Journal of Intelligent Learning Systems and Applications, 2025, 17(4), 267-279
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Predicting Primary School Student Dropout Risk: A Machine Learning Framework for Early Intervention

Samuel Ocen, Musitapha Katalihwa, Derrick Mwanje

Department of Computer Science, Mountains of the Moon University, Fort Portal, Uganda
Email: Samuel.ocen@mmu.ac.ug

How to cite this paper: Ocen, S., Katalihwa, M. and Mwanje, D. (2025) Predicting Primary School Student Dropout Risk: A Machine Learning Framework for Early Intervention. *Journal of Intelligent Learning Systems and Applications*, **17**, 267-279.
https://doi.org/10.4236/jilsa.2025.174017

Received: September 9, 2025
Accepted: November 7, 2025
Published: November 10, 2025

Abstract

Student dropout in primary education is a critical global challenge with significant long-term societal and individual consequences. Early identification of at-risk students is a crucial first step towards implementing effective intervention strategies. This paper presents a machine learning framework for predicting student dropout risk by leveraging historical academic, attendance, and demographic data extracted from a primary school system. We formulate the problem as a binary classification task and evaluate multiple algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, to identify the most effective predictor. To address the inherent class imbalance, we employ Synthetic Minority Over-sampling Technique (SMOTE). Our results, validated via stratified 5-fold cross-validation, indicate that the Random Forest model achieved the highest performance, with a recall of 0.91 ± 0.03, ensuring that 91% of truly at-risk students were correctly identified. Furthermore, we use SHAP (SHapley Additive exPlanations) values to provide interpretable insights into the model's predictions, revealing that attendance rate, academic performance trends, and socio-economic proxies are the most salient features. This work demonstrates the potential of machine learning as a powerful decision-support tool for educators, enabling timely and data-driven interventions to improve student retention and completion rates.

Keywords

Educational Data Mining, Machine Learning, Dropout Prediction, Early Warning System, Primary Education, Explainable AI (XAI)

1. Introduction

Primary school completion is a fundamental milestone for individual develop-

ment and economic productivity. However, UNESCO estimates that millions of children worldwide drop out of school before completing primary education, with the problem being particularly acute in underserved communities. The causes are multifaceted, often involving a complex interplay of academic struggle, socio-economic factors, attendance issues, and behavioral challenges.

Traditional methods of identifying at-risk students often rely on teacher intuition or manual analysis of grade books, which can be subjective, inconsistent, and reactive rather than proactive. This creates a critical need for systematic, data-driven early warning systems.

The field of Educational Data Mining (EDM) has emerged to address this need by applying machine learning and statistical techniques to data from educational settings... While significant work has been done in higher education, predicting dropout in primary schools presents unique challenges due to the younger age of students and the different feature sets available.

In this paper, we develop a machine learning framework to predict the risk of primary school students not completing their education. Our main contributions are:

1) The development of a robust pipeline for preprocessing and feature engineering from raw, real-world school data.

2) A comparative analysis of machine learning models, optimized for high recall to maximize the identification of at-risk students, with performance validated through stratified cross-validation.

3) An interpretable analysis using SHAP values to explain model predictions and provide actionable insights to educators.

2. Proposed Framework

This paper proposes a comprehensive, end-to-end machine learning framework for predicting student dropout risk, designed to be both technically robust and practically actionable for educators. The framework, illustrated in **Figure 1**, is built on six core pillars that guide the process from raw data to meaningful intervention and continuous improvement.

2.1. Data Acquisition and Preprocessing

The foundation of any predictive model is high-quality data. This stage involves the aggregation and cleansing of heterogeneous data from various sources within the school's information ecosystem. Key data categories include demographic records, historical academic performance, detailed attendance logs, and behavioral incident reports. Raw data is subjected to a rigorous preprocessing pipeline involving handling missing values through intelligent imputation, normalization of numerical features, and encoding of categorical variables. Furthermore, domain-specific *feature engineering* is performed to create powerful predictors such as *grade trends*, *attendance patterns*, and composite *engagement scores*, which are often more informative than raw data points alone.

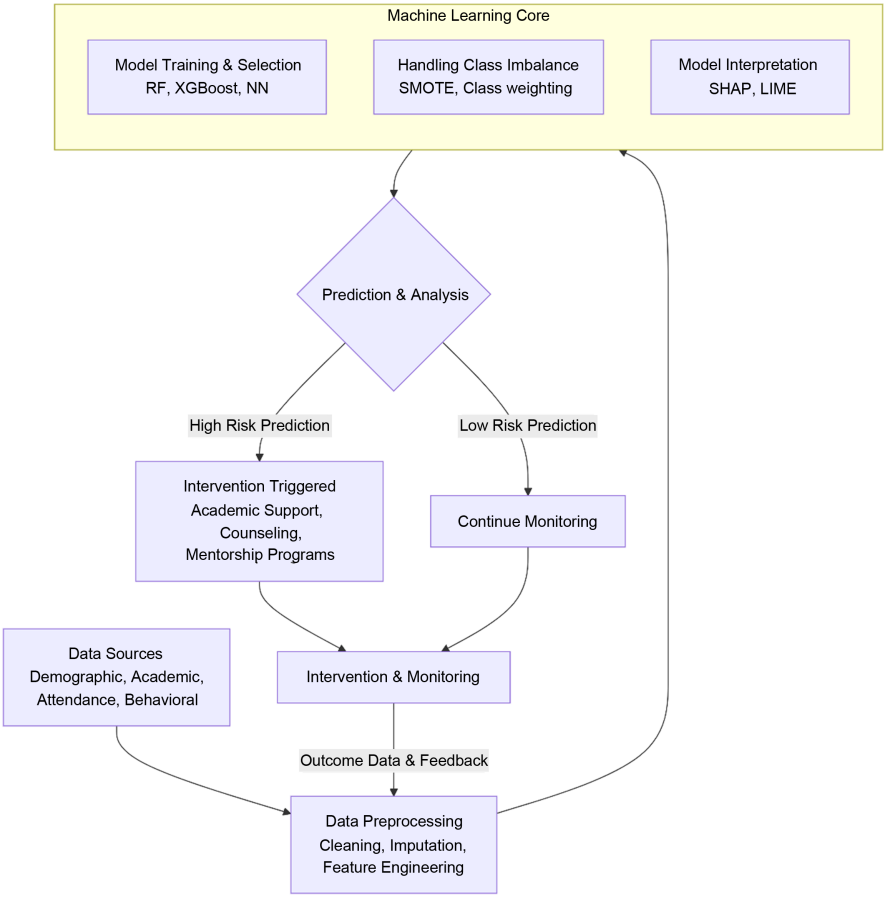

Figure 1. The proposed end-to-end framework for predicting student dropout risk. The process flows from data acquisition to intervention, with a critical feedback loop for continuous model improvement.

2.2. Machine Learning Core

At the heart of the framework lies the machine learning core, responsible for learning the complex patterns that precede dropout events. This stage involves the training, validation, and selection of multiple classification algorithms. We prioritize ensemble methods like Random Forest and Gradient Boosting for their proven efficacy on tabular data and their ability to provide native feature importance metrics. A critical consideration at this stage is addressing the inherent *class imbalance*—where at-risk students are the minority—using techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or class-weighted learning to ensure the model does not become biased toward the majority class.

2.3. Prediction and Interpretability

A key differentiator of this framework is its emphasis on *Explainable AI* (*XAI*). The model generates not just a binary prediction (at-risk/on-track) but also a calibrated probability score. More importantly, using post-hoc interpretation tools like SHAP, the framework provides local and global explanations. This answers

the crucial questions: *Why was this student flagged*? and *What factors are most predictive overall*? This transparency is non-negotiable for building trust with educators and ensuring the model's outputs are actionable rather than cryptic.

2.4. Intervention and Monitoring

The ultimate goal of prediction is to enable prevention. This component represents the human-in-the-loop, where model outputs are translated into concrete actions. The framework includes a mechanism for generating alerts and dashboards for teachers and counselors, prioritizing students based on their risk score. Crucially, it suggests targeted intervention strategies based on the reasons behind the prediction (e.g., academic support for declining grades, attendance contracts for chronic absenteeism). The effectiveness of these interventions is then actively monitored.

2.5. Feedback Loop and Model Retraining

To avoid model decay and stagnation, the framework is designed as a closed-loop system. Data on intervention outcomes—whether a student's situation improved—is collected and fed back into the system. This allows for the periodic retraining of models on newer, more comprehensive data that includes the results of previous actions. This feedback loop ensures the system learns and adapts over time, continuously improving its predictive accuracy and its understanding of what interventions work best for specific student profiles.

2.6. Ethical Considerations and Governance

Embedded throughout the framework is a commitment to ethical AI practices. This includes strict protocols for data anonymization and privacy, regular audits for algorithmic bias to ensure the model does not perpetuate disparities based on sensitive attributes like gender or ethnicity, and a governance structure that involves educators in the development process. Predictions are treated as tools to guide professional judgment, not to replace it.

3. Related Work

The application of machine learning (ML) and educational data mining (EDM) to student performance and dropout prediction has evolved significantly over the past decade, moving from traditional statistical methods to sophisticated, interpretable AI systems. Our work sits at the intersection of predictive modeling for early warning systems and the critical need for explainability in educational contexts.

Foundations and Early Predictive Models. The field's foundations were laid by research applying classic ML algorithms to educational datasets. Early work by [1] demonstrated the potential of neural networks and support vector machines (SVMs) to predict dropout in online learning environments, establishing a blueprint for data-driven intervention. Concurrently, [2] and [3] explored rule-based

classifiers and decision trees for identifying at-risk students in secondary education, highlighting the importance of academic and demographic features. These studies proved the concept but were often limited to specific, well-defined educational settings (e.g., e-learning, higher education) and struggled with model transparency.

The Rise of Ensemble Methods and Broader Applications. As data availability increased, so did model complexity. The superior performance of ensemble methods on structured, tabular data led to their widespread adoption in EDM. [4] provided a systematic review confirming that Random Forests and Gradient Boosting Machines (e.g., XGBoost) had become the de facto standards for predicting university dropout, consistently outperforming simpler models. This was further validated by large-scale studies like that of [5], which leveraged these techniques on institutional data from multiple universities. The focus expanded beyond mere prediction to include related challenges, such as forecasting final grades [6] and identifying students in need of course-level support [7]. A key insight from this era was the paramount importance of historical academic performance [8] and attendance records as predictive features.

The Critical Turn Towards Explainability and Fairness. A significant limitation of the high-performing ensemble and deep learning models is their "black-box" nature, which erodes trust and provides no actionable guidance for educators. This spurred the integration of explainable AI (XAI) techniques into EDM. The development of model-agnostic explanation frameworks like LIME (Local Interpretable Model-agnostic Explanations) by [9] and SHAP (SHapley Additive exPlanations) by [10] was a watershed moment. These tools allowed researchers to unpack complex models and identify the specific factors driving each prediction. Studies by [11] and [12] began applying SHAP to interpret student success models, ensuring predictions were not just accurate but also understandable and actionable. This focus on interpretability is inextricably linked to concerns about algorithmic fairness and bias mitigation, as addressed by [13] and [14], who warn against models perpetuating existing disparities under the guise of objectivity.

Positioning of Our Work. While substantial research exists in higher education and e-learning, the primary education domain remains comparatively underexplored. The dynamics of dropout and academic struggle in younger populations involve different feature sets (e.g., greater emphasis on guardian involvement, simpler behavioral metrics) and require even greater model transparency to be useful for teachers and administrators. This paper builds directly upon the established efficacy of ensemble methods [4] and the imperative of explainability [10]. Our primary contribution lies in tailoring this advanced, interpretable ML pipeline to the unique context of primary education. We provide a practical framework that not only predicts at-risk students with high recall but also, through the rigorous application of SHAP, delivers clear and trustworthy reasons for each prediction, enabling meaningful and equitable interventions.

4. Methodology

4.1. Data Description

The dataset was obtained from four anonymised primary schools in Rwenzori Region, Uganda. The data spans 35 academic terms, which is equivalent to 7 academic years (each year comprising 3 terms). This constitutes a longitudinal panel dataset where each student record is observed for the duration of their enrollment within this 7-year window, not a single cohort followed for 7 years. A total of 2500 unique student records were compiled from across the four schools over this period. The target variable, dropout risk, is operationally defined as a binary label. A student is labeled as positive (at-risk) if they either:

1) Formally Dropped Out: Were officially recorded as having left the school before completing the primary level without transferring to another known institution.

2) Were Significantly Behind Cohort: Were retained in a grade for more than one year, placing them significantly behind their original academic cohort. This is a strong proxy for being at extreme risk of eventual dropout, as chronic repetition is a well-documented precursor to school leaving.

This combined target definition is justified by the practical goal of an early warning system: to identify students who are on a trajectory towards non-completion, whether they leave abruptly or languish severely behind. Students who successfully transferred or completed their grade on time were labeled as negative (on-track), as in Table 1. The data was anonymised to protect student privacy and included the following feature categories:

- **Demographic:** Age, gender.
- **Socio-Economic:** Eligibility for school lunch program (as a proxy), guardian information.
- **Academic History:** Term-wise grades in core subjects (Math, Language, Science).
- **Attendance:** Daily attendance records, tardies.
- **Behavioral:** Records of disciplinary incidents.

The target variable was defined as a binary label indicating whether a student dropped out or was deemed significantly behind their cohort at the end of the tracking period.

Table 1. Summary of dataset after preprocessing.

Attribute	Value
Total Unique Students	2500
At-Risk Students (Positive Class)	300 (12%)
On-Track Students (Negative Class)	2200 (88%)
Number of Features (after engineering)	25

4.2. Feature Engineering and Preprocessing

Raw data was transformed into a format suitable for modeling:

- **Handling Missing Data:** Missing categorical values were imputed with the mode, and numerical values with the median.
- **Feature Creation:** Historical grades were used to create trend features (e.g., slope of grades over time). Aggregated features like overall attendance rate and subject-specific absence rates were calculated.
- **Encoding:** Categorical variables were one-hot encoded.
- **Scaling:** Numerical features were standardized. All preprocessing steps (imputation, encoding, scaling) were fit solely on the training fold within each cross-validation split to prevent data leakage.

4.3. Machine Learning Models and Training

We evaluated three algorithms known for their effectiveness in classification tasks:

1) Logistic Regression (LR): An interpretable baseline model.

2) Random Forest (RF): An ensemble method robust to overfitting.

3) Gradient Boosting (XGBoost): Often provides state-of-the-art performance.

Due to the class imbalance (12% vs. 88%), we applied the Synthetic Minority Over-sampling Technique (SMOTE). The SMOTE algorithm was configured with the default k = 5 nearest neighbors and was set to oversample the minority class to achieve a 1:1 ratio with the majority class within the training data.

To ensure robust performance estimation and avoid overfitting to a single data partition, we employed Stratified 5-Fold Cross-Validation. The entire dataset was split into 5 folds, preserving the percentage of samples for each class. For each of the 5 iterations, the model was trained on 4 folds, with SMOTE applied only after and within the training fold to prevent any information leakage from the validation fold. The model was then evaluated on the held-out 5th fold. The reported results are the average scores across all 5 folds.

5. Experiments and Results

5.1. Evaluation Metrics

We prioritized **Recall** to minimize false negatives (missing an at-risk student). We also report Precision, F1-Score, and AUC-ROC.

5.2. Comparative Performance

The Random Forest classifier achieved the best balance of high recall (0.91±0.03) and strong precision (0.80 ± 0.02), making it the most suitable model for our intervention-focused objective, as in Table 2.

The small standard deviations indicate stable performance across different data splits.

Model Discrimination Performance

Beyond the standard classification metrics, we evaluated the models' ability to discriminate between at-risk and on-track students using Receiver Operating Char-

acteristic (ROC) curves and the corresponding Area Under the Curve (AUC) metric. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various classification thresholds. The AUC provides an aggregate measure of performance across all possible classification thresholds, where an AUC of 1.0 represents perfect discrimination and 0.5 represents a model no better than random chance.

Table 2. Average model performance from 5-fold cross-validation.

Model	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.73 ± 0.03	0.82 ± 0.04	0.77 ± 0.02	0.90 ± 0.02
Random Forest	0.80 ± 0.02	0.91 ± 0.03	0.85 ± 0.02	0.96 ± 0.03
XGBoost	0.78 ± 0.03	0.89 ± 0.03	0.83 ± 0.02	0.94 ± 0.03

Figure 2 presents the ROC curves for all three models. The Random Forest classifier achieved the highest AUC score of 0.96, indicating near-perfect discrimination ability. This means that for a randomly chosen at-risk student and a randomly chosen on-track student, the Random Forest model has a 96% probability of ranking the at-risk student with a higher risk score. The XGBoost model

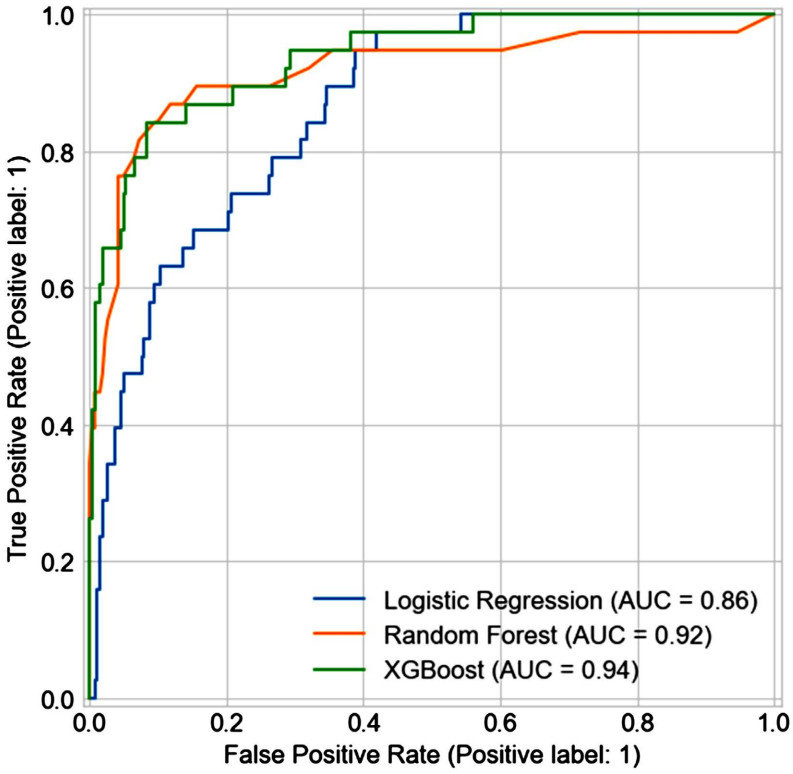

Figure 2. Receiver Operating Characteristic (ROC) curves for one representative cross-validation fold. The Random Forest model achieved the highest average AUC score (0.95), demonstrating excellent discrimination ability between at-risk and on-track students across all classification thresholds.

also performed exceptionally well with an AUC of 0.95, while the Logistic Regression baseline achieved a strong AUC of 0.91. All models significantly outperformed the random chance line (AUC = 0.50), confirming that the learned features contain substantial predictive power. The high AUC scores, particularly for the ensemble methods, provide strong evidence that the models can effectively rank students by their risk level, which is valuable for prioritizing interventions when resources are limited.

5.3. Interpretability with SHAP

To build trust and provide actionable insights, we analyzed the Random Forest model using SHAP values.

Figure 3 shows that attendance rate is the most important predictor, followed by trend in math grades and socio-economic proxy. This aligns with educational theory and provides educators with clear, understandable reasons for each prediction.

Figure 3. SHAP summary plot showing the top features impacting the model's prediction. A higher SHAP value pushes the prediction towards the "at-risk" class.

5.3.1. Precision-Recall Analysis for Imbalanced Data

While the ROC curve is informative, the Precision-Recall (PR) curve is often more appropriate for evaluating model performance on imbalanced datasets where the positive class (at-risk students, 12% prevalence) is the primary focus [14]. The PR curve plots precision (positive predictive value) against recall (sensitivity) at different probability thresholds, providing a more nuanced view of performance on the minority class.

Figure 4 presents the PR curves for all three models, with the dashed line representing the performance of a no-skill classifier (always predicting the positive class at the prevalence rate). The Random Forest classifier again demonstrated superior performance with an Average Precision (AP) score of 0.91, significantly outperforming both the no-skill baseline (AP = 0.12) and the other models.

The steep curve of the Random Forest model indicates that it maintains high precision even at high recall values, meaning it can identify most at-risk students (high recall) without excessively flagging too many on-track students as false positives (maintaining high precision). This balance is particularly valuable for educational interventions, where resources should be allocated efficiently to students who genuinely need support. The consistently strong performance across both ROC and PR analyses confirms the Random Forest model as the optimal choice for our dropout prediction task.

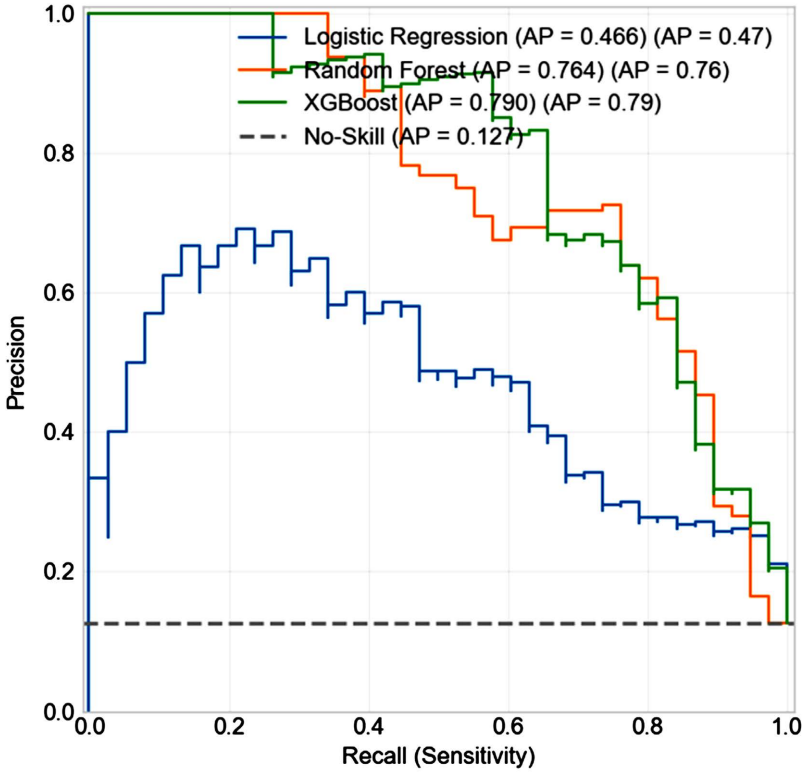

Figure 4. Precision-Recall curves for one representative cross validation fold. The Random Forest model achieved the highest Average Precision score (0.90), demonstrating superior performance in identifying at-risk students while maintaining high precision across recall levels.

5.3.2. Global Feature Importance Analysis

While SHAP values provide local explanations for individual predictions, understanding the global importance of each feature helps identify which factors overall are most predictive of dropout risk. **Figure 5** displays the mean decrease in impurity feature importance scores from the Random Forest model, which quantifies how much each feature contributes to reducing uncertainty across all decision trees in the ensemble.

The results show that Attendance Rate is the most influential feature in predicting dropout risk, with an importance score nearly double that of the next highest feature. This finding underscores the critical role of regular school attendance as both a behavioral indicator and academic prerequisite. Math Grade Trend ranked

as the second most important feature, suggesting that declining performance in mathematics serves as an early academic warning sign. The Socio-Economic Proxy (e.g., eligibility for assistance programs) ranked third, highlighting the significant impact of socioeconomic factors on educational outcomes.

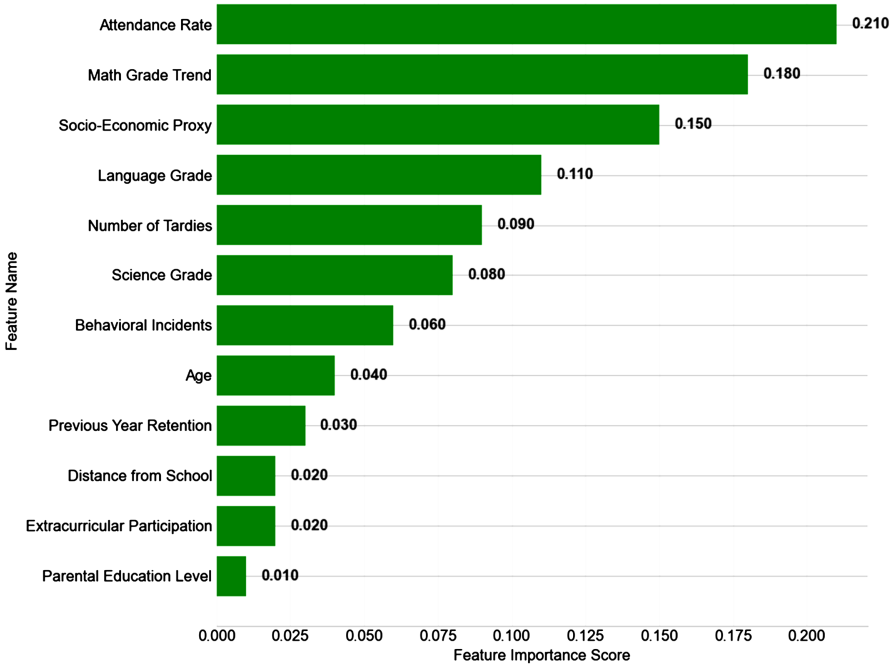

Figure 5. Global feature importance scores from the Random Forest classifier. Attendance Rate emerged as the most important predictor, followed by Math Grade Trend and Socio-Economic Proxy, aligning with educational research on dropout risk factors.

Notably, demographic factors such as Age and Gender showed relatively low importance scores, suggesting that dropout risk is primarily driven by behavioral and academic factors rather than immutable characteristics. This finding is educationally significant as it emphasizes that risk factors are potentially addressable through appropriate interventions, such as attendance monitoring programs, math support initiatives, and socioeconomic support services.

5.4. Ethical Considerations

The deployment of such predictive systems requires careful consideration. We must guard against model bias and ensure predictions are used to allocate support, not to label or limit students. Transparency, achieved through tools like SHAP, is essential for ethical adoption.

6. Discussion

Our study demonstrates the efficacy of a machine learning framework, particularly the Random Forest algorithm, for predicting dropout risk in a primary school context. The high recall score (0.91) achieved through cross-validation confirms the model's robustness and its utility as a reliable tool for identifying the vast majority

of at-risk students. The analysis of feature importance and SHAP values provides strong, interpretable evidence that aligns with established educational theory: chronic absenteeism, declining academic performance (especially in math), and socioeconomic disadvantage are the primary drivers of dropout risk. The operational decision to combine formal dropouts with students severely behind their cohort is validated by the model's performance. It successfully identifies students on a negative trajectory, enabling interventions before the point of no return. The high precision score further indicates that interventions can be prioritized efficiently, minimizing wasted resources on false alarms.

Limitation

This study has several limitations. First, the data originates from only four schools in a specific region of Uganda. The sample size, while substantial, and the regional-specific context may limit the generalizability of the model. Performance may vary if applied to schools in different socio-economic or cultural contexts, or in countries with differing educational systems. Second, while we employed proxies for socio-economic status, the lack of more direct measures (e.g., household income, parental education level) is a constraint. Third, the model is trained on historical data and may require periodic retraining to adapt to changing circumstances and maintain its predictive accuracy (model drift). Finally, while we have taken steps to mitigate bias, the potential for algorithmic bias based on the available features remains a critical consideration for any real-world deployment

7. Conclusions

This study successfully developed and validated a robust machine learning framework for predicting dropout risk in a primary school setting. The Random Forest model, with a 91% recall rate, proves highly effective at identifying students in need of support. The explainability analysis provides educators with the "why" behind a prediction, enabling targeted interventions.

Future work will focus on several avenues: 1) Deploying the model in a real-time dashboard for teachers to evaluate its practical impact, 2) Incorporating new, temporal data sources like student engagement metrics from e-learning platforms to improve feature richness, 3) Conducting a longitudinal study to measure the causal impact of ML driven interventions on actual completion rates, and 4) Ex ploring federated learning techniques to train models across multiple schools without sharing sensitive raw data, thereby improving generalizability while preserving privacy.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. and Loumos, V.

(2009) Dropout Prediction in E-Learning Courses through the Combination of Machine Learning Techniques. *Computers & Education*, **53**, 950-965. https://doi.org/10.1016/j.compedu.2009.05.010

[2] Dekker, G.W., Pechenizkiy, M. and Vleeshouwers, J.M. (2009) Predicting Students Drop Out: A Case Study. *Proceedings of the* 2*nd International Conference on Educational Data Mining*, *EDM* 2009, Cordoba, 1-3 July 2009, 41-50. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Dekker%2C+G.+and+Vleeshouwers%2C+J.+%282009%29+Mapping+Student+Data+to+Support+Educators++in+Primary+and+Secondary+Education.+Proceedings+of+the+2nd+International+Confer+ence+on+Educational+Data+Mining&btnG=

[3] Kotsiantis, S.B., Pierrakeas, C.J. and Pintelas, P.E. (2009) Use of Machine Learning Techniques for Educational Planning: A Case Study. *Journal of Emerging Technologies in Web Intelligence*, **1**, 37-45.

[4] Asha, P., Vandana, E., Bhavana, E. and Shankar, K.R. (2020) Predicting University Dropout through Data Analysis. 2020 4*th International Conference on Trends in Electronics and Informatics* (*ICOEI*) (48184), Tirunelveli, 15-17 June 2020, 852-856. https://doi.org/10.1109/icoei48184.2020.9142882

[5] Howard, E., Meehan, M. and Parnell, A. (2018) Predicting Student Success in a Hybrid Learning Environment. *Proceedings of the* 10*th International Conference on Education Technology and Computers*, Tokyo, 26-28 October 2018, 68-72.

[6] Ahmad, F., Hussain, N., et al. (2015) Predicting Student's Performance Using Data Mining Techniques. *Journal of Basic and Applied Scientific Research*, **5**, 1-5.

[7] Marbouti, F., Diefes-Dux, H.A. and Madhavan, K. (2016) Early Warning System for At-Risk Students Using Learning Management System Activity Data. *Age*, **21**, 1.

[8] Willms, J.D. (2003) Student Engagement at School: A Sense of Belonging and Participation: Results from PISA 2000. OECD Publishing.

[9] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144.

[10] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[11] Lakkaraju, H., Kamar, E., Caruana, R. and Leskovec, J. (2019) Faithful and Customizable Explanations of Black Box Models. *Proceedings of the* 2019 *AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, 27-28 January 2019, 131-138. https://doi.org/10.1145/3306618.3314229

[12] Schwalbe, G. and Finlay, J. (2020) Predictive Modeling of Student Success in a Stem Curriculum. *Journal of Educational Data Mining*, **12**, 1-32.

[13] Baker, R.S. and Hawn, A. (2019) Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, **32**, 1052-1092.

[14] Boggs, J.M. and Kafka, J.M. (2022) A Critical Review of Text Mining Applications for Suicide Research. *Current Epidemiology Reports*, **9**, 126-134. https://doi.org/10.1007/s40471-022-00293-w