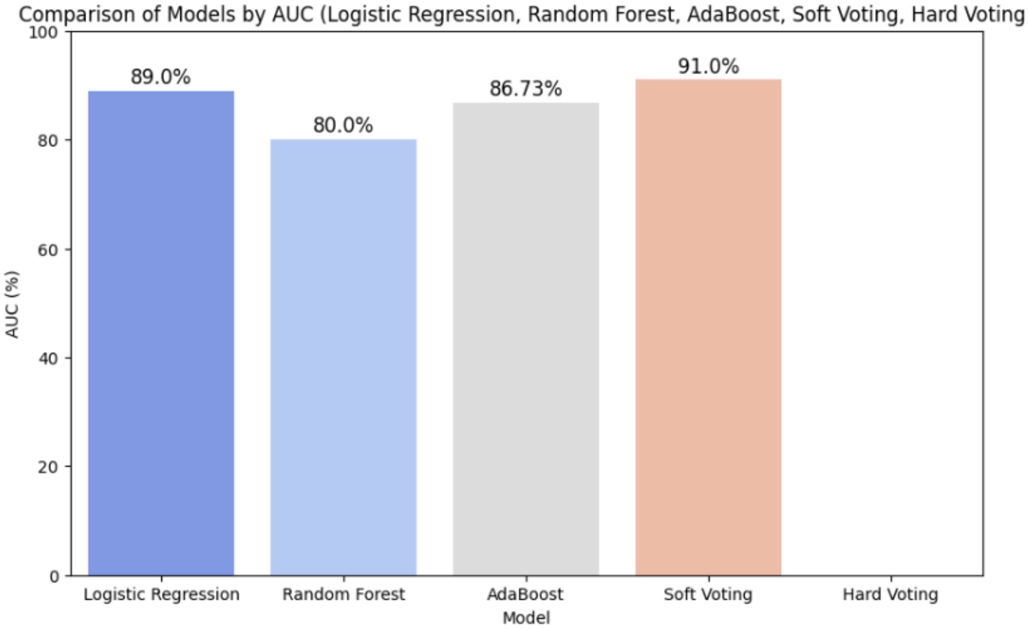

Figure 5: Comparison of Models by AUC: Logistic Regression, Random Forest, AdaBoost, Soft Voting, Hard Voting

4.4 Feature Importance Analysis

The donut chart reveals the distribution of students based on their academic status, showing that the majority (2,209) graduated, while a significant number (1,421) dropped out, highlighting a notable dropout issue. The remaining 794 students are still enrolled and are at risk of either graduating or dropping out in the future. The bar chart identifies the top 10 features influencing student dropout, with academic performance being the most significant predictor, particularly the number of curricular units approved and grades obtained in both semesters. Non-academic factors, such as age, father's occupation, tuition fee status, and course of study, also contribute to dropout risk, indicating the need for interventions that address both academic and socio-economic factors. In analyzing student academic performance through a correlation heatmap, factors like marital status, application order, and daytime attendance show weak to moderate correlations with academic performance. Prior qualifications, parental education/occupation, and financial factors also exhibit weak to moderate positive correlations. Stronger positive correlations are seen with curricular factors like credits and grades. Demographic and macroeconomic variables have minimal impact.

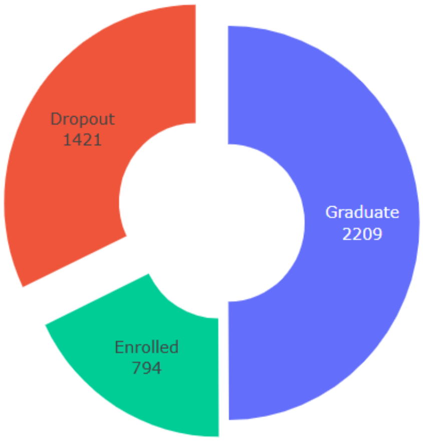

Figure 6: Distribution of students based on their academic status: graduates, dropouts, and currently enrolled

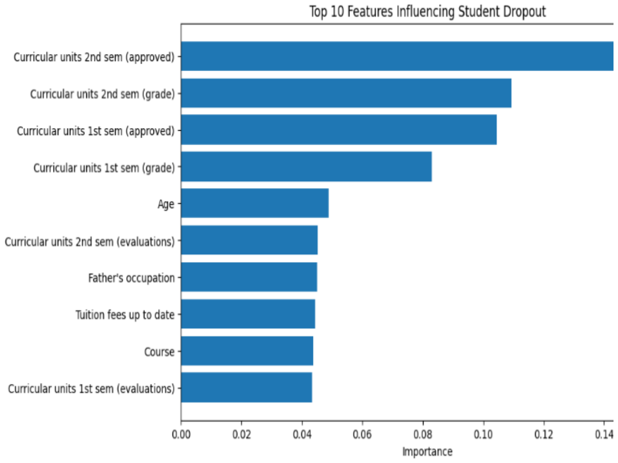

Figure 7: The top 10 features that most significantly influence student dropout

5. Conclusion

This research developed a predictive model using ensemble learning techniques to identify students at risk of dropping out of higher learning institutions (HLIs). By combining models such as Random Forest, Logistic Regression, and AdaBoost through soft and hard voting ensembles, the study demonstrated that ensemble methods enhance both prediction accuracy and model robustness. The Random Forest model outperformed others in terms of accuracy, while Logistic Regression exhibited superior discriminative power, as evidenced by its high Area Under the Curve (AUC). The ensemble model, combining these classifiers using soft voting, achieved the highest accuracy (80.56%) and AUC (91%), surpassing individual models. The feature importance analysis revealed key factors influencing dropout decisions, including academic performance, financial issues, and demographic factors, offering institutions insights to implement targeted interventions for at-risk students. However, the research also highlighted gaps in the existing literature, particularly the lack of larger, more generalized datasets and the underexplored potential of advanced ensemble techniques like stacked ensembles in dropout prediction, suggesting a promising area for future research.

6. Recommendations

Future research on dropout prediction models should focus on improving predictive accuracy by exploring advanced ensemble techniques and hybrid models. While current models have provided valuable insights, there is room for enhancing their performance, especially in multi-class classification scenarios. Researchers should aim to fine-tune models for better generalizability across diverse datasets, particularly those that reflect the varied educational contexts of different institutions. Furthermore, collecting localized datasets is essential to enhance the relevance and applicability of dropout prediction models. Many existing models are trained on data from developed countries, which may not fully represent the socio-economic, cultural, and institutional dynamics of educational systems in developing regions. By gathering data from a range of institutions, particularly those in underrepresented areas, future research can ensure that predictive models are more globally applicable and accurate. This approach will provide a deeper understanding of the specific factors affecting student retention, ultimately improving the accuracy and actionability of dropout prediction models in diverse educational environments.

References

i. Aditya, M. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. Retrieved March 30, 2025, from https://towardsdatascience.com/: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

ii. Bank, W. (2022). *The State of Global Learning Poverty: 2022 Update*. Retrieved April 25, 2025, from https://www.worldbank.org: https://www.worldbank.org/en/topic/education/publication/state-of-global-learning-poverty

iii. Chauhan, N. S. (2020, May 28). *Model Evaluation Metrics in Machine Learning*. Retrieved March 28, 2025, from https://www.kdnuggets.com/: https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html

iv. Terra, J. (2024, Aug 13). *What is a ROC Curve, and How Do You Use It in Performance Modeling?* Retrieved from https://www.simplilearn.com/: https://www.simplilearn.com/what-is-a-roc-curve-and-how-to-use-it-in-performance-modeling-article

v. UNESCO. (2022). *Higher education global data report*. Retrieved April 24, 2025, from https://unesdoc.unesco.org/: https://unesdoc.unesco.org/ark:/48223/pf0000389859

vi. UNESCO. (2023, September). *Education Data Release 2023*. Retrieved March 26, 2025, from https://uis.unesco.org/: https://uis.unesco.org/en/news/education-data-release

vii. Assegie, T. A., Salau, A. O., Chhabra, G., Kaushik, K., & Braide, S. L. (2024). Evaluation of Random Forest and Support Vector Machine Models in Educational Data Mining. *2024 2nd International Conference on Advancement in Computation &amp; Computer Technologies (InCACCT)*, 131–135. https://doi.org/10.1109/InCACCT61598.2024.10551110

viii. Bako, H. S., Ambursa, F. U., Galadanci, B. S., & Garba, M. (2023). PREDICTING TIMELY GRADUATION OF POSTGRADUATE STUDENTS USING RANDOM FORESTS ENSEMBLE METHOD. *FUDMA JOURNAL OF SCIENCES*, *7*(3), 177–185. https://doi.org/10.33003/fjs-2023-0703-1773

ix. Bäulke, L., Grunschel, C., & Dresel, M. (2022). Student dropout at university: A phase-orientated view on quitting studies and changing majors. *European Journal of Psychology of Education*, *37*(3), 853–876. https://doi.org/10.1007/s10212-021-00557-x

x. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen’s Kappa and Brier Score in Binary

Classification Assessment. *IEEE Access*, *9*, 78368–78381. https://doi.org/10.1109/ACCESS.2021.3084050

xi. Glick, D., Cohen, A., & Chang, C. (Eds.). (2020). *Early Warning Systems and Targeted Interventions for Student Success in Online Courses:* IGI Global. https://doi.org/10.4018/978-1-7998-5074-8

xii. Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, *8*, 1244686. https://doi.org/10.3389/feduc.2023.1244686

xiii. Islam, M., Islam, M. M., Ali, Md. S., Niloy, N. T., Chowdhury, A., & Avik, S. C. (2024). Ensemble Method for Predicting Student Performance and Dropout Risk. In J. K. Mandal, M. Hinchey, & S. Chakrabarti (Eds.), *Recent Advances in Artificial Intelligence and Smart Applications* (pp. 269–278). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-3485-6_21

xiv. Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

xv. Lorenzo-Quiles, O., Galdón-López, S., & Lendínez-Turón, A. (2023). Factors contributing to university dropout: A review. *Frontiers in Education*, *8*, 1159864. https://doi.org/10.3389/feduc.2023.1159864

xvi. Neupane, B. (2024). Causes of Dropout in Higher Education: An Analysis of Student Dropouts in Bachelor of Education from Marsyangdi Multiple Campus. *Marsyangdi Journal*, 1–14. https://doi.org/10.3126/mj.v4i1.67750

xvii. Xiao, J., Li, Y., Xie, L., Liu, D., & Huang, J. (2018). A hybrid model based on selective ensemble for energy consumption forecasting in China. *Energy*, *159*, 534–546. https://doi.org/10.1016/j.energy.2018.06.161