© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715880

Predictive Model for Student Dropout Rates Using Machine Learning Techniques

NWAJIOBI KOSISO PRECIOUS[1], RIDWAN KOLAPO[2], MUHAMMAD IBRAHIM NURUDEEN[3], TEMITOPE ATOYEBI[4], PREMA KIRUBAKARAN[5]

[1, 2, 4, 5]*Information Technology Department, Nile University of Nigeria.*

[3]*Cyber Security Department. Nile University of Nigeria.*

Abstract- ***Student dropout remains a persistent challenge in higher education, particularly in developing countries where institutional support systems are often limited. This study develops a machine learning based predictive model for early identification of students at risk of dropping out within Nigerian universities. A quantitative research design was adopted using institutional data comprising academic performance, demographic characteristics, and behavioural indicators. The dataset was pre-processed through cleaning, encoding, and feature selection, and subsequently divided into training and testing subsets.Multiple classification algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors were implemented and evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The results indicate that ensemble and kernel-based methods outperform traditional linear models, achieving accuracy levels exceeding 80% while significantly improving recall in identifying at-risk students. Key predictors of dropout include academic performance trends, attendance patterns, and student engagement indicators. The findings demonstrate the effectiveness of machine learning techniques in enabling early detection of student attrition risk. The study recommends the integration of interpretable predictive models into institutional information systems to support timely intervention strategies. Furthermore, it highlights the need for robust data governance frameworks to ensure ethical and sustainable deployment of predictive analytics in higher education.***

Index Terms- ***Students, Predictive, Dropout.***

I. INTRODUCTION

Student dropout has become a critical concern for higher education institutions worldwide, with far-reaching implications for individuals, institutions, and national development. There has been an increasing interest in big data applications in education amongst the researchers who wish to take part in predictive modelling as touching student dropout rates. [1] envisioned that between 2016 and 2021, there has been a gradual rise in EDM publications, and they attribute this, in their opinion, to the stimulating effects of technological advancements and policy changes towards evidence-based education. [2] have similarly pointed out that there has been a major trend moving towards data-driven education, where decisions involve teaching, learning, and institutional planning, all being mediated by analytics. The most enthusiastic optimism, however, has also been challenged. Potentially, [3] argues that such a system-wide coordination of data collection may give rise to privacy and ethical issues since students may remain oblivious about what happens to their data. Thus, predictive analytics requires efficacy evaluation with data protection and informed consent. If done properly, data-informed education can make internal reforms in these institutions expedient, like the early detection of students with academic risk.

In Nigerian universities, students often struggle with complex concepts, especially fresh undergraduates navigating their academic journey. However, most courses taught are designed in theories and practicals as the primary medium of instruction, there is a need for additional strategies to enhance students' learning experiences[4]. Still, predictive models provide real benefit in empirical research into retention for current actual students. [5] longitudinally studied the undergraduate student outcome from 2018 to 2021 using a deep learning model. The dropout risk has been predicted in them with an accuracy level of over 90%, which surpassed the traditional linear classifiers. An example by [6] is that of a hybrid ensemble framework for predictive modelling successfully implemented in the Chilean

generalization school systems, which proved that indeed local machine learning models could be applied. Their approach was limited in generalization, as the principle would be institutional datasets with very few behavioral attributes; thus, generalization would be narrowed. [7] argued that it should also include emotional engagement and motivation for the models because these two aspects are just as important regarding retention as is academic performance, thus bringing in the whole debate of involving other methodological approaches in the debate regarding the synthesized quantitative versus qualitative understanding in predictive research. The views above can be synthesized as the best recipe for an ideal dropout prediction system: cognitive, behavioral, and demographic indicators through a transparent analytical framework.

These have produced advances in the field of machine learning in education as well as innovations in algorithms that have emerged. Among the dropout predictors adopted in the emerging markets include those that are tree-based, random forests, logistic regression, and most popularly, neural networks [8]

II. LITERATURE REVIEW

Currently, dropout prediction from institutions tends to suffer from imitation of empirical data and requires performance improvement, as there is a high dependence and also ongoing development in machine learning to build interventions against the limitations of traditional statistical methods. Dropout is an occurrence that has many complexities across different educational contexts, with various causes and factors from the academic, demographic, behavioral, and socioeconomic domains. Initial attempts at crafting methods have been more descriptive and regression-focused; however, new trending enquiries are focusing on supervised machine learning techniques in order to foster high accuracy in prediction and timely interception of high-risk students [9]. This is by far quite representative of larger developments in educational data mining and learning analytics, as the volume of data institutions have collected on them allows for much more sophisticated modelling of student trajectories.

By characterizing the most commonly used classification methods, hardly any of the survey studies are found to have conveyed the use of any forms of decision trees, logistic regression, support vector machines, random forests, or neural networks. Random forests, especially, enjoy prominence in performance prediction, as they grapple with any non-linear relations in features. Probably all of these were conducted in Asia, Africa, Europe, and South America, and they reported accuracy of Random Forest around 71 to more than 96 per cent, depending on data quality and feature selection strategy [10]. Other ensemble methods such as gradient boosting and XGBoost might perform excellently, but there is something above 95 per cent in accuracy with regard particularly to being fitted against feature engineering and sampling techniques with a focus on class imbalance [11].

Measurable, widely applicable, and the simplest to apply, especially in institutional environments where transparency has become mainstream, logistic regression is still the most widely applicable method. Its predictive performance is low against the ensemble methods and neural-orientated models, but some authors have cited studies confirming that for early dropout detection, recall and F1-score reporting can be competitive [12]. In this regard, support vector machines have also been said to be studied, and empirical studies have shown that they perform quite decently in high- dimensional spaces; however, they often receive criticism for their sensitivity to parameter tuning and scalability issues [13].

In fact, through that neural networking and teaching probe far advanced in adaptation or investigation on large datasets of studies about different countries, these models always seem to outperform traditional classifiers on counts of accuracy and recall regarding dropout predictions at the school and university level [14]. The performance measures, however, have never been equal to quell most authors' reservations on inferability, computational cost, and willingness of institutions to roll out such great models [15]. Hence, it is much more open in the literature to crave for a compromise in selecting models against predictive strength and operational ease.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715880

III. METHODOLOGY

The data for this research originated from the Kaggle repository for Higher Education Predictors of Student Retention in an undergraduate student sample across a variety of indicators of institutional persistence. This dataset consists of 4424 student record cases with 35 attributes describing various academic, demographic, and socioeconomic features, including marital status, nationality, parental qualifications, previous qualifications, etc. These variables collectively represent a picture of a student's profile with respect to dropout scenarios in a mixed-type dataset through numerical values for some of the quantitative measures (for example, age) and categorical codes for classifications (for example, application mode and status), and also signal the expected outcome of student retention in predictive models in effect, a binary outcome.

Table 1: Objective vs Methodology Matching

Objectives	Proposed Methodology	Expected Results
Objective (1)	Data collection and preprocessing	Clean, reliable dataset
Objective (2)	Model design and training (Random Forest, XGBoost, etc.)	Predictive model for dropout
Objective (3)	Model evaluation and comparison	Best-performing

Table 1 provides a structured mapping of the research objectives to their respective methodologies and expected results. The progression from data collection and pre-processing to model development and finally to evaluation and comparison, reflects a systematic approach designed to produce a robust predictive model and identify the most effective technique.

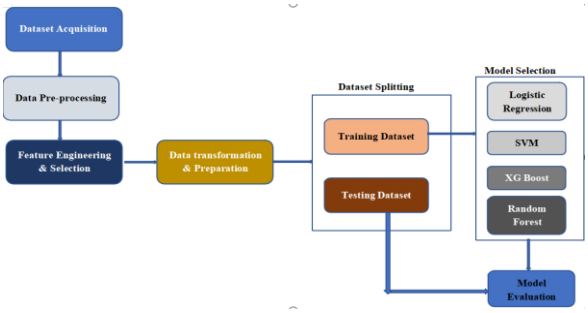

Figure 1: Student Dropout Rates Model Predictive Framework

Figure 1 illustrates the overall research workflow adopted in this study. The process begins with dataset acquisition and pre-processing, followed by feature engineering and data transformation. The dataset is then split into training and testing sets, after which multiple machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), XGBoost, and Random Forest, are applied. Finally, model evaluation is conducted to determine the best-performing predictive model.

IV. RESULTS AND DISCUSSIONS

The dataset used in this study comprised 3,630 student records after pre-processing, with missing values effectively handled to ensure completeness and consistency. The features included academic, demographic, and financial variables, providing a comprehensive representation of student profiles.

Descriptive analysis indicated that most students were within the traditional university entry age, with a mean age of approximately 23 years. Academic performance variables, particularly first- and second-semester grades and approved curricular units, exhibited relatively high variability, reflecting differences in student performance levels. Financial indicators such as tuition fee status and scholarship availability also showed meaningful variation across the dataset.

The target variable distribution revealed a moderate class imbalance, with approximately 60% graduates and 40% dropouts. To address potential bias in model training, data balancing techniques (e.g., SMOTE) were applied during preprocessing, ensuring that both classes were adequately represented.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715880

Model Performance Evaluation

A summation of Table 2: includes the performances evaluation of the four machine classifiers Logistic Regression, Random Forest, SVM, and XGBoost, through standard metrics. Thus, these de-limitations address Objective(i) and (ii). Accuracy values of 0.938 for Logistic Regression and SVM indicate absolute classification capability almost universally in favour of acceptance, whereas Logistic Regression exhibited good balance performance-wise between precision (0.919) and recall (0.923) with the highest F1-Score (0.921) , thus meaning it was good at catching dropouts but not mis-classifying graduates. Random Forest was less precise but more informative, whereas XGBoost had a slight difference in accuracy retaining its recall and ROC-AUC values, suggesting that it performed quite well in discriminating between the groups. This set of results altogether verified across various modelling experiments for the student dropout prediction while trading off caution against accepting specificity.

Table 2: Performance results of each machine learning model

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.938017	0.919298	0.922535	0.920914	0.973193
Random Forest	0.92562	0.922794	0.883803	0.902878	0.971054
SVM	0.938017	0.947566	0.890845	0.91833	0.969704
XGBoost	0.910468	0.87372	0.901408	0.887348	0.970676

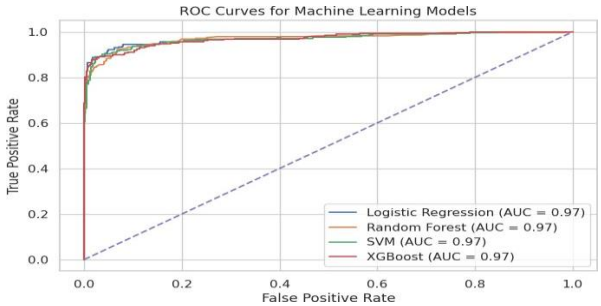

Figure 2: ROC Curves for Predictive Models

CONCLUSION

The findings revealed that academic, demographic, and behavioural indicators may be overlapping with performance progress variables, which still proved to be the strongest determinants of dropout. Here, dropout is viewed as termination from study before fulfilling all requirements for program completion, while retention is continuation of enrolment till degrees earned by students. Quite high correlations between dropout rate and performance in the first and the second semester seem to imply that learning might provide some early participation proxy measurement of adjustment by students to the institution's demands. In most other studies, however, bases on these demographic characters as prime determinants, especially age and gender. This limited thinking seems to rather be transferred across contexts as indicated by findings within the present study that demographic factors are bogus single predictors through which indirect effects between academic outcomes.

Under the complex models and algorithms emerged not much of difference in the results emanating from performances with respect to accuracy scores. The logistic regression and SVM were indeed on par in terms of accuracies, as well as the ROC-AUC as shown in figure 2 above, with those from ensemble methods; thus defying the myth that strong predictive performance results from more complex algorithms, contrary to claims that dropout prediction gets a much better yield through gradient boosting or even by deep learning strategies.

REFERENCES

[1] C. Baek and T. Doleck, “Educational data mining: A bibliometric analysis of an emerging field,” *IEEE Access*, vol. 10, pp. 31289–31296, 2022.

[2] D. D. Figaredo, J. Reich, and J. A. Ruipérez-Valiente, “Learning analytics and data-driven education: A growing field,” *Revista Iberoamericana de Educación a Distancia*, vol. 23, no. 2, pp. 33–39, 2020.

[3] I. Potgieter, “Privacy concerns in educational data mining and learning analytics,” *The International Review of Information Ethics*, vol. 28, 2020.

[4] T. O. Atoyebi, J. O. Olayiwola, A. N. Eru, R. Kolapo, and P. Kirubakaran, "Gamification: An educational strategy to increase students' motivation and academic performance," *FUDMA Journal of Sciences*, vol. 9, no. 12, pp. 36–41, 2025, doi: 10.33003/fjs-2025-0912-4164.

[5] Y. T. Shiao *et al.*, "Reducing dropout rate through a deep learning model for sustainable education: Long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021," *Smart Learning Environments*, vol. 10, no. 1, p. 55, 2023.

[6] P. Rodríguez, A. Villanueva, L. Dombrovskaia, and J. P. Valenzuela, "A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: The case of Chile," *Education and Information Technologies*, vol. 28, no. 8, pp. 10103–10149, 2023.

[7] S. A. Sulak and N. Koklu, "Predicting student dropout using machine learning algorithms," *Intelligent Methods in Engineering Sciences*, vol. 3, no. 3, pp. 91–98, 2024.

[8] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Sciences*, vol. 10, no. 3, p. 1042, 2020.

[9] L. Kemper, G. Vorhoff, and B. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, pp. 28–47, 2020, doi: 10.1080/21568235.2020.1718520.

[10] D. K. Dake and C. Buabeng-Andoh, "Using machine learning techniques to predict learner dropout rate in higher educational institutions," *Mobile Information Systems*, vol. 2022, Art. no. 2670562, 2022.

[11] T. Akter *et al.*, "Dropout prediction of university students in Bangladesh using machine learning," Sep. 25, 2024, doi: 10.1109/COMPAS60761.2024.10797033.

[12] E. E. Osemwegie, F. Amadin, and O. M. Uduehi, "Student dropout prediction using machine learning," *FUDMA Journal of Sciences*, Dec. 31, 2023, doi: 10.33003/fjs-2023-0706-2103.

[13] M. A. Dewi *et al.*, "Machine learning algorithms for early predicting dropout student online learning," Nov. 7, 2023, doi: 10.1109/ICCED60214.2023.10425359.

[14] S. Anwar, "Predicting school dropout risk using machine learning models: A comparative study of random forest, gradient boosting, and neural network," *Jurnal Ekonomi, Teknologi dan Bisnis*, Jul. 25, 2025, doi: 10.57185/jetbis.v4i6.193.

[15] J.-P. A. Barthès, "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, Jul. 1, 2023, doi: 10.1016/j.eswa.2023.120933.