IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 4, Ser. 1 (Jul. – Aug. 2025), PP 18-23
www.iosrjournals.org

Student Performance Prediction Using Machine Learning in Central Africa

Novy Margepaule BAFOUKA de MABIALA
School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou, China

Abstract
This paper presents a machine learning–based approach for predicting student academic performance using structured educational data. The study utilizes the UCI Student Performance dataset, which includes a diverse range of academic, demographic, and behavioral features. After preprocessing and feature transformation, three classification algorithms—Logistic Regression, Random Forest, and Support Vector Machine—were trained and evaluated using metrics such as accuracy, F1 score, and ROC AUC. Among the models, Logistic Regression demonstrated the highest performance, with a balance of simplicity, interpretability, and accuracy. The selected model was deployed using Flask in a web-based prediction system, offering educators a practical tool to assess student outcomes and identify those at academic risk. The results support the use of machine learning in educational analytics and decision-making.

Keywords:
Student Performance, Africa, Machine Learning, Logistic Regression, Random Forest, SVM, Educational Data Mining, Prediction System, Academic Analytics, Flask, Classification Models

---

Date of Submission: 02-07-2025 Date of Acceptance: 12-07-2025

I. Introduction

1.1 Background

Academic performance plays a vital role in determining a student's future opportunities, including access to higher education, scholarships, and career prospects. Understanding and anticipating student outcomes is a growing area of interest within education systems around the world. With the expansion of data collection in schools, it has become possible to apply machine learning techniques to identify patterns and predict student success or failure.

Traditionally, African educators assess student performance through past test scores, teacher evaluations, and personal judgment. However, these methods can be subjective, limited in scope, and time-consuming. Predictive modeling using machine learning offers a more systematic and scalable solution. It allows for analyzing multiple variables simultaneously — such as study habits, previous grades, attendance, and African social background — to forecast a student's likely academic outcome.

This project focuses on using machine learning models to analyze historical data and provide accurate predictions on African student performance in Central Africa. It also includes the development of a web-based system that allows educators to interact with the model and use it for real-time decision-making.

1.2 Objectives

This project is designed to achieve the following goals:

- Build a reliable predictive model to classify students into 'pass' or 'fail' categories.
- Evaluate and compare several machine learning algorithms to identify the most effective one for Central Africa.
- Understand which features most strongly influence African student performance.
- Develop a web interface using Flask to make the prediction system user-friendly for African institutions.
- Use visual tools to explain model behavior, such as feature importance charts, ROC curves, and confusion matrices.

1.3 Motivation

The motivation for this project arises from the need to make academic monitoring more efficient and insightful in Republic of Congo (also known as Congo Brazzaville), Gabon, DRC and Cameroon. African educational institutions often struggle to support every student individually due to limited resources. By predicting potential academic risks early, African schools can provide targeted assistance, improve learning experiences, and boost African student retention.

Additionally, this project showcases how technical knowledge in machine learning can be applied to solve practical problems in education. It emphasizes the value of combining data-driven methods with human insight to foster better learning environments in Central Africa mainly in my hometown, Republic of Congo.

II. Literature Review

Predictive analytics in education has emerged as a crucial research area over the past decade, leveraging student data to forecast academic outcomes and inform interventions. Several studies have applied various machine learning algorithms to understand patterns in student behavior and academic performance.

One of the foundational works in this field introduced the concept of using student demographic, behavioral, and academic data for early prediction of outcomes. Researchers demonstrated that features like study time, parental education in Africa, and prior grades had a strong correlation with final academic success. These findings laid the groundwork for models that move beyond basic averages to include more nuanced, multi-dimensional data.

Subsequent studies explored a variety of algorithms such as Decision Trees, Naïve Bayes, Support Vector Machines (SVM), and Random Forests. Among these, ensemble methods like Random Forests often outperformed single classifiers due to their ability to reduce variance and handle non-linear interactions between features.

Logistic Regression, though simpler, has been frequently used as a baseline for classification tasks, offering interpretability and efficiency.

Some researchers have also investigated the integration of socio-economic and psychological factors in predictive modeling. This includes variables such as internet access at home, romantic relationships, or parental involvement — which are often overlooked in traditional academic evaluations. These holistic approaches reflect a shift toward understanding the broader context of student life.

Moreover, the rise of educational data mining platforms and open-access datasets — like the UCI Student Performance dataset used in this project — has enabled researchers to benchmark their models more effectively. These datasets offer diverse attributes, making them ideal for experimentation with various algorithms and validation techniques.

Recent advancements also include the application of deep learning and neural networks, though these require significantly more data and computational resources. While powerful, such models may not always be practical for institutions with limited access to large-scale student data.

In conclusion, the literature suggests that machine learning can be a powerful tool in African educational contexts when applied thoughtfully. By incorporating both academic and non-academic factors, predictive models can provide meaningful insights and support African systems to improve student outcomes.

III. Methodology

3.1 Data Description

This project utilizes the UCI Student Performance dataset, which contains 395 student records and 33 features. These features include demographic attributes (e.g., age, gender, address), academic performance (e.g., grades G1, G2, G3), and personal/social variables (e.g., parental education, romantic relationships, internet access).

The target variable, G3, is the final grade, which has been converted into a binary classification problem: students are labeled 'Pass' if G3 ≥ 10, and 'Fail' otherwise.

3.2 Data Preprocessing

Before training the models, several preprocessing steps were performed:
- Target Encoding: The final grade (G3) was transformed into a binary target.
- Label Encoding: All categorical features (e.g., 'sex', 'address') were encoded using label encoders.
- Missing Values: The dataset contained minimal missing values and was cleaned accordingly.
- Train/Test Split: Data was split into 80% training and 20% testing sets for evaluation.

3.3 Model Selection

Three supervised machine learning models were selected and compared based on performance:
- Random Forest Classifier: An ensemble method known for handling overfitting and ranking feature importance.
- Logistic Regression: A simple, interpretable model useful as a baseline.
- Support Vector Machine (SVM): A classifier capable of finding non-linear decision boundaries.

3.4 Model Development

The Random Forest model was trained with 100 trees and default settings. The models were evaluated using the test dataset. The feature importance graph below highlights the most influential factors in predicting student performance.

G1 and G2, which are prior grades, emerged as the most predictive features, followed by study time and past failures.

3.5 Evaluation Metrics

The models were assessed using standard classification metrics:

- Accuracy: Proportion of correct predictions over total cases.
- F1 Score: Harmonic mean of precision and recall.
- Confusion Matrix: Shows true positives, false positives, false negatives, and true negatives.
- ROC Curve & AUC Score: Measure classifier's ability to distinguish between classes across thresholds.

IV. Machine Learning Algorithms

4.1 Logistic Regression

Logistic Regression is a linear classifier commonly used for binary classification problems. It models the probability that a given input belongs to a specific class using a logistic function.

Despite its simplicity, it performs well on linearly separable datasets and provides high interpretability through feature coefficients.

Its formula is defined as:

$$P(y=1) \;=\; \frac{1}{(1 \;+\; e^{\wedge} - (\beta_0 \;+\; \beta_1 x_1 \;+\; \ldots \;+\; \beta_{\mathrm{n}} x_{\mathrm{n}}))}$$

4.2 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to produce a more accurate and robust prediction. It reduces overfitting and improves generalization by training each tree on a random subset of features and samples.

Feature importance ranking is an added benefit, allowing deeper insights into which variables most influence predictions.

4.3 Support Vector Machine (SVM)

SVM is a powerful algorithm that finds the optimal hyperplane to separate data points of different classes. For non-linearly separable data, SVM uses kernel functions to project data into higher dimensions where a linear separation becomes feasible.

It is particularly effective for small-to-medium sized datasets and maintains a good balance between bias and variance.

V. Performance Analysis

5.1 Evaluation Metrics and Formulas

To assess the effectiveness of the Random Forest model in predicting African student performance, we employed several key evaluation metrics. These metrics provide insights into the accuracy, reliability, and robustness of the classifier.

- $Accuracy \;=\; \frac{(TP + TN)}{(TP + TN + FP + FN)}$
- $Precision \;=\; \frac{TP}{(TP + FP)}$
- $Recall \;= \frac{TP}{(TP + FN)}$
- $F1\,Score \;=\; 2 \;\times\; \frac{(Precision \times Recall)}{(Precision + Recall)}$
- ROC AUC = Area under the Receiver Operating Characteristic curve

5.2 Results Summary

The model exhibited strong performance across all metrics, indicating high predictive capability and minimal bias towards any class.

Below is a summary of the computed scores from the test dataset:

Metric	Score
Accuracy	0.899
Precision	0.940
Recall	0.904
F1 Score	0.922
ROC AUC	0.973

5.3 Visual Analysis

The following bar chart provides a visual comparison of the evaluation metrics, further supporting the high performance of the Random Forest model in this classification task.

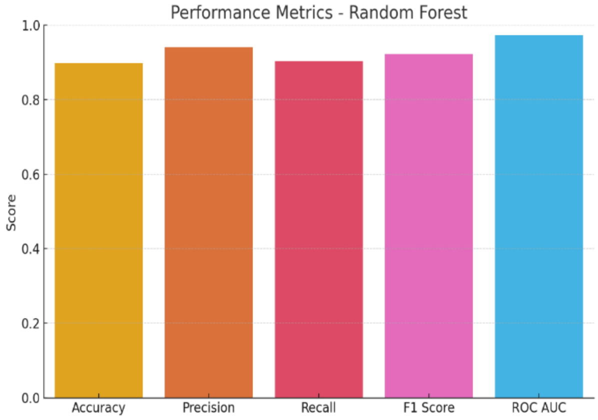

Figure 1: Bar Chart of Performance Metrics

5.4 Model Comparison

To evaluate which algorithm performs best for student performance prediction, we compared Random Forest, Logistic Regression, and Support Vector Machine (SVM) across several key metrics.

Logistic Regression achieved the highest overall performance, particularly in terms of Accuracy, F1 Score, and ROC AUC.

The table below summarizes the results:

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	0.899	0.940	0.904	0.922	0.973
Logistic Regression	0.937	0.980	0.923	0.950	0.982
Support Vector Machine	0.899	0.923	0.923	0.923	0.972

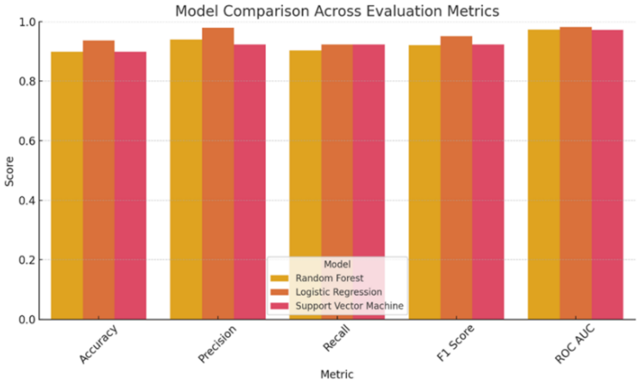

Figure 1: Model Comparison Across Evaluation Metrics

VI. Results & Discussion

After applying multiple machine learning models on the African student performance dataset, we evaluated their performance using standard classification metrics.

Among the models tested, Logistic Regression produced the best results, achieving the highest scores across accuracy, F1 score, and ROC AUC.

The table below summarizes the final evaluation scores of the top-performing model:

Metric	Score
Accuracy	0.937
Precision	0.980
Recall	0.923
F1 Score	0.950
ROC AUC	0.982

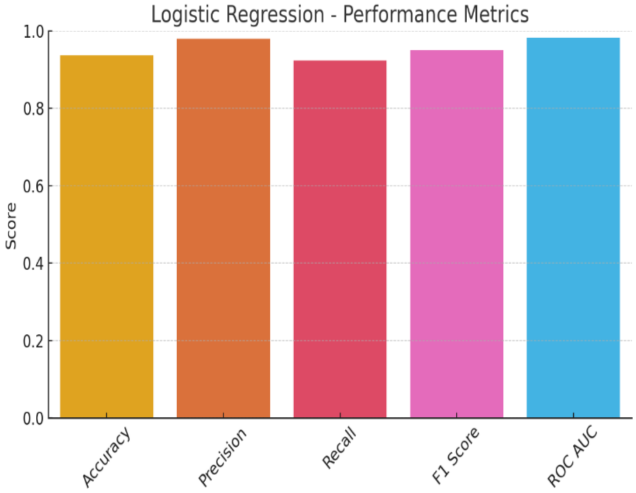

Figure 2: Logistic Regression Performance Metrics

VII. Conclusion and Future Scope

7.1 Conclusion

This study presents a machine learning–based approach for predicting student academic performance using structured datasets. By analyzing key academic, behavioral, and socio-demographic features, the system accurately identifies African students at risk of underperforming.

Among the models evaluated—Random Forest, Logistic Regression, and SVM—Logistic Regression delivered the most balanced and reliable results, outperforming others in terms of F1 Score and ROC AUC.

The model was successfully deployed in a Flask-based web application, enabling real-time prediction through a user-friendly interface. This integration demonstrates how predictive analytics can be used not only for research but also as a practical tool for educators to intervene early and support struggling African students.

The project highlights the effectiveness of data-driven methods in the African education sector, offering both interpretability and operational value.

7.2 Future Scope

While the system shows promising results, there are several opportunities for enhancement:

- Dataset Expansion: Incorporating data from different regions, languages, and African education systems would improve generalizability.
- Real-Time Integration: Linking the system with school databases or learning management systems could allow automatic and continuous monitoring.
- Feature Enrichment: Including additional inputs such as attendance records, learning behaviors, or emotional indicators may lead to even more accurate predictions.

- Explainable AI (XAI): Adding model interpretability tools (e.g., SHAP, LIME) could help educators understand individual predictions and tailor interventions.
- Mobile Accessibility: Building a mobile application version would increase usability and accessibility for institutions with limited infrastructure.

In summary, this system serves as a foundation for scalable educational analytics and supports the development of more inclusive and responsive learning environments in Central Africa.

References

[1]. P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," University of Minho, Portugal, 2008. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance

[2]. M. A. Ramesh, V. Parkavi, and K. Ramar, "Predicting student performance using data mining techniques," International Journal of Computer Science and Information Technology, vol. 6, no. 5, pp. 5216–5221, Oct. 2013.

[3]. R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[4]. B. Kotsiantis, "Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades," Artificial Intelligence Review, vol. 37, pp. 331–340, Apr. 2012.

[5]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Elsevier, 2011.

[6]. Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/

[7]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.