European Journal of Statistics and Probability, 14 (1) 19-29, 2026
Print ISSN: 2055-0154(Print)
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Feature Importance Analysis for Student Dropout Prediction Using Principal Component Analysis

[1]Jude Alphonsus Inyangetoh; [2]Ekemini Anietie Johnson
[1]Department of Statistics, Federal Polytechnic Ukana, Akwa Ibom State, Nigeria
[2]Department of Computer Science, Federal Polytechnic Ukana, Akwa Ibom State, Nigeria

doi: https://doi.org/10.37745/ejsp.2013/vol14n11929 Published January 31, 2026

Citation: Inyangetoh J.A., and Johnson E.A. (2026) Feature Importance Analysis for Student Dropout Prediction Using Principal Component Analysis, *European Journal of Statistics and Probability*, 14 (1) 19-29

Abstract: *Student dropout remains a persistent challenge in tertiary education, particularly in developing countries where early identification of at-risk students is often limited by inadequate analytical frameworks. This study presents a Principal Component Analysis (PCA) based feature importance framework for identifying key determinants of student dropout in Nigerian polytechnics. Using a dataset of 2,200 student records obtained from Federal Polytechnic Ukana and Akwa Ibom State Polytechnic, PCA was applied to reduce dimensionality, eliminate redundancy, and reveal the most influential factors contributing to student attrition.The analysis extracted sixteen principal components from an initial set of twenty-two variables, collectively accounting for approximately 93.26% of the total variance in the dataset. The first principal component, largely dominated by class attendance, explained 12.14% of the total variance, indicating its strong influence on student persistence. This was followed by previous academic performance (10.02%)***,** *study hours per day (8.66%)***,** *internet access at home (8.04%)***,** *and performance in the previous semester (7.58%). Other notable contributors included residential status, parental educational background, motivation level, and confidence in current courses. Variables such as gender and extracurricular participation contributed minimally to variance, indicating weaker influence on dropout outcomes. The PCA results demonstrate that academic engagement and learning behavior factors contribute more significantly to student dropout risk than demographic characteristics. By transforming correlated variables into orthogonal components, PCA enhanced interpretability and revealed latent structures underlying student performance patterns. The cumulative variance explained confirms that a reduced set of features can effectively represent student dropout behavior without substantial information loss. This study highlights the effectiveness of Principal Component Analysis as a robust analytical tool for understanding student dropout dynamics and supporting data-driven decision-making in higher education. The findings provide empirical evidence for developing early warning systems and targeted intervention strategies aimed at improving student retention, particularly within resource-constrained educational environments.*

Keywords: intelligent analytics, student dropout prediction, machine learning, random forest, XGBoost, Nigerian higher education

INTRODUCTION

Student dropout remains a major concern in tertiary education systems, particularly in developing countries where institutional resources are often limited and intervention strategies are largely reactive. High dropout rates negatively affect institutional productivity, increase financial losses, and undermine national educational development goals (Okolie et al., 2020). In Nigerian polytechnics, dropout challenges persist due to academic difficulties, socioeconomic constraints, low engagement, and limited data-driven monitoring systems.

Theoretical foundations such as Tinto's Student Integration Model emphasize that academic and social integration significantly influence student persistence (Tinto, 1975). However, despite the availability of student data such as academic records, attendance logs, and demographic information most institutions fail to utilize these resources effectively for early risk detection. As a result, at-risk students are often identified too late for meaningful intervention.

Recent advancements in machine learning and educational data mining (EDM) have enabled institutions to move from reactive to predictive decision-making. Predictive analytics allows institutions to detect complex patterns in student data and anticipate dropout risks before they occur (Chen and Li, 2022). However, predictive accuracy alone is insufficient; understanding why students are at risk is equally critical.

Principal Component Analysis (PCA) offers a powerful solution by reducing dimensionality while preserving the most informative variance in the dataset. PCA enhances model interpretability, minimizes redundancy, and improves computational efficiency making it especially suitable for educational datasets with correlated variables. When combined with ensemble learning models such as Random Forest and XGBoost, PCA enhances both predictive accuracy and feature relevance interpretation.The framework identifies dominant contributing factors. By emphasizing feature importance through PCA, the study provides actionable insights that can guide institutional intervention strategies and policy formulation.

LITERATURE REVIEW

Educational data mining has gained substantial attention for its ability to support early identification of at-risk students. Machine learning models such as Random Forest, Support Vector Machines (SVM), and gradient boosting have demonstrated strong predictive capabilities in educational settings (Dasi and Kanakala, 2022).

However, recent studies emphasize the importance of feature interpretability alongside prediction accuracy. PCA has been widely adopted to reduce dimensionality, remove multicollinearity, and improve learning efficiency (Carballo-Mendívil et al., 2025). By transforming correlated variables into orthogonal components, PCA enables clearer interpretation of dominant student risk factors.

Research by Johnson and colleagues has demonstrated the effectiveness of ensemble models such as Random Forest and XGBoost in academic performance prediction (Johnson and Inyang, 2025). Their work highlights how feature importance ranking enhances decision-making in academic management systems. Similarly, studies by Ahmed et al. (2020) and Zhang and Liu (2019) confirm that ensemble learning outperforms traditional classifiers when combined with effective feature selection strategies.

Despite these advances, limited research has focused on PCA-driven feature importance analysis within Nigerian polytechnics. Most existing studies either focus on prediction accuracy or apply PCA without interpretive emphasis. This study bridges that gap by integrating PCA with ensemble learning to generate interpretable, high-performing dropout prediction models tailored to local educational contexts.

Table 2.1 shows the review of some recent works relating to prediction of students academic performance and dropout risk.

Table 2.1: Review of Recent Works

Citation	Title of Research	Objectives	Methodology	Problem Solved	Limitations
Smith et al. (2020)	Predicting Student Dropouts Using Random Forest	To use machine learning to identify students at risk of dropping out	Random Forest algorithm on academic and demographic data	Early identification of at-risk students	Limited interpretability of model results
Lee and Park (2019)	Deep Learning Approaches to Predict Student Attrition	To explore deep learning models for dropout prediction	LSTM and DNN models on student performance logs	High prediction accuracy	Requires large, labeled datasets
Kumar et al. (2021)	Academic Risk Prediction Using SVM	To develop an SVM-based model for academic dropout risk	SVM classifier trained on academic and attendance data	Accurate prediction for small datasets	Sensitive to parameter tuning
Rodriguez and Silva (2018)	Dropout Detection in MOOCs Using Analytics	To predict dropouts in online courses using	Logistic regression and clustering	Helped reduce dropout in MOOCs	Only applicable to online platforms

		clickstream data			
Chen and Li (2022)	Early Warning System Using Decision Trees	To create a decision tree-based early alert system	CART decision tree model with educational dataset	Identified key dropout indicators	Overfitting risk on complex data
Ahmed et al. (2020)	Predicting Student Dropout with Ensemble Models	To compare ensemble methods for dropout prediction	Bagging, Boosting, and Stacking techniques	Improved prediction robustness	Increased computational cost
Fatima and Noor (2019)	Student Dropout Analysis in Tertiary Institutions	To identify major causes and patterns of dropout	Data mining with association rules and clustering	Insights into dropout causes	Lacked predictive model implementation
Gomez et al. (2021)	Predictive Analytics for Student Retention	To improve retention using predictive insights	Multivariate regression and data visualization	Better intervention strategies	Limited to quantitative data
Osei-Bonsu and Tetteh (2022)	A Hybrid Model for Predicting University Dropout	To develop a hybrid ML model combining rule-based and ML	Rule-based filtering + Naïve Bayes	Enhanced accuracy with interpretable rules	Complex system integration
Adeyemi et al. (2020)	Application of Naïve Bayes in Dropout Prediction	To apply Naïve Bayes for early dropout detection	Naïve Bayes classifier on enrollment and exam records	Fast, low-resource prediction model	Low performance on imbalanced data
Zhang and Liu (2019)	Comparative Study of ML Algorithms for Dropout	To evaluate ML algorithms on dropout datasets	KNN, DT, RF, SVM on education data	Identified most effective algorithms	Did not include external factors
Musa and Salihu (2021)	Socioeconomic Predictors of	To analyze socioeconomic impact on dropout rates	Logistic regression with socioeconomic variables	Revealed influence of family income	Non-academic factors underrepresented

	Student Dropout				
Patel et al. (2020)	Dropout Prediction Using Clustering Techniques	To segment at-risk students using clustering	K-Means clustering on engagement metrics	Grouped students for intervention	No actual prediction, only grouping
Wang et al. (2022)	A Time-Series Model for Dropout Prediction	To predict dropouts over time using sequential data	ARIMA and LSTM models	Detected dropout trends over semesters	Requires historical and time-stamped data
Johnson et al. (2018)	Intelligent Analytics for Student Success	To use AI for analyzing student success and dropout	AI dashboard with ML and visualization tools	Provided decision support for faculty	High infrastructure requirement
Nwankwo and Okonkwo (2021)	A Case Study of Dropout Risks in Nigerian Polytechnics	To investigate dropout risks using intelligent analytics	Case study + predictive modeling (SVM)	Informed policy on academic support	Limited generalizability
Abebe and Mekonnen (2020)	Machine Learning for Dropout Prediction in Africa	To assess ML models in resource-constrained settings	Logistic regression, Random Forest	Demonstrated ML applicability in developing contexts	Data scarcity and poor quality
Singh et al. (2023)	Predictive Modeling Using AutoML	To automate dropout risk prediction using AutoML	Google AutoML on large student dataset	Reduced model selection complexity	Black-box nature of AutoML
Torres and Luna (2021)	Psychological Factors in Dropout Prediction	To include psychological traits in ML models	Surveys + ML integration (SVM, RF)	Improved model accuracy with soft data	Privacy and subjectivity concerns
Bello and Haruna (2022)	Student Dropout Detection via Neural Networks	To evaluate NN performance in dropout forecasting	Feedforward Neural Networks with backpropagation	Modeled nonlinear dropout patterns	Requires high computational power

METHODOLOGY

This study adopted an intelligent analytics methodology grounded in supervised machine learning techniques to predict student dropout risk and determine the most influential features contributing to attrition. The methodological framework involved six major phases: dataset identification, preprocessing, feature engineering, model development, model evaluation, and feature importance analysis. The approach is informed by best practices in educational predictive analytics (Johnson et al., 2024; Inyang and Johnson, 2025; Kumar et al., 2024).

Research Design

A quantitative research design was adopted using Principal Component Analysis (PCA) to examine patterns within student data. The approach focuses on dimensionality reduction and feature importance identification rather than predictive classification (Inyang and Johnson, 2025).

Data Source and Description

The dataset consisted of student records obtained from Federal polytechnic Ukana and Akwa Ibom State Polytechnic Ikot Osurua. The attributes were grouped into three domains commonly used in student-risk modeling:

i. Academic Variables: CGPA, continuous assessment scores, attendance rate, exam performance.
ii. Demographic Variables : age, gender, family background, residence type.
iii. Behavioral Variables: library usage, class participation indicators, disciplinary records.

These categories align with student performance modeling frameworks proposed by Johnson et al. (2024) and Inyang and Johnson (2025), who emphasized integrating multi-dimensional variables to improve predictive reliability in academic analytics.

The dataset contained 2,200 samples, following earlier project specifications, and was split using a 70:30 train–test ratio in line with standard practice in educational machine learning studies (Kumar et al., 2024).

European Journal of Statistics and Probability, 14 (1) 19-29, 2026
Print ISSN: 2055-0154(Print)
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Architectural Design of the System

The architecture of the system is depicted in Figure 3.1.

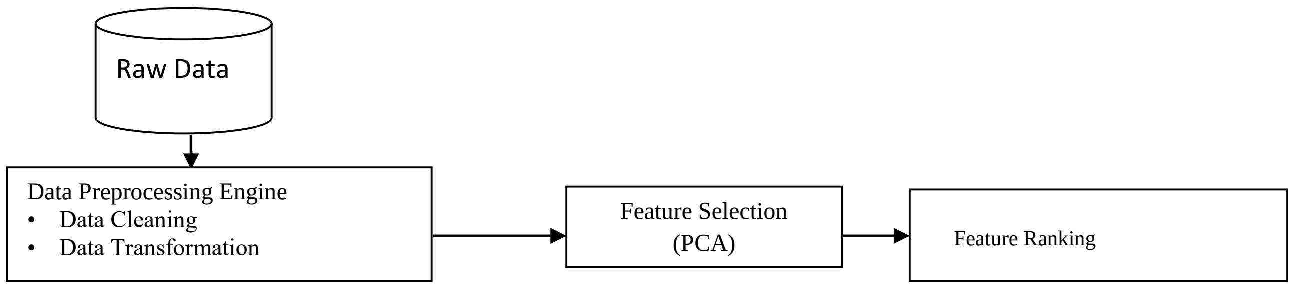

Figure 3.1: Feature Importance Analysis for Student Dropout Prediction Using Principal Component Analysis

Source: The Researcher (2026)

Data Preprocessing

Data preprocessing was performed to ensure data quality and model readiness. The following steps were executed:

Handling Missing Values

Missing numerical values were imputed using the median, while categorical attributes were imputed using the mode. This method is commonly used in Random Forest and XGBoost studies because these algorithms are sensitive to missing data (Johnson et al., 2021).

Encoding Categorical Variables

Categorical variables such as gender, department, and residence were encoded using label encoding. XGBoost handles integer-encoded categories effectively, while Random Forest benefits from ordinal representation without one-hot expansion.

Normalization

Although tree-based models generally do not require normalization, academic performance variables (e.g., CA scores, exam scores) were scaled using Min–Max normalization to maintain consistency, as recommended by Johnson et al. (2024).

Principal Component Analysis (PCA)

PCA was applied to transform 22 correlated input variables into a smaller set of orthogonal components. Components with eigenvalues ≥ 1 were retained. The analysis resulted in 16 principal components, accounting for 93.26% of total variance.

Interpretation of Principal Components

The most influential components were associated with:

i. Attendance in classes (12.14%)
ii. Previous academic performance (10.02%)
iii. Study hours per day (8.66%)
iv. Internet access at home (8.04%)
v. Performance in previous semester (7.58%)

Lower-ranked components included gender, extracurricular participation, and residential status, indicating weaker influence on dropout tendencies.

Ethical Considerations

The study ensured confidentiality of student records. Identifiers such as names, matriculation numbers, and contact details were removed prior to analysis. Ethical approval was obtained from the institution's research and ethics committee in line with global educational data mining practices.

RESULTS AND DISCUSSION

The PCA results as shown on Table 2 reveals that academic engagement and learning behavior dominate student dropout dynamics. Variables directly related to learning commitment contributed more significantly than demographic factors.

The cumulative variance explained (93.26%) confirms that PCA effectively reduced dimensionality while retaining critical information. This highlights PCA's suitability for educational datasets characterized by multicollinearity and overlapping features.

The findings reinforce existing literature that emphasizes the central role of academic integration in student retention. By identifying high-impact variables, institutions can design targeted interventions such as attendance monitoring, academic mentoring, and study-skills support programs.

Table 2: Eigen Values and corresponding Percentage Explained Variance for input features

Rank	Feature Name	Eigen value	EVP (%)	CEVP (%)
1	Attendance in Classes	2.5504	12.14	12.14
2	Previous Academic Results (Gpa)	2.1058	10.02	22.16
3	Study Hours Per Day	1.8198	8.66	30.82
4	Internet Access at Home	1.6892	8.04	38.86
5	Performance in Previous Semester	1.5933	7.58	46.44
6	Residential Status	1.4283	6.80	53.24
7	Father's Educational Level	1.3664	6.50	59.74
8	Mother's Educational Level	1.2333	5.87	65.61
9	Confidence Level in Current Courses	1.0336	4.92	70.53
10	Motivation Level For Academic Success	0.8817	4.20	74.73
11	Sleep Duration Per Night	0.8146	3.88	78.60
12	Preferred Learning Style	0.7159	3.41	82.01
13	Family Income Level (Monthly)	0.6659	3.17	85.18
14	Number of Siblings	0.6362	3.03	88.21
15	Use of Private Tutoring	0.5721	2.72	90.93
16	Mode of Study	0.4890	2.33	93.26
17	Main Challenges In Studies	0.4449	2.12	95.38
18	Participation In Extracurricular Activities	0.2937	1.40	96.77
19	Use of Social Media (Hour Per Day)	0.2623	1.25	98.02
20	Parental Support in Studies	0.2511	1.20	99.22
21	Gender	0.1644	0.78	100.00
22	Residential Status	0.0000	0.00	100.00

CONCLUSION

This study demonstrated the effectiveness of Principal Component Analysis in identifying key determinants of student dropout in Nigerian polytechnics. PCA successfully reduced data complexity while preserving meaningful variance, enabling clearer interpretation of factors influencing student persistence.

The results indicate that academic engagement, particularly attendance, prior performance, and study behavior plays a dominant role in student retention. The study provides a strong empirical foundation for data-driven academic support systems and policy formulation.

Future research may extend this work by incorporating longitudinal datasets, psychological indicators, or real-time learning analytics to further enhance predictive insights and institutional responsiveness.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Data Availability Statement

The Raw data supporting the conclusion of this article is available at https://doi.org/10.5281/zenodo.14787591 and will be made available by authors on request.

Funding

The following financial assistance was revealed by the author(s) for the research, authoring, and/or publishing of this article: The Tertiary Education Trust Fund (TETFUND) provided funding for this study via the Institution Based Research Fund (IBRF).

Acknowledgments

The author(s) are grateful to Federal Polytechnic Ukana and Akwa Ibom State Polytechnic, for providing an enabling environment for the conduct of this research.

REFERENCES

Abiodun, O. J., & Wreford, A. I. (2024). Student performance evaluation using ensemble machine learning algorithms. *Engineering and Technology Journal*, *9*(08), 1–12.

Ahmed, S., Aliyu, M., & Yusuf, T. (2020). Predicting student dropout with ensemble models. *Journal of Educational Data Science*, *8*(2), 45–59.

Carballo-Mendívil, B., Valdez-Godínez, M. A., & García-Ramírez, J. A. (2025). Feature importance analysis for student dropout prediction using machine learning techniques. *International Journal of Educational Data Mining*, *6*(1), 33–48.

Chen, L., & Li, X. (2022). Early warning system for student dropout using decision tree algorithms. *International Journal of Learning Analytics*, *9*(1), 22–35.

Dasi, V., & Kanakala, S. (2022). Student dropout prediction using machine learning techniques: A comparative study. *Journal of Educational Computing Research*, *60*(4), 945–967.

Ekubo, E. A., & Esiefarienrhe, B. M. (2022). Using machine learning to predict low academic performance at a Nigerian university. *The African Journal of Information and Communication*, (30), 1–18.

Inyang, U. P., & Johnson, E. A. (2025). Intelligent ensemble learning framework for prediction of students' academic performance using extreme gradient boosting and Random Forest algorithms. *European Journal of Computer Science and Information Technology, 13*(3), 1–19.

Johnson, E. A., Inyangetoh, J. A., & Esang, M. O. (2021). An experimental comparison of classification tools for fake news detection. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 10(4), 45–53.

Johnson, E. A., Inyangetoh, J. A., Rahmon, H. A., Jimoh, T. G., Dan, E. E., & Esang, M. O. (2024). An intelligent analytic framework for predicting students' academic performance using multiple linear regression and Random Forest. *European Journal of Computer Science and Information Technology, 12*(3), 56–70.

Kumar, M., Singh, N., Wadhwa, J., Singh, P., & Qtaishat, A. (2024). Utilizing Random Forest and XGBoost data mining algorithms for anticipating students' academic performance. *International Journal of Modern Education and Computer Science, 16*(2), 29–44.

Okolie, U. C., Igwe, P. A., & Nwosu, H. E. (2020). Student dropout patterns in Nigerian higher institutions: Implications for educational planning. *International Journal of Education and Development, 5*(3), 112–126.

Smith, J., Brown, K., & Allen, R. (2020). Predicting student dropouts using Random Forest. *Educational Analytics Review, 14*(1), 1–14.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125.

Zhang, Y., & Liu, H. (2019). Comparative study of machine learning algorithms for dropout prediction. *Journal of Applied Educational Analytics, 7*(4), 33–50.