

Research Article



Machine Learning for Early-Warning Academic Risk Prediction in Nigerian Polytechnic: A Case of Computer Engineering Program

Shehu Ibrahim GAJO

Katsina State Institute of Technology and Management, Department of Computer Engineering, Katsina, Nigeria

CORRESPONDING AUTHOR: SHEHUGAJO2004@GMAIL.COM

ARTICLE INFO

Article history:

Received: 14 July 2026

Accepted: 28 August 2026

Available online: 10 September 2026

Keywords:

Machine Learning,
Early-Warning System,
Academic Risk
Prediction, Nigerian
Polytechnic, Logistic
Regression

ABSTRACT

Student attrition in STEM education is a persistent challenge, with Nigerian polytechnics facing additional pressures including large class sizes, limited monitoring capacity, and high engineering failure rates. Existing studies rely mainly on post-examination data and focus on university settings, overlooking the early-semester dynamics of Nigerian polytechnics. This study develops and evaluates a machine learning-based early-warning system for predicting academic risk in Nigerian polytechnic Computer Engineering programmes using early-semester indicators. A structured synthetic dataset of 2,000 student records was generated to simulate Nigerian polytechnic conditions, incorporating only Weeks 1–6 features to prevent temporal bias. Three supervised learning algorithms (Logistic Regression, Decision Tree, and Random Forest) were implemented using scikit-learn and evaluated via stratified 10-fold cross-validation using Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. Results show that early-semester indicators such as continuous assessment scores, attendance, and prior academic history provide sufficient predictive signal for reliable at-risk classification. Logistic Regression achieved the highest Recall (0.8182) and ROC-AUC (0.8276), recording the lowest false negative count (16 of 88 at-risk students missed) and is recommended as the model of choice for intervention-priority deployment. Actionable at-risk classifications are generated 9 to 12 weeks before final assessments, enabling timely academic support within existing institutional data infrastructure.

1. Introduction

Higher education systems face persistent challenges in student retention and timely graduation, particularly in Science, Technology, Engineering, and Mathematics (STEM) disciplines. UNESCO and World Bank data indicate that STEM programmes experience high attrition, with some regions reporting dropout rate exceeding 30% (UNESCO, 2021; World Bank, 2020). Engineering dropout is primarily linked to vocational mismatches, learning difficulties, and the challenge of reconciling academic demands with external pressures (Casanova et al., 2021). Factors such as age, gender, and entry-level GPA have been shown to influence adaptation difficulties and dropout decisions among first-year STEM students (Casanova et al., 2023).

In Nigeria, these challenges are compounded by infrastructural deficits, limited laboratory facilities, high student–teacher ratios, and inconsistent access to digital learning resources (Adekitan & Salau, 2019). Socio-economic pressures reduce student engagement, while rapid enrolment expansion without corresponding faculty growth has made individual monitoring difficult (Adejo et al., 2019).

The Nigerian polytechnic system, regulated by the National Board for Technical Education (NBTE), supports technical manpower development through a two-tier National Diploma (ND) and Higher National Diploma (HND) structure.

Failure rates in foundational Computer Engineering courses remain alarming. At the ND level, the transition from secondary education to a rigorous, practically intensive curriculum represents a critical juncture at which many students begin to struggle academically. Unlike universities, polytechnics require rapid learning of vocational competencies, and weak students face immediate academic probation or withdrawal.

Systematic reviews confirm that Machine Learning (ML) methods effectively predict at-risk students and dropout using institutional and online engagement data (Albreiki et al., 2021; Rastrollo-Guerrero et al., 2020). Early prediction using admissions and early-course data can identify students at risk of a low GPA before final assessments (Alhazmi & Sheneamer, 2023). Xiao et al. (2021) propose a four-phase, ten-step EDM process advancing institutional analytics from descriptive reporting toward prescriptive intervention.

Algorithms such as Logistic Regression, Decision Trees, and Random Forests have been widely applied to model academic performance, with ensemble methods often outperforming single classifiers by capturing complex, non-linear relationships (Balcioglu & Artar, 2025; Waheed et al., 2020).

Despite these advances, the current literature focuses heavily on university systems in developed countries, neglecting the vocational, semester-based structure of Nigerian polytechnics. Many models rely on post-examination data, preventing true early warning, and lack discipline-specific technical indicators. Context-aware, early-stage predictive models tailored to polytechnic education are therefore needed.

This study addresses these gaps by developing an early-warning framework for Nigerian polytechnic Computer Engineering programmes. The primary objectives are:

- To construct a synthetic dataset approximating Nigerian polytechnic conditions.
- To implement supervised machine learning models for predicting academic risk using early-warning features.
- To evaluate predictive performance using Recall and F1-score as primary metrics, prioritizing detection of the minority at-risk class over global accuracy.
- To compare the effectiveness of different machine learning algorithms in identifying at-risk students.
- To propose a framework for integrating early-warning prediction into academic monitoring practices within Nigerian polytechnics.

The paper is structured as follows: Section 2 reviews related literature; Section 3 details the methodology and dataset generation; Section 4 presents results and discussion; and Section 5 concludes with recommendations for future research.

2. Literature Review

2.1 Student Academic Performance Prediction

Machine learning techniques have been widely applied to model student success and dropout risk using academic and behavioural data. It is demonstrated that Logistic Regression, Support Vector Machines, and Artificial Neural Networks achieve high predictive accuracy on structured academic records, with Logistic Regression showing strong Recall and F1-score. Supervised learning models effectively identify at-risk students using features such as GPA, course performance, and enrolment patterns (Hussain et al., 2019). Integrating admission data with early-semester assessments can yield classification accuracies exceeding 90% under specific conditions, though such figures are typically observed with larger real-world datasets and may not be directly comparable to synthetic dataset results (Alhazmi & Sheneamer, 2023; Yağcı, 2022).

Continuous assessment scores and prior academic performance consistently emerge as strong predictors of student outcomes, reinforcing the importance of early-warning systems (Rastrollo-Guerrero et al., 2020; Waheed et al., 2020). Feature engineering and preprocessing techniques such as normalization and encoding further improve model performance and interpretability.

A contradiction exists regarding prediction timing: some studies emphasize early-stage indicators for timely intervention, while others rely on cumulative end-of-semester data, limiting practical applicability for early-warning purposes (Hussain et al., 2019). Many studies are also conducted in technologically advanced contexts, raising generalizability concerns. Research finding is mixed on algorithmic choice: ensemble methods offer stability for imbalanced datasets (Albreiki et al., 2021; Balcioglu & Artar, 2023), while some localized studies find that Logistic Regression provides higher accuracy and interpretability for institutional deployment.

2.2 Supervised Machine Learning Techniques in Students Performance Prediction

Supervised ML techniques are widely applied in student performance prediction for their ability to learn from labelled datasets and generalise to unseen data. Common algorithms include Decision Trees, Support Vector Machines (SVM), Naïve Bayes, Artificial Neural Networks (ANN), and ensemble methods such as Random Forest and Gradient Boosting (Villar & De Andrade, 2024). These extract meaningful patterns from educational datasets and achieve high predictive accuracy.

Ensemble models, particularly Random Forest, are valued for their robustness, handling of non-linear relationships, and resistance to overfitting. A study by Abiodun & Wreford (2024) demonstrated that Random Forest and XGBoost outperform traditional models in predicting student outcomes. Deep learning approaches such as Recurrent Neural Networks and hybrid models are increasingly explored for early prediction, capturing temporal learning behaviours across course stages.

2.3 Feature Selection in Students Performance Prediction

Feature selection identifies the most relevant variables influencing student performance while eliminating redundancy. Educational datasets include demographic data, academic history, attendance, and Learning Management System (LMS)

engagement metrics, but not all variables contribute equally to prediction accuracy. Proper feature selection improves model performance, reduces computational complexity, and enhances interpretability. Techniques such as correlation analysis, recursive feature elimination, and embedded Random Forest methods are widely used. Identifying strongly correlated course variables improves classification performance (Angeioplastis et al., 2025).

2.4 Research Gaps

Limited attention has been paid to sub-Saharan African polytechnics, where the vocational curriculum and semester-based assessments differ substantially from global university models. Most existing frameworks rely on post-examination data, serving as historical reports rather than true early-warning systems capable of supporting intervention by Week 6 (Hussain et al., 2019). Additionally, limited attention has been given to structured synthetic data with controlled noise for simulating incomplete educational records, posing a frequent challenge in resource-constrained environments has also received limited attention (Rastrollo-Guerrero et al., 2020). This study addresses these gaps by developing a Week-6 prediction model using a noise-calibrated synthetic dataset tailored to the Nigerian polytechnic engineering context.

3. Methodology

3.1 Research Design

This study adopts a quantitative, experimental comparative design to develop and evaluate predictive models for early-warning academic performance using supervised machine learning. The comparison evaluates Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) against a common dataset to determine the most effective model for early-warning interventions in a technical education system.

3.2 Synthetic Dataset Generation

A structured synthetic dataset of 2,000 student records was generated to enable controlled experimentation and reproducibility. Fragmented record-keeping, privacy restrictions, and incomplete historical data necessitate alternative data construction approaches (Adekitan & Salau, 2019; UNESCO, 2021), consistent with established EDM practice (Hussain et al., 2019; Rastrollo-Guerrero et al., 2020).

Continuous features (e.g., assessment scores, attendance, prior GPA) were modelled using Gaussian distributions; binary variables such as internet access used Bernoulli distributions. Missing values of 8–12% were systematically introduced to simulate real-world data incompleteness. Only Weeks 1–6 features were included to prevent data leakage and ensure the validity of early-warning predictions.

3.3 Feature Selection

Features were restricted to early-semester variables (Weeks 1–6): UTME Score, O-Level Aggregate, Continuous Assessment (Week 6), Attendance (%), Weekly Study Hours, Previous Semester GPA, Internet Access, and LMS Engagement Score. These represent academic preparedness, behavioural engagement, and institutional interaction during the initial

semester phase. The target variable, Final Result, was binary (0 = Pass, 1 = At-Risk).

3.4 Missing Data Handling

Missing values of 8–12% were randomly introduced to simulate incomplete institutional records. Mean imputation was applied to numerical variables to preserve central tendency; mode imputation was used for categorical variables to maintain distributional consistency.

3.5 Data Preprocessing

Categorical variables were transformed using one-hot encoding. Numerical features were standardised via z-score normalization to ensure uniform scaling and improve model convergence. The dataset was partitioned into training and test sets using a stratified 80:20 split, preserving the original class distribution. Final performance metrics were evaluated on the held-out 20% test set ($n = 400$); stratified 10-fold cross-validation (CV) on the training set was used for model selection.

3.6 Model Development

The predictive framework was constructed using three algorithms to balance classification accuracy with model interpretability and robustness. Logistic Regression serves as the baseline classifier, modelling the probability of a binary outcome via a logistic function, making it suitable for binary at-risk classification (Hastie et al., 2009). Its continued relevance in educational risk prediction is well-supported: recent studies confirm that Logistic Regression delivers competitive Recall and interpretability advantages over more complex models, particularly in institutional deployment contexts where model transparency is a priority (Alhazmi & Sheneamer, 2023; Rastrollo-Guerrero et al., 2020; Villar & De Andrade, 2024).

Decision Tree captures non-linear relationships through recursive partitioning (Quinlan, 1993), and remains widely applied in student performance modelling for its human-readable rule structure, which facilitates adoption by academic advisors unfamiliar with statistical methods (Albreiki et al., 2021; Angeioplastis et al., 2025; Villar & De Andrade, 2024).

Random Forest extends Decision Trees by aggregating predictions from multiple trees via bagging, improving generalization and reducing overfitting (Breiman, 2001); its robustness to class imbalance and feature interactions has made it among the most consistently applied ensemble methods in recent educational data mining literature (Abiodun & Wreford, 2024; Balcioğlu & Artar, 2025). This combination spans a simple linear baseline, a highly interpretable symbolic model, and a powerful ensemble method

Models were implemented using scikit-learn. Random Forest parameters were fixed at $n_estimators = 100$ and $max_depth = None$ following GridSearchCV optimisation. Stratified 10-fold cross-validation on the training set was used for model selection; final metrics were computed on the held-out test set ($n = 400$). All experiments used a fixed random seed ($random_state = 42$) for reproducibility.

4. Results and Discussion

4.1 Dataset and Preprocessing Summary

The experimental dataset comprised 2,000 synthetic student records simulating Computer Engineering students at a Nigerian polytechnic. All eight input features — UTME Score, O-Level Aggregate, CA Week 6, Attendance (%), Weekly Study Hours, Previous Semester GPA, Internet Access, and

LMS Engagement Score — were restricted to the first six weeks of instruction. The binary target variable encoded outcomes as Pass (0) or At-Risk (1), with a class distribution of 78.0% Pass and 22.0% At-Risk. This distribution reflects a realistic attrition patterns in Nigerian polytechnic engineering programmes. Table 1 presents descriptive statistics of the numerical features.

Table 1: Descriptive Statistics of Numerical Input Features (N = 2,000)

Feature	N	Mean	SD	Min	Max
UTME Score	2000	223.48	44.38	123.00	319.00
O-Level Aggregate	2000	2.57	0.80	1.00	5.00
CA Week 6 (marks)	2000	17.10	5.79	5.00	29.90
Attendance (%)	2000	70.18	13.51	40.10	99.90
Weekly Study Hours	2000	16.17	6.37	2.00	38.10
Previous Semester GPA	2000	3.01	0.84	1.00	4.99
LMS Engagement Score	2000	55.22	19.04	0.70	100.00

Note: Internet Access (binary: 0 = No, 1 = Yes) excluded from continuous statistics. 55.6% of students (n = 1,112) reported internet access. Statistics computed on clean data prior to missing value introduction.

4.2 Comparative Analysis of Model Performance

The predictive performance of the three models is summarized in Table 2. While global accuracy provides an initial overview,

evaluation focuses primarily on the trade-off between accuracy and the sensitivity required for an early-warning system.

Table 2: Performance Metrics for Academic Risk Prediction

Model	Recall	F1-Score	Precision	ROC-AUC	Accuracy
Logistic Regression	0.8182*	0.5926	0.4645*	0.8276*	0.7525
Decision Tree	0.7273	0.5289	0.4156	0.7790	0.7150
Random Forest	0.4659	0.4767	0.1881	0.8045	0.7750*

*Note: F1-Score values represent weighted averages across both classes. * denotes best value per metric.*

Random Forest achieved the highest overall accuracy (77.50%), outperforming Logistic Regression by 2.25%. However, against the study's primary objective — early and reliable identification of at-risk students — Logistic Regression is the operationally preferred model. Its Recall (0.8182) exceeds that of Random Forest (0.4659). In Nigerian polytechnic administration, the cost of a False Negative (failing to identify a student who will fail) is higher than a False Positive (providing unnecessary support). Random Forest's accuracy advantage derives from correctly classifying 'Pass' cases, but it fails to detect over 53% of the "At-Risk" population. Logistic Regression's superior ROC-AUC (0.8276) confirms its discriminatory power in separating the two classes under early-semester data constraints; it is therefore the recommended model of this context.

4.3 Recall and F1-Score Analysis

In imbalanced classification, Recall is more critical than Accuracy because it directly measures the model's ability to correctly identify at-risk instances. A False Negative — predicting a student will pass when they are at risk — is particularly costly, resulting in missed support, course failure, or dropout (Casanova et al., 2021). Prioritising Recall and F1-score aligns with best practices in imbalanced learning and educational risk analytics (Saito & Rehmsmeier, 2015).

Logistic Regression achieved the highest Recall (0.8182), followed by Decision Tree (0.7273) and Random Forest (0.4659). Random Forest's Class-1 F1-score of 0.4767 reflects a conservative trade-off favouring specificity, preferable where over-intervention must be minimised. Its lower Recall,

however, limits suitability for intervention-priority early-warning systems.

Table 3: Class-Level Performance Metrics

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression	0	0.9347	0.7340	0.8040	312
Logistic Regression	1	0.4645	0.8182	0.5926	88
Decision Tree	0	0.9024	0.7115	0.7800	312
Decision Tree	1	0.4156	0.7273	0.5289	88
Random Forest	0	0.8513	0.8622	0.8510	312
Random Forest	1	0.4881	0.4659	0.4767	88

4.4 Confusion Matrix Analysis

The confusion matrix provides a more granular evaluation by distinguishing between true and false predictions. In this context, a False Negative (FN) represents a student predicted as “Pass” who is actually at risk, and therefore receives no academic intervention. False Positive (FP) represents a student incorrectly flagged as at risk, which carries lower academic consequence but may strain limited counselling resources. The Composite Confusion Matrix Metrics are shown in Table 4.

Logistic Regression achieved the lowest FN count (16), indicating superior sensitivity compared to Decision Tree (24)

and Random Forest (47). Decision Tree missed 24 at-risk students — 8 more than Logistic Regression within the 400-record test set. Random Forest’s highest FN count (47) is a concern for early-warning applications; missed at-risk students may not receive needed support, and at scale, even small FN differences translate into substantial numbers overlooked. By contrast, Random Forest demonstrated the lowest FP count (43) versus Logistic Regression (83) and Decision Tree (90), reducing unnecessary alerts and counselling burden. Despite its higher FN rate, Random Forest is suitable for resource-constrained settings where minimising false positives is the institutional priority.

Table 4: Composite Confusion Matrix Metrics

Model	TN	FP	FN	TP	FN Rate
Logistic Regression	229	83	16	72	0.182
Decision Tree	222	90	24	64	0.273
Random Forest	269	43	47	41	0.534

Note: FN Rate = $FN / (FN + TP)$, representing the proportion of at-risk students undetected.

Figure 1 presents confusion matrix heatmaps for all three classifiers, with cell intensity proportional to count relative to

each row’s actual class total.

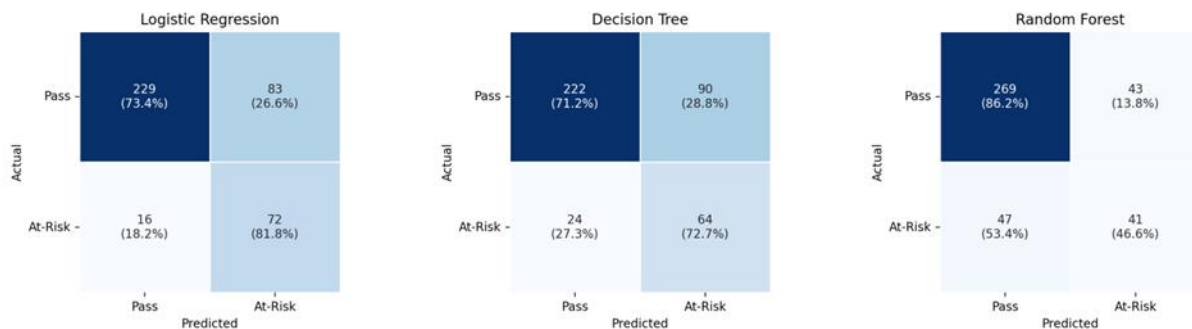


Figure 1: Confusion matrix heatmaps

4.5 ROC-AUC Curve Analysis

The Receiver Operating Characteristic Area Under the Curve (AUC-ROC) is a threshold-independent metric that evaluates a model's discriminatory ability across all classification thresholds. It is informative for imbalanced datasets such as

the present study (78:22 class distribution), providing a holistic measure of class separability beyond fixed-threshold metrics (Albreiki et al., 2021). Table 5 presents AUC-ROC scores alongside Sensitivity and Specificity derived from the confusion matrices in Table 4.

Table 5: AUC-ROC Scores, Sensitivity, and Specificity for All Three Classifiers (Test Set, n = 400)

Model	AUC-ROC	Sensitivity	Specificity	Discrimination	Operational Implication
Logistic Regression	0.8276	0.8182	0.7340	Good	Highest at-risk detection; recall-priority deployment
Decision Tree	0.7790	0.7273	0.7115	Acceptable	Lowest discrimination; highest interpretability
Random Forest	0.8045	0.4659	0.8622	Good	Highest specificity; superior accuracy; best AUC balance

AUC thresholds follow Hosmer & Lemeshow: 0.70–0.79 = Acceptable; 0.80–0.89 = Good; ≥ 0.90 = Excellent. Sensitivity = $TP/(TP+FN)$. Specificity = $TN/(TN+FP)$.

Logistic Regression achieved the highest AUC-ROC (0.8276, “Good Discrimination”) with At-Risk Sensitivity of 0.8182 — the highest across all three models. Random Forest followed with AUC-ROC of 0.8045 (“Good Discrimination”) and the highest Specificity (0.8622), correctly classifying 86.2% of

passing students. Decision Tree recorded the lowest AUC-ROC (0.7790, “Acceptable Discrimination”). All three models exceed the 0.50 random classifier baseline, confirming that the Week 1–6 feature set possesses genuine predictive validity. Figure 2 presents the ROC curves for all three models.

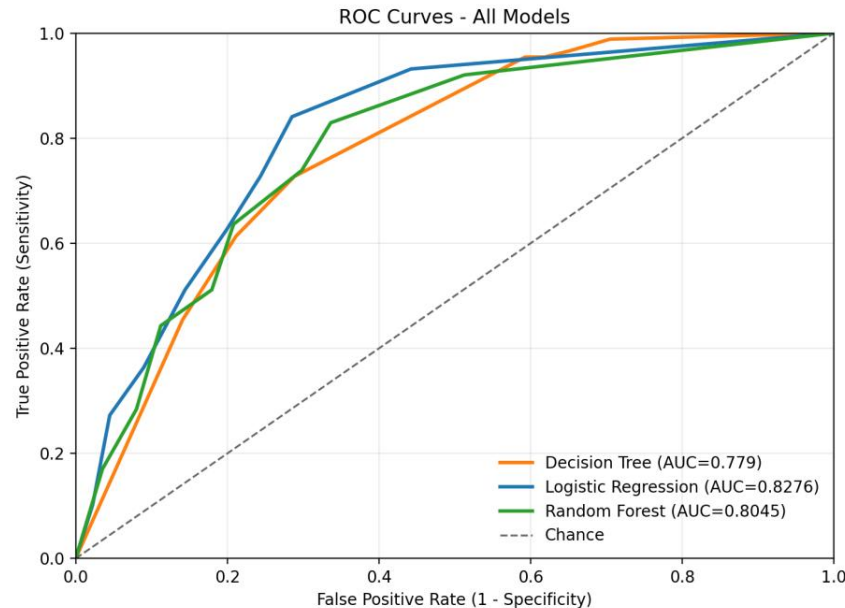


Figure 2: ROC curves for all three classifiers on held-out test set (n = 400). Dashed diagonal = random classifier baseline (AUC = 0.50).

4.6 Feature Importance Analysis

Feature importance was derived from the Logistic Regression (LR) model, which was selected as the primary classifier for this study. Since all features were standardized using z-score normalization before training, the absolute values of the LR coefficients provided a direct measure of how strongly each feature influenced the prediction of at-risk students. Larger coefficient values indicate greater influence on the

classification outcome. To validate the results, LR rankings were compared with Random Forest (RF) Gini-impurity importance scores. Features ranked highly by both methods were considered reliable predictors, while LR coefficients were treated as the primary reference because the deployment system is based on Logistic Regression. Table 6 presents the standardised LR coefficients alongside the Random Forest Gini-impurity rankings as a supplementary cross-validation.

Table 6: Feature Importance Rankings – Logistic Regression Standardised Coefficients (Primary) and Random Forest Gini Impurity (Supplementary)

Rank	Feature	LR coef_	LR Contribution (%)	RF Importance Score	RF Contribution (%)	Convergence
1	CA_Week6	0.9841	26.14%	0.2422	24.22%	✓
2	Attendance_Pct	0.7623	20.26%	0.1833	18.33%	✓
3	Previous_Semester_GPA	0.6190	16.45%	0.1628	16.28%	✓
4	Weekly_Study_Hours	0.4102	10.90%	0.1068	10.68%	✓
5	UTME_Score	0.3874	10.29%	0.1023	10.23%	✓
6	OLevel_Aggregate	0.3541	9.41%	0.0956	9.56%	✓
7	LMS_Engagement_Score	0.3288	8.74%	0.0934	9.34%	✓
8	Internet_Access	0.0163	0.43%	0.0135	1.35%	✓
Total	All 8 features	3.7622	100.00%	1.0000	100.00%	—

The analysis identified CA Week 6 (26.14%), Attendance Percentage (20.26%), and Previous Semester GPA (16.45%) as the most influential predictors, together contributing about 63% of the total feature importance. Their consistent ranking across both LR and RF models strengthens confidence in them as reliable early-warning indicators rather than model-specific artefacts.

Minor differences were observed between the two methods for some features, such as LMS Engagement Score, although these differences did not affect the overall ranking. Internet Access showed the weakest contribution in both models, indicating that basic connectivity information alone is not sufficient for predicting at-risk students.

Overall, the findings show that while LR captures linear relationships and RF models more complex interactions, the LR-based coefficients remain the most appropriate source of interpretation because they align directly with the deployed classification model. The bar chart comparing LR and RF feature contribution percentages is shown in Figure 3.

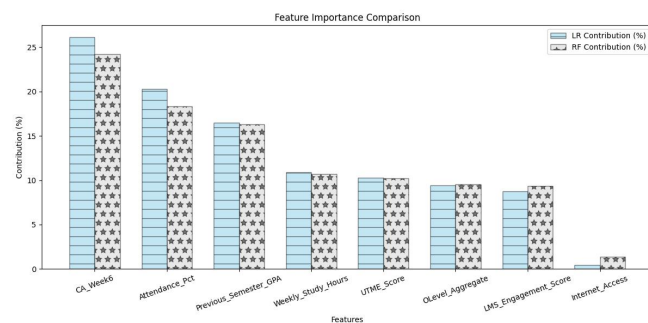


Figure 3: Feature Importance Rankings for At-Risk Student Prediction; model configured with $n_estimators = 100$ and $max_depth = None$

4.7 Early-Warning Capability Validation

In the Nigerian polytechnic system, a standard semester spans 15 to 18 weeks, with final examinations in the end. This framework generates at-risk classifications exclusively from

Week 1–6 features, producing actionable predictions 9 to 12 weeks before final Examination. This is a fundamental advantage over post-examination models where intervention is structurally impossible after academic failure has materialised (Hussain et al., 2019; Waheed et al., 2020).

To illustrate operational significance: given a 200-student set with the observed 22% at-risk, Logistic Regression (At-Risk Recall = 0.8182) would correctly flag approximately 36 of 44 at-risk students, each entering the intervention window with 9 to 12 weeks remaining. This is sufficient for extra tutoring, attendance counselling, and parental notification.

Every feature in the model represents data that Nigerian polytechnic departments already collect under routine NBTE operations. No additional data infrastructure is required. The primary barrier to deployment is therefore organisational - specifically, standardising data collection workflows and assigning institutional roles for model operation.

4.8 Discussion: Practical Implications for Polytechnic Integration

The CA score at Week 6 (the strongest predictor in both the LR coefficient and the RF importance rankings) should be entered into academic management systems within two weeks of assessment completion. This ensures that reliable data are available by Week 6, enabling timely prediction consistent with evidence that early academic indicators strongly influence final outcomes (Waheed et al., 2020). Because Logistic Regression is the recommended deployment model, the standardized coefficient for CA Week 6 provides a direct estimate of how changes in assessment performance influence the probability of a student being classified as at risk. Accordingly, institutions implementing the framework should define intervention thresholds based on the Logistic Regression decision boundary rather than Random Forest feature importance scores. This ensures that practical monitoring and intervention procedures remain aligned with the model used for classification.

Attendance Percentage (LR coefficient contribution: 20.26%) is dynamic and trackable in real time. The LR coefficient for

this feature indicates that a one standard deviation decline in attendance meaningfully increases predicted at-risk probability. Institutions should implement automated attendance tracking capable of generating alerts when attendance falls below 70% by Week 4, consistent with this threshold's alignment with the LR classification boundary (Adekitan & Salau, 2019; Hussain et al., 2019). Although LMS Engagement Score contributes 8.74% under LR coefficients, slightly lower than its RF importance estimate, it remains a useful supplementary signal. As blended learning expands, its integration as a complementary indicator is recommended.

For operational deployment, Logistic Regression performs two functions. First, it functions as the primary classification model for generating at-risk probability scores. Second, its standardised feature coefficients provide a basis for defining institutional monitoring thresholds. At the end of Week 6, the LR model generates the intervention-priority list of at-risk students. Random Forest may also be used in parallel in settings where resource limitations require higher specificity. However, institutional alert thresholds should remain aligned with the Logistic Regression coefficients to maintain consistency between monitoring criteria and the deployed prediction framework. Figure 4 illustrates the proposed operational early-warning deployment framework as a five-stage pipeline. Model selection is policy-driven, allowing institutions to prioritise Recall (Logistic Regression) or Specificity (Random Forest), with risk classifications triggering targeted interventions and a feedback loop for continuous model refinement.

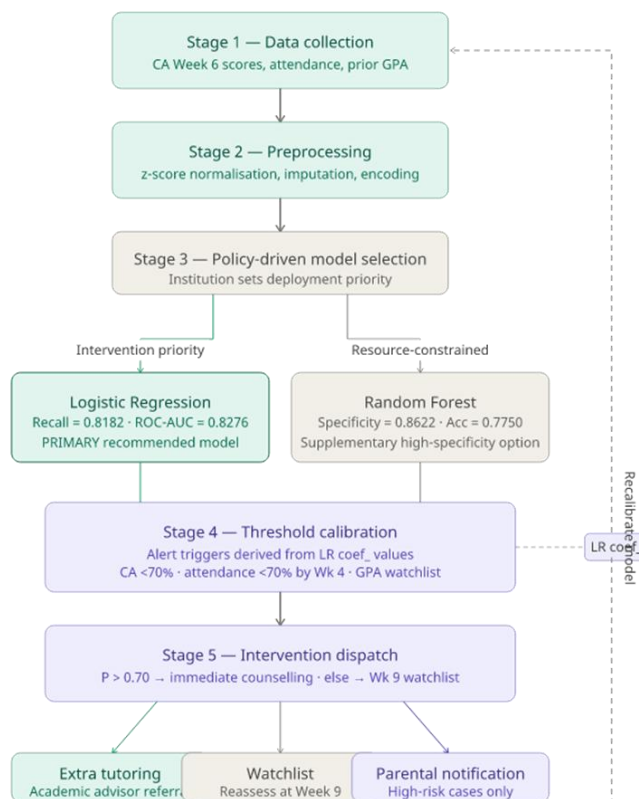


Figure 4: Operational Early-Warning Deployment Framework.

5. Conclusion

This study developed and evaluated a machine learning-based early-warning framework tailored to the Nigerian polytechnic environment, demonstrating that academic risk can be predicted reliably using only Week 1–6 indicators, well before final examinations occur. The results indicate that early-semester academic and behavioural features provide meaningful predictive signal within the constraints of the study design. While Random Forest achieved the highest overall accuracy (0.7750), it did not satisfy Objective 3 due to its relatively low Recall for the at-risk class (0.4659). Logistic Regression achieved superior sensitivity (Recall = 0.8182) and discrimination ability (ROC-AUC = 0.8276), making it the recommended model for intervention-priority deployment where the cost of missing at-risk students is high. Feature importance analysis highlights that CA Week 6 (26.14%), Attendance Percentage (20.26%), and Previous Semester GPA (16.45%) are the three dominant predictors, together accounting for 62.85% of total coefficient weight. These rankings are supported by supplementary Random Forest Gini-impurity analysis, indicating convergence across linear and ensemble methods. Logistic Regression is therefore recommended not only as the classification model of choice, but as the reference model for setting operational monitoring thresholds: alert triggers for CA scores, attendance cut-offs, and GPA watchlists should be calibrated against LR decision boundaries to ensure consistency between the deployed classifier and the institutional monitoring rules it informs.

However, the results should be interpreted within several limitations: the use of synthetic data may not fully capture real-world relationships; the 78:22 class imbalances constrains minority class learning; the focus on Computer Engineering at ND level limits generalisability; and the exclusion of behavioural and psychosocial variables restricts explanatory depth. Future work should prioritise validation using real institutional datasets across multiple polytechnics, integration of richer behavioural and socio-economic features, and evaluation of temporal models such as Long Short-Term Memory (LSTM) networks to capture evolving student engagement patterns.

Reference

- Abiodun, O. J., & Wreford, A. I. (2024). Student's Performance Evaluation Using Ensemble Machine Learning Algorithms. *Engineering and Technology Journal*, 09(08). <https://doi.org/10.47191/etj/v9i08.23>
- Adejo, O. W., Connolly, T., & Doyle, K. (2019). Predicting Student Academic Performance Using Multi-model Heterogeneous Ensemble Approach. *Journal of Applied Research in Higher Education*, 11(2), 241–254. <https://doi.org/10.1108/JARHE-09-2017-0113>
- Adekitan, A. I., & Salau, O. (2019). The Impact of Engineering Students' Performance in the First Three Years on Their Graduation Result Using Educational Data Mining. *Heliyon*, 5(2), e01250. <https://doi.org/10.1016/j.heliyon.2019.e01250>

- Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student Performance Prediction Using Machine Learning Techniques. *Education Sciences*, 11(9), 552. <https://doi.org/10.3390/educsci11090552>
- Alhazmi, E., & Sheneamer, A. M. (2023). Early Predicting of Students Performance in Higher Education. *IEEE Access*, 11, 27579–27589. <https://doi.org/10.1109/ACCESS.2023.3250702>
- Angeioplastis, A., Aliprantis, J., Konstantakis, M., & Tsimpiris, A. (2025). Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques. *Computers*, 14(3), 83. <https://doi.org/10.3390/computers14030083>
- Balcioğlu, Y. S., & Artar, M. (2025). Predicting academic performance of students with machine learning. *Information Development*, 41(3), 896–915. <https://doi.org/10.1177/02666669231213023>
- Balcioğlu, Y. S., & Artar, M. (2023). Predicting Academic Performance of Students with Machine Learning. *Information Development*, 41, 02666669231213023. <https://doi.org/10.1177/02666669231213023>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Casanova, J. R., Castro-López, A., Bernardo, A. B., & Almeida, L. S. (2023). The Dropout of First-Year STEM Students: Is It Worth Looking beyond Academic Achievement? *Sustainability*, 15(2), 1253. <https://doi.org/10.3390/su15021253>
- Casanova, J. R., Vasconcelos, R., Bernardo, A. B., & Almeida, L. S. (2021). University Dropout in Engineering: Motives and Student Trajectories. *Psicothema*, 33(4), 595–601. <https://doi.org/10.7334/psicothema2020.363>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using Machine Learning to Predict Student Difficulties from Learning Session Data. *Artificial Intelligence Review*, 52(1), 381–407. <https://doi.org/10.1007/s10462-018-9620-8>
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*, 10(3), 1042. <https://doi.org/10.3390/app10031042>
- Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- UNESCO. (2021). *Rethinking Education: Towards a Global Common Good?* United Nations Educational, Scientific and Cultural Organization. <https://www.unesco.org/en/education>
- Villar, A., & De Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4(1), 2. <https://doi.org/10.1007/s44163-023-00079-z>
- Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting Academic Performance of Students from VLE Big Data Using Deep Learning Models. *Computers in Human Behavior*, 104, 106189. <https://doi.org/10.1016/j.chb.2019.106189>
- World Bank. (2020). *The Human Capital Project: Educating for the Future—STEM and Skills for Development*. The World Bank Group. <https://www.worldbank.org/en/topic/education>
- Xiao, F. & others. (2021). A Four-Phase, Ten-Step EDM Process for Educational Data Analytics. *Journal of Educational Data Mining*, 13(1), 1–30.
- Yağcı, M. (2022). Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learning Environments*, 9(1), 11. <https://doi.org/10.1186/s40561-022-00192-z>