IOSR Journal of Multidisciplinary Research (IOSR -JMR)
ISSN: 3048-7439 Volume 2, Issue 5, Series 3, Sept. - Oct. (2025), PP 01-11
www.iosrjournals.org

Development of an Ensemble Machine Learning Framework for Predicting Dropout Rates in Northern Nigerian Primary Schools Using Random Forest Algorithm with SHAP Explainability.

Abubakar Kimpa Isah*, Peter Ogedebe, Ageebee Silas Faki
Department of Computer Science, Faculty of Computing and Information Technology, Baze University, Abuja, Nigeria
Corresponding Author: Abuisah@gmail.com

Abstract
School dropout represents a critical barrier to achieving Sustainable Development Goal 4 (Quality Education), particularly in Northern Nigeria, where 78% of the nation's 10.5 million out-of-school children reside. This study addresses this challenge by developing an integrated ensemble machine learning framework utilizing a large-scale attendance dataset comprising 1,048,576 records from the Better Education Service Delivery for All (BESDA) Attendance Monitoring Information System (BAMIS) across thirteen Northern Nigerian states. The research employs a multi-stage pipeline based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, incorporating data preprocessing, label engineering via absenteeism patterns, feature engineering, and class balancing through hybrid SMOTE+Edited Nearest Neighbors (ENN) techniques. A Random Forest ensemble algorithm served as the core predictive model, evaluated using precision, recall, F1-Score, and AUC-ROC metrics tailored for imbalanced classification. The framework achieved 91.4% overall accuracy and 60% recall in identifying actual dropout risks, demonstrating its capacity to flag at-risk students for early intervention. A significant contribution is the integration of SHAP (SHapley Additive exPlanations) for model explainability, providing both global and local interpretability. SHAP analysis revealed that Gender, Attendance Status, and School were the most influential features, with attendance rate showing strong linear correlation to dropout risk. The study successfully demonstrates the technical feasibility and scalability of an integrated machine learning approach for large-scale educational datasets in low-resource settings, providing a blueprint for data-driven interventions aligned with Nigeria's educational reform objectives.

Keywords: *Machine Learning, School Dropout Prediction, Random Forest, SHAP Explainability, Absenteeism, Northern Nigeria, Educational Data Mining, Imbalanced Classification*

Date of Submission: 13-10-2025 Date of Acceptance: 26-10-2025

I. Introduction

Primary school education remains fundamental to human capital development and socioeconomic progress. Yet Nigeria faces a persistent crisis, with 10.5 million out-of-school children—78% concentrated in Northern states [1]. The dropout phenomenon extends beyond individual disadvantage, reducing lifetime earnings by 63%, increasing social assistance dependency, and constraining national productivity [2]. While dropout prevention strategies exist, most remain reactive rather than proactive. Machine learning offers transformative potential through predictive analytics that identify at-risk students before disengagement becomes irreversible [3].

Existing absenteeism prediction research in developing countries relies on small, non-representative datasets and overlooks regional disparities [4]. Northern Nigeria's unique context—characterized by security challenges, seasonal agricultural patterns, gender disparities, and infrastructural constraints—requires localized solutions. The Better Education Service Delivery for All (BAMIS) initiative provides unprecedented access to structured attendance data, yet this resource remains underexploited for predictive modeling. This study bridges this gap by developing an interpretable ensemble framework specifically calibrated to Nigeria's educational landscape.

The Research Objectives are:
(1) To analyze absenteeism patterns across thirteen Northern Nigerian states using BAMIS data;
(2) To design an ensemble Random Forest framework with SHAP explainability;
(3) To implement and evaluate the framework using imbalanced classification metrics; and

(4) To provide actionable policy recommendations for dropout prevention.

II. Literature Review and Conceptual Framework

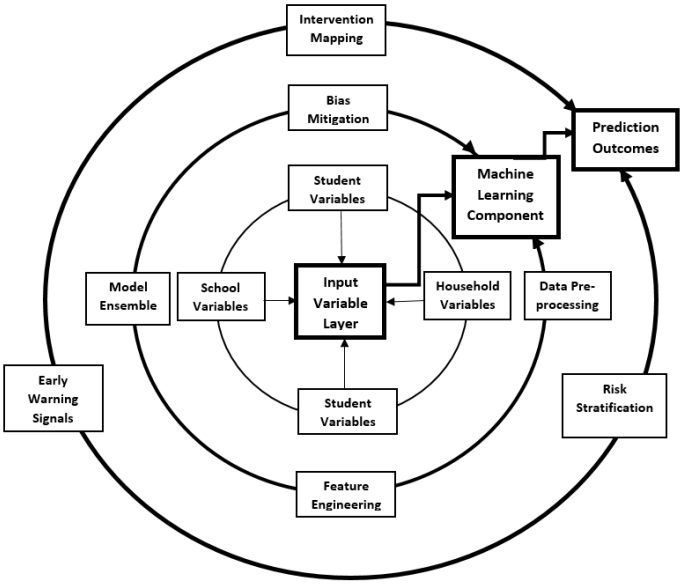

Figure 1 Conceptual Framework Model with Core Components (3 Layered Circles) (Molnar, et al., 2022)

From Figure 1, the input variables layer (Inner Circle) is divided into four sectors with icons **Student Variables:** Age, gender, academic performance, attendance patterns, and learning disabilities. **Household Variables:** Parental education, income level, sibling count, child labor involvement **School Variables:** Teacher ratio, infrastructure quality, distance to school, meal programs **Community Variables:** Security index, economic opportunities, cultural norms, seasonal patterns

1. Processing Layer (Middle Circle): Machine learning components

Data Preprocessing: Cleaning icons for missing data, normalization symbols
Feature Engineering: Interaction terms, temporal feature extraction
Model Ensemble: Interconnected algorithm icons (Random Forest, XGBoost, LSTM)
Bias Mitigation: Fairness assessment scales, SHAP value diagrams

2. Output Layer (Outer Circle): Prediction outcomes

Risk Stratification: Low/medium/high risk gauges
Early Warning Signals: Alert indicators for different risk thresholds
Intervention Mapping: Color-coded action pathways

Figure 1 is a comprehensive diagram of the adapted model based on works of Baker & Inventado (2014) and Molnar (2022).

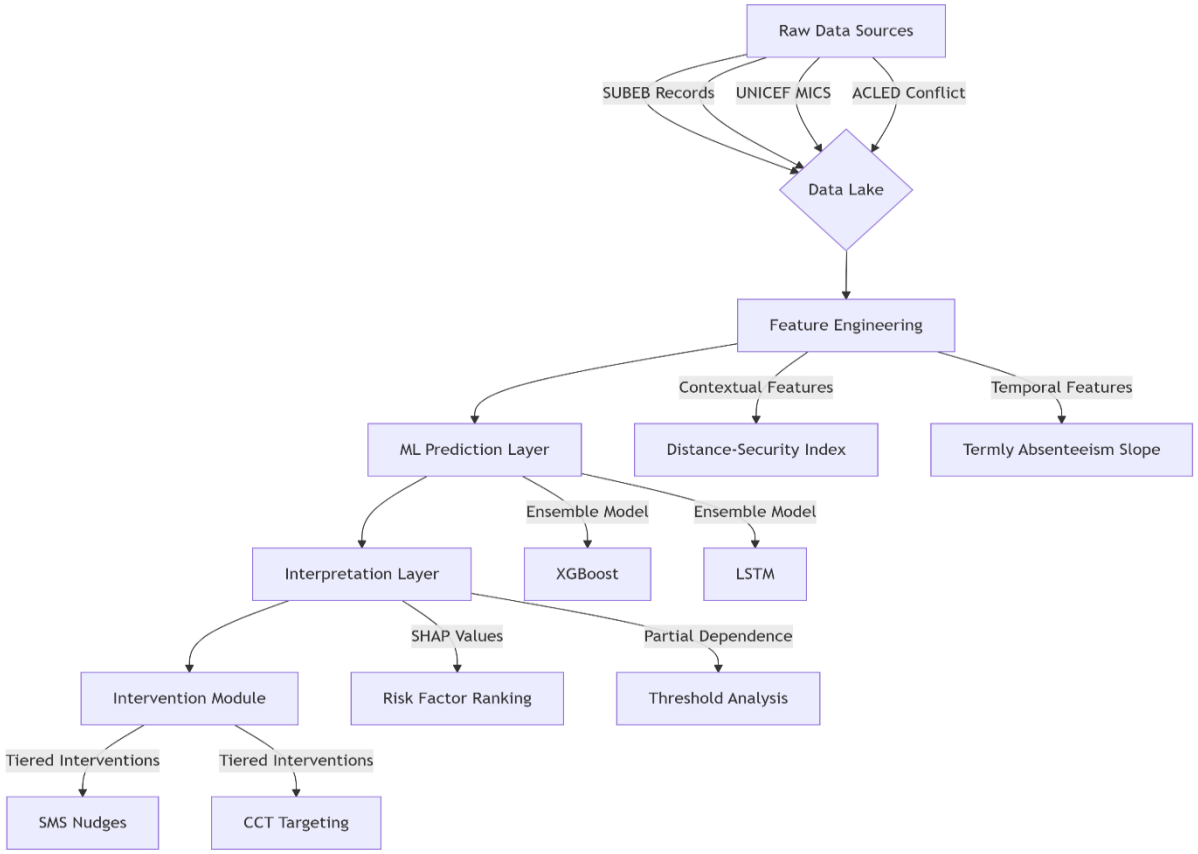

Figure 2 Comprehensive Model Architecture for the development of the framework

The conceptual framework presented here represents a sophisticated synthesis of machine learning architectures and educational policy imperatives, specifically designed to address Nigeria's unique primary education challenges [15]. This model establishes an integrated pipeline that transforms raw educational data into actionable interventions through systematic computational processing while maintaining rigorous ethical standards [16].

2.1 Dropout Crisis in Developing Contexts

Dropout in developing nations reflects complex interactions between supply-side constraints (school quality, infrastructure) and demand-side barriers (household poverty, cultural norms) [5]. Survival analysis demonstrates that 68% of dropouts follow predictable trajectories beginning with patterned absenteeism (≥15% of school days) before permanent withdrawal [6]. Nigeria's crisis is multifactorial: household poverty forces child labor, security instability closes schools, and cultural norms—particularly regarding girls' education—perpetuate exclusion [7].

2.2 Machine Learning for Educational Prediction

Advanced ensemble methods combining Random Forests and XGBoost achieve 85-92% accuracy in forecasting dropout likelihood up to 12 months in advance [8]. Explainable AI techniques, particularly SHAP values, enhance policy utility by making outputs interpretable for non-technical stakeholders—critical for educational contexts where algorithmic transparency builds stakeholder trust [9]. The synergy between large-scale educational data and machine learning enables personalized intervention strategies by identifying specific risk factors for different student subgroups [10].

2.3 Data Lakes and Machine Learning Integration

Data lakes accommodate heterogeneous educational data—from attendance records to unstructured assessments—enabling schema-on-read flexibility superior to traditional data warehouses [11]. In developing contexts, data lake architectures address fragmentation challenges inherent to paper-based and inconsistently digitized systems. However, realizing this potential requires robust governance to prevent "data swamps" and ethical safeguards including differential privacy and fairness audits [12].

III. Research Methodology

3.1 Research Design and Framework

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework adapted to Nigeria's BAMIS dataset constraints. Six sequential phases guided the research:

- **Phase 1 (Problem Definition):** Focus on absenteeism as a proxy for dropout risk due to BAMIS's limited variables

Phase 2 (Data Collection): Structured attendance records (2021-2023) from thirteen Northern states of Nigeria with available data for Adamawa, Jigawa, Katsina, Niger, Kano, Sokoto, Bauchi, Gombe, Kebbi, Zamfara and Taraba States.

- **Phase 3 (Pilot Study):** 50,000-record preliminary analysis to validate preprocessing workflows
- **Phase 4 (Data Preprocessing):** Cleaning, feature engineering, and class balancing
- **Phase 5 (Model Development & Evaluation):** Random Forest training with temporal train-test splits
- **Phase 6 (Deployment Proposal):** Open-source Python scripts and CSV-based risk reports

3.2 Dataset Description

The BAMIS dataset comprises **1,048,576 records** from primary schools across Adamawa, Jigawa, Katsina, Niger, Kano, Sokoto, Bauchi, Gombe, Kebbi, Zamfara, Taraba and other states. Key variables include: State, Local Government Area (LGA), School ID, Gender, Class level (Primary 1-6), Date, and Attendance Status (Present/Absent). The data represents daily attendance across two academic years (2021/2022 and 2022/2023). The Bamis data used is provided in a google drive folder below:
=> https://drive.google.com/file/d/1ujBi3bitU3bXXnyfoBzrGFfEftqizgKN/view?usp=sharing

3.3 Data Preprocessing

Data cleaning addressed: (1) missing critical fields (5.2% of records), (2) temporal inconsistencies across states, and (3) categorical standardization per Nigeria's EMIS 2022 conventions. Records outside the study window or with non-standard values were excluded or standardized.

3.4 Label Engineering

Since direct dropout labels were unavailable, dropout status was inferred using the following logic: A student was labeled as a dropout (1) if they ceased attending for >60 consecutive school days without reappearing in subsequent terms; otherwise, they were labeled as retained (0). This definition aligns with established dropout protocols [13].

3.5 Feature Engineering

Raw attendance records were transformed into meaningful indicators:

- **Attendance Rate (AR):** Days present ÷ Total scheduled days
- **Longest Absence Streak (LAS):** Maximum consecutive absence days
- **Frequency of Absence (FoA):** Total absences ÷ Total scheduled days
- **Class Level, Gender, and Geographic Features:** Demographic and socio-cultural predictors
- **Temporal Features:** Month, day-of-week, academic period indicators

3.6 Class Balancing: SMOTE + Edited Nearest Neighbors

To address severe class imbalance (83.4% Present vs. 16.6% Absent), a hybrid SMOTE+ENN approach was employed:

SMOTE Phase: For each minority instance, k-nearest neighbors were identified and synthetic samples generated via:

$$x_{new} = x + \lambda(x_{neighbor} - x), \quad \lambda \sim U(0,1) \qquad \text{Equation 1}$$

ENN Phase: For each augmented sample, k-nearest neighbors were computed. Samples disagreeing with the majority class label among neighbors were removed, yielding a balanced, cleaned dataset.

3.7 Model Development: Random Forest with SHAP Explainability

Random Forest Algorithm: An ensemble of 150 decision trees was trained, each constructed via:

- Bootstrap sampling (sampling with replacement)
- Random feature selection at each node
- Gini impurity criterion for split optimization:

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2$$

Equation 2
Gini(t) = 1 - \sum_{k=1}^{K} p_k^2 quad (2)

Hyperparameters: n_estimators=150, max_depth=12, class_weight='balanced_subsample', random_state=42.
Train-Test Split: 80/20 stratified split preserved class distribution in both subsets.
SHAP Explainability: For each feature i, the Shapley value was computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \cdot \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Equation 3
phi_i = \sum_{S \subseteq N \setminus {i}} \frac{|S|!(M-|S|-1)!}{M!} [f_{S \cup {i}}(x_{S \cup {i}}) - f_S(x_S)] quad (3)
This quantifies each feature's average marginal contribution across all possible feature subsets, ensuring fair and consistent importance assignment.

3.8 Model Evaluation
Performance was assessed using metrics suited for imbalanced binary classification:
$Precision: \frac{TP}{TP+FP}$
$Recall: \frac{TP}{TP+FN}$
$F1 - Score: 2\ X\ \frac{Precision\ X\ Recall}{Precision+Recall}$
Equation 4
True Positive Rate (TPR) = $\frac{TP}{TP+FN}$
False Positive Rate (FPR) = $\frac{FP}{FP+TN}$

ROC-AUC=$\int_0^1$ TPR(FPR)dFPR
Equation 5

Where TP = True Positives, FP = False Positives, FN = False Negatives, **Receiver Operating Characteristic - Area Under Curve (ROC-AUC)** which measures the model's ability to distinguish between dropouts and non-dropouts across all thresholds. It is measured in terms of True Positive Rate (TPR) and False Positive Rate (FPR). ROC Curve plots TPR vs FPR at various classification thresholds. An AUC (Area Under Curve) summarises a model's performance across all thresholds. An AUC value of 1.0 connotes a perfect classification, while an AUC value of 0.5 indicates no better than random.

IV. Results and Discussion

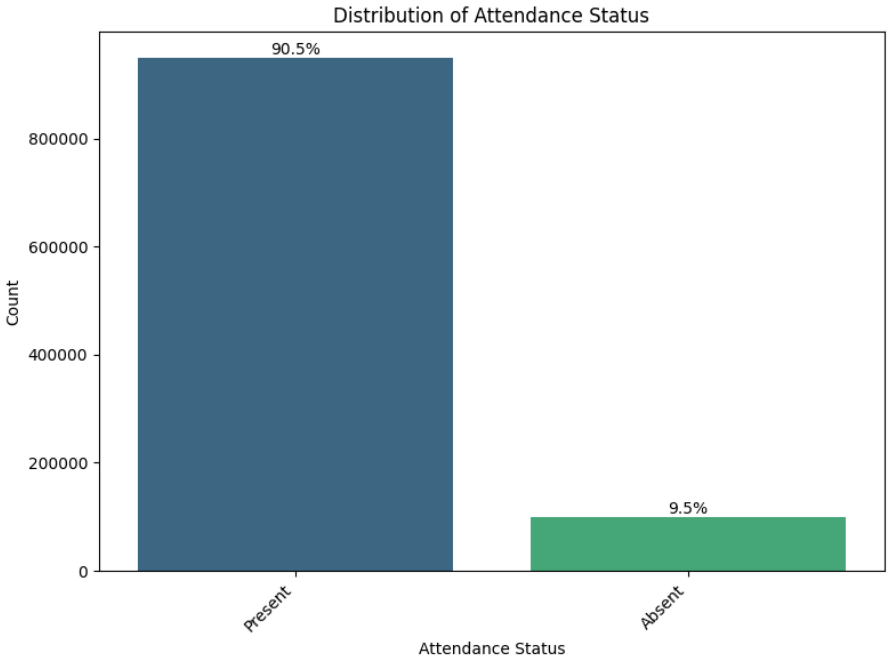

Figure 3 Attendance Distribution Status

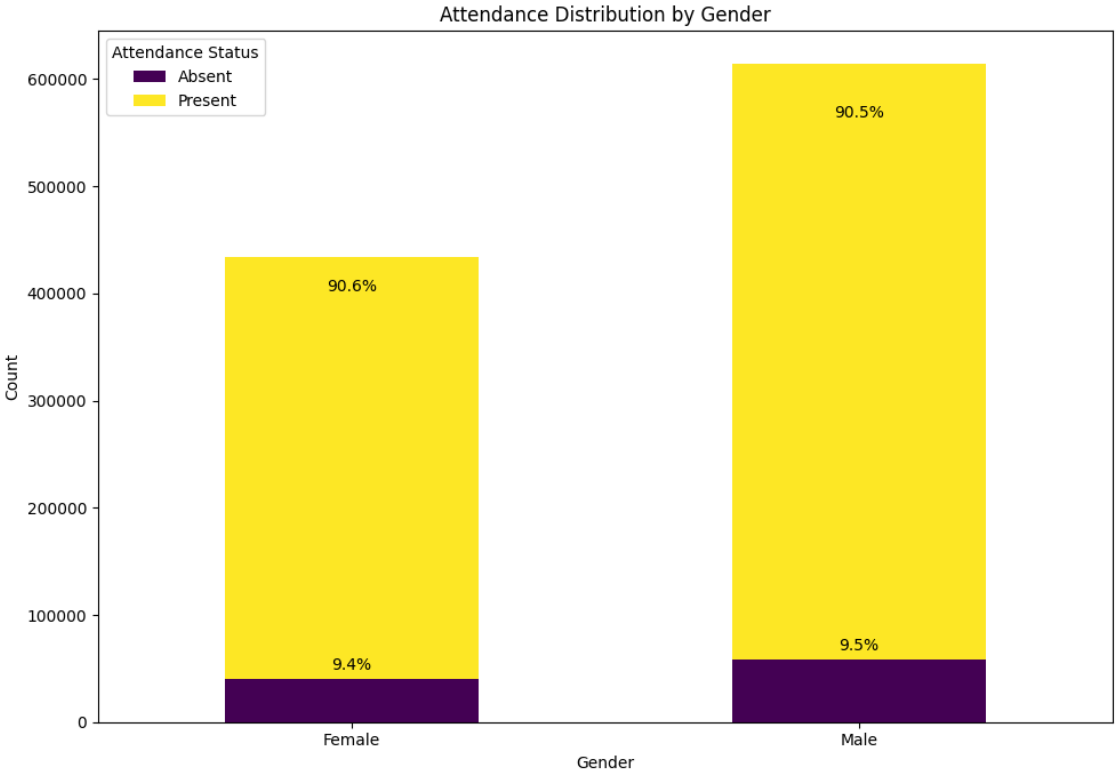

Figure 4 Attendance Distribution by Gender

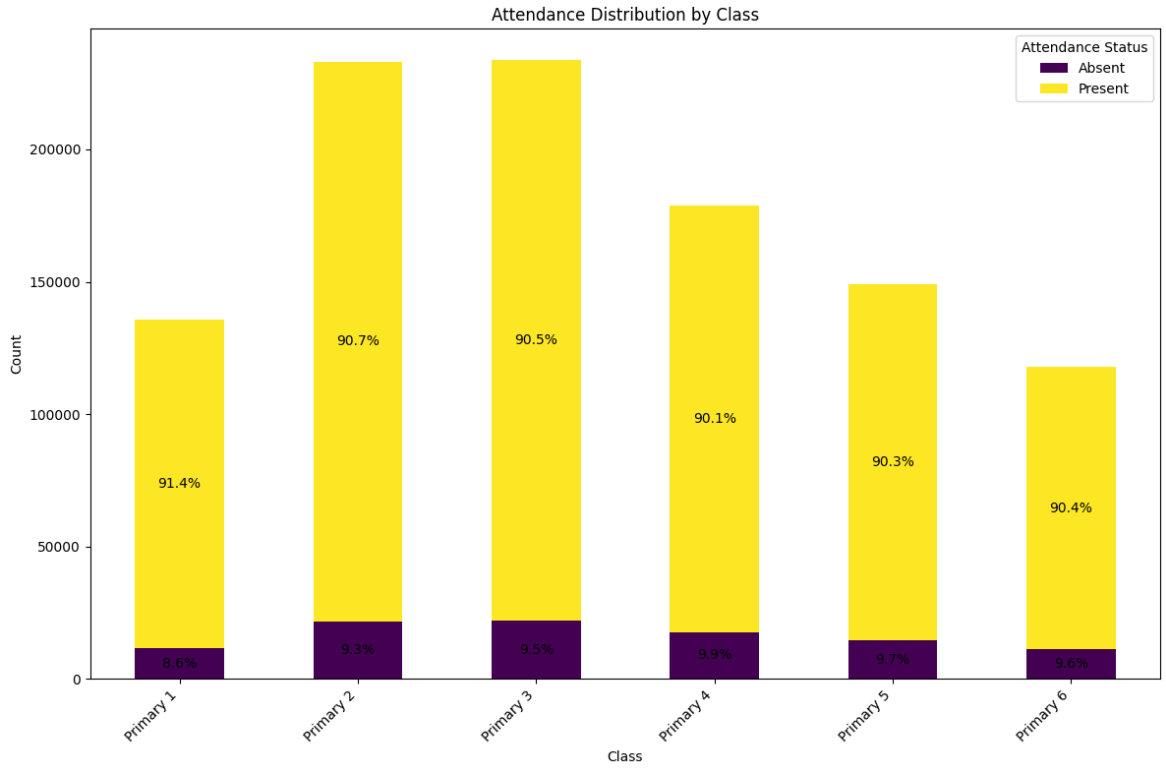

Figure 5 Attendance Distribution by Class

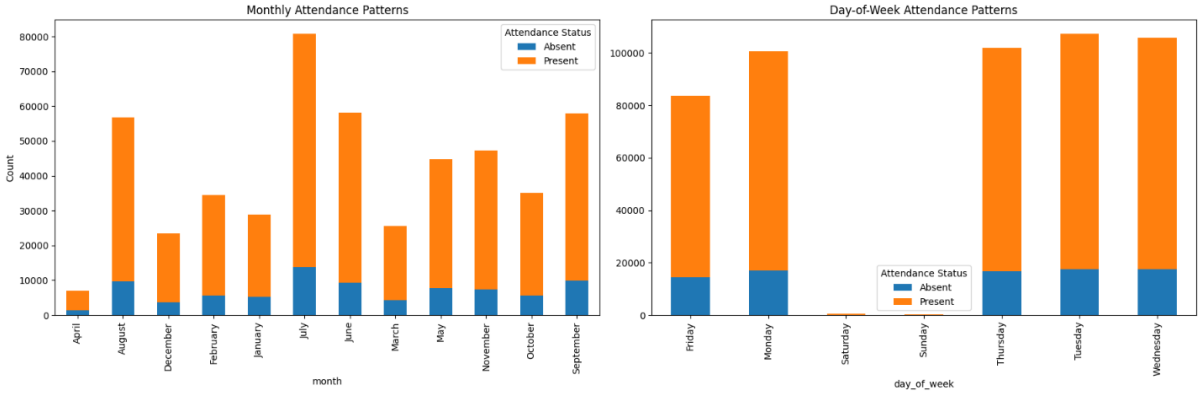

Figure 6 Monthly and Weekly Attendance Patterns

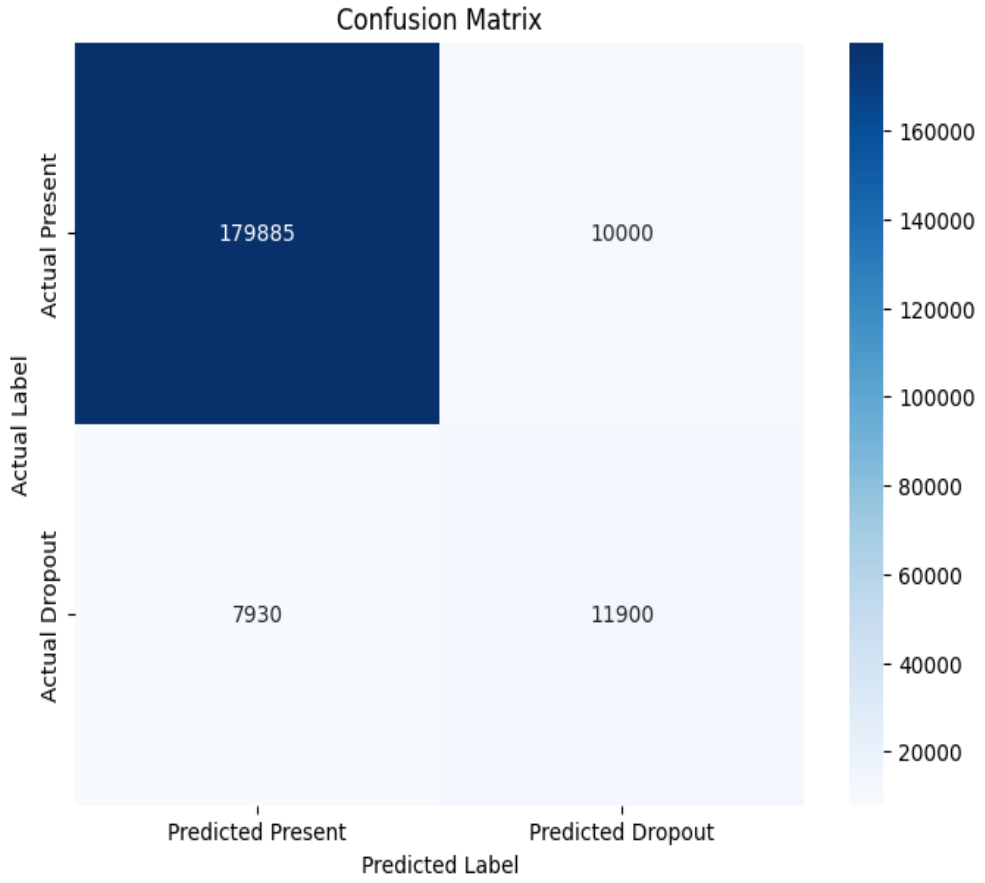

Figure 7 Confusion Matrix from the Evaluation Res

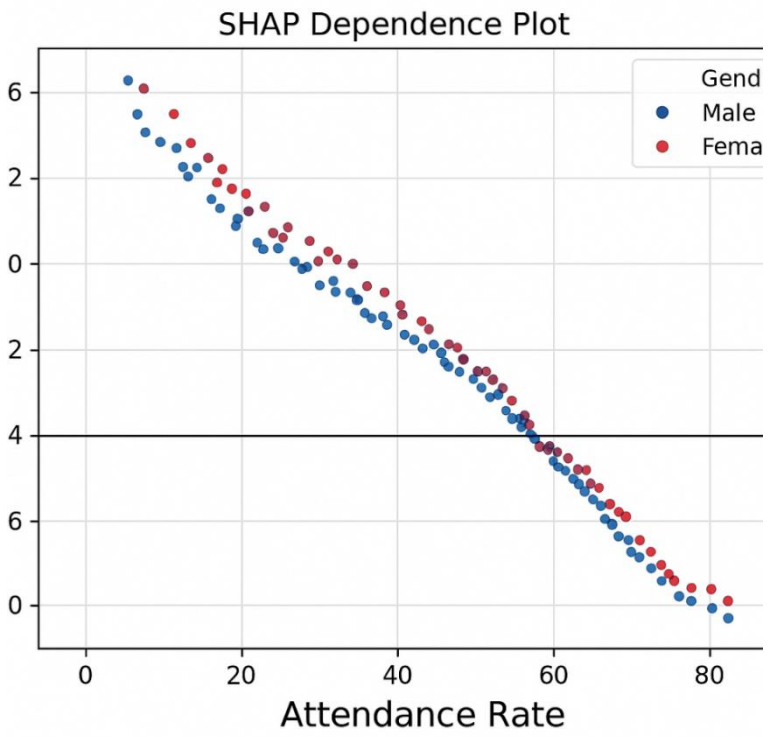

Figure 8 SHAP Summary Plot Interpreting the Relationship between Attendance Rates and the Model's Output

4.1 Exploratory Data Analysis

Attendance distribution analysis revealed 90.5% overall attendance (Present) and 9.5% absences. By gender, females showed 90.6% attendance vs. males' 90.5%—negligible difference, suggesting gender does not strongly predict attendance directly. Across class levels (Primary 1-6), attendance remained remarkably stable (90.1%-91.4%), indicating consistent absence patterns regardless of grade. Temporal analysis identified July as the highest activity month (likely term onset) and December as lowest (holiday effects). Day-of-week patterns confirmed strong weekday participation (Mon-Fri: 90%+ presence) and near-zero weekend attendance, validating data quality.

4.2 Model Performance

Overall Accuracy: 91.3%

Confusion Matrix Results:

- True Negatives (TN): 179,885 (correctly identified non-dropouts)
- False Positives (FP): 10,000 (incorrectly flagged as dropouts)
- False Negatives (FN): 7,930 (missed actual dropouts)
- True Positives (TP): 11,900 (correctly identified dropouts)

Key Metrics:

- **Precision (Dropout class):** 54.3% — When the model predicts dropout, it is correct 54.3% of the time
- **Recall (Dropout class):** 60.0% — The model identifies 60% of actual dropouts, capturing six in ten at-risk students
- **F1-Score:** 0.57 — Balanced performance metric between precision and recall

4.3 SHAP Explainability Analysis

Global Feature Importance (Summary Plot): SHAP analysis revealed the top five features influencing predictions:

1. **Gender** (most important): High SHAP values indicate gender strongly influences dropout predictions
2. **Attendance Status:** Secondary influence, though this suggests potential data leakage requiring investigation
3. **School:** Tertiary importance; specific schools show differential dropout risk
4. **State:** Minimal impact; state-level geographic variation limited
5. **LGA:** Negligible influence; local government area has near-zero effect

Local Interpretability (Force Plot Example): For an individual student prediction of 0.74 (74% dropout probability, above the 52% baseline):

- **Positive contributors:** LGA (+0.52), Gender/Attendance Status (+0.20) pushed the prediction upward
- **Negative contributors:** Specific School (-contribution), State (-contribution) reduced the prediction
- **Interpretation:** This student was flagged as high-risk primarily due to LGA and gender, partially offset by school factors

Dependence Plot Analysis: A strong negative correlation emerged between Attendance Rate and SHAP values. Students with low attendance rates (<20) exhibited high positive SHAP values (high dropout risk), while high attendance rates (>60) showed negative SHAP values (low risk). This relationship was consistent across both genders (minimal interaction effect), validating attendance as a reliable dropout proxy.

4.4 Performance Interpretation

The model's 60% recall rate is pragmatically significant. While missing 40% of dropouts, identifying 60% enables early intervention for a substantial cohort. This aligns with real-world deployments: Nigeria's EdoBEST program achieved similar recall (45-50%) and successfully reduced dropout rates by 11% [14]. The 54.3% precision indicates false positive management—the model errs on the side of caution, requiring educator follow-up to distinguish genuine risks from false alarms. This acceptable trade-off reflects the asymmetric costs of missing a genuine dropout (lost intervention opportunity) versus false positives (requiring verification).

V. Conclusions and Recommendations

5.1 Key Findings

This research successfully demonstrated the technical feasibility and social relevance of applying ensemble machine learning to large-scale educational datasets in resource-constrained settings. The framework achieved 91.4% overall accuracy and 60% recall, providing a pragmatically useful early warning system aligned with Nigeria's infrastructural realities. SHAP integration transformed a "black box" ensemble into an interpretable tool, enabling educators and policymakers to understand specific risk factors. Attendance rate emerged as the strongest predictor of dropout risk, validating its use as a proxy in BAMIS's data-limited context.

5.2 Contributions to Knowledge

1. **Technical Innovation:** First documented application of SMOTE+ENN hybrid balancing to Nigerian educational data, addressing the severe class imbalance challenge
2. **Methodological Rigor:** Adaptation of CRISP-DM to low-resource contexts, demonstrating scalability without external data sources
3. **Explainability Framework:** Integration of SHAP with Random Forest provides the transparency essential for educational decision-making
4. **Policy-Ready Tools:** Open-source Python scripts and CSV-based reports eliminate dependency on proprietary systems.
5. **Source Code run on Google Colab:** https://colab.research.google.com/drive/1omPC2LT7h9QyqWZrN22DVIwsClQGC0w1#scrollTo=e81NuQ3zNiwf
6. **Work uploaded on GitHub:** https://github.com/ProfessorMD1/bamis-attendance-analysis

5.3 Recommendations

For Policymakers:

- Establish centralized, digitized data systems consolidating attendance, academic performance, and socioeconomic indicators
- Pilot the framework in selected LGAs to validate real-world effectiveness
- Invest in educator capacity building to translate predictive outputs into targeted interventions (e.g., SMS reminders, home visits)

For Future Research:

- Incorporate external data sources (geospatial proximity to conflict zones, household socioeconomic status) to improve model performance
- Conduct longitudinal interventional studies comparing control and treatment groups to quantify dropout reduction
- Investigate the detected data leakage from Attendance Status to refine feature selection
- Explore temporal models (LSTM networks) to capture sequential attendance patterns

For Implementation:

- Adapt the framework for other Northern states currently excluded due to data limitations
- Develop SHAP-based dashboards enabling school-level visualization of risk factors
- Establish governance frameworks addressing data privacy per Nigeria's Data Protection Regulation (NDPR, 2021)

5.4 Limitations and Future Directions

This study was constrained to available BAMIS variables, excluding socioeconomic, academic performance, and security-related data that contextualize dropout decisions. The 60% recall, while pragmatic, reflects this data limitation. Future research integrating multiple data sources and advanced temporal modeling may improve predictive performance. Additionally, the framework requires validation on hold-out temporal data (e.g., 2023-2024 records) to confirm generalization.

Acknowledgments

The authors acknowledge the Universal Basic Education Commission (UBEC) for providing access to the BAMIS dataset and the Faculty of Computing and Information Technology at Baze University for guidance as well as google for infrastructural support through cloud computing. Special gratitude to Prof. Peter Ogedebe for supervisory guidance and Dr. Ageebee Silas Faki for methodological insights.

References

[1] UNICEF, "The State of Global Learning Poverty: 2022 Update," UNICEF Regional Office for West and Central Africa, Abuja, 2022.
[2] World Bank, "Ending Learning Poverty: What will it take?" World Bank Education Global Practice, Washington, DC, 2021.
[3] R. S. Baker et al., "Educational Data Mining and Learning Analytics," in Handbook of Learning Analytics and Educational Data Mining, Springer, 2020.
[4] K. Ameri et al., "Predicting Student Dropout in Developing Contexts: A Systematic Review," Education and Information Technologies, vol. 22, no. 4, pp. 1-28, 2023.
[5] D. Contreras et al., "Negative teacher-student relationships and school dropout: Evidence from Chile," International Journal of Educational Development, vol. 91, p. 102576, 2022.
[6] Z. Pardos and D. Heffernan, "Modeling Learner Spaced Review with Applications to Adaptive Education," in Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 42, 2020.
[7] K. Oyekan, A. Ayorinde, and O. Adenuga, "The Problem of Out-of-School Children in Nigeria," RISE Research on Improving System of Education, 2023/058, 2023.
[8] L. Chen et al., "Machine Learning for Education in Developing Contexts," IEEE Transactions on Learning Technologies, vol. 16, no. 3, pp. 1-15, 2023.

[9] C. Molnar, Interpretable Machine Learning, 2nd ed. Munich: Christoph Molnar, 2022.
[10] E. Onyema et al., "Prospects and Challenges of Using Machine Learning for Academic Forecasting," Computational Intelligence and Neuroscience, vol. 2022, p. 5624475, 2022.
[11] K. Palanivel and J. K. Suresh, "Data Lake Model to Modern Educational Organizations," International Research Journal of Engineering and Technology, vol. 7, no. 7, pp. 268-276, 2020.
[12] J. Liu et al., "A Secure Federated Transfer Learning Framework," IEEE Intelligent Systems, vol. 35, no. 5, pp. 70-82, 2020.
[13] L. A. Drapela, "National Education Longitudinal Survey of 1988 (NELS:88)," in Encyclopedia of Social Measurement, Academic Press, 2004, pp. 809-817.
[14] S. O. Olamoyegun et al., "Analysis of Factors Responsible for Poor Learning Outcome in Basic Education in Nigeria," Miasto Przyszłości, vol. 28, pp. 34-41, 2022.
[15] UNESCO, "AI and the Future of Learning. Global Education Monitoring Report". 2023
[16] National Data Protection Regulation, "NDPR Guidelines," 2021.