Prediction of Student Performance

Team Innovate

Data Fest AFrica Hackathon

October 9, 2024


iStock
Credit: epicurean
182061552

Outline

- Introduction
- Aim and objectives
- Significance of the study
- Methodology
- Results and discussion
- Conclusion and recommendation

Introduction

- Importance of Student Performance: Predicting academic success is crucial for shaping the future of students, especially in public schools where resources are limited.
- Limitations of Traditional Methods: Traditional prediction methods often rely on limited factors and may not account for the full range of influences on student outcomes.
- Machine Learning for Improved Accuracy: This project utilizes machine learning to analyze comprehensive student data, offering a more reliable method for early identification of students at risk of underperforming, enabling timely interventions.

Aim and objectives

Aim:

- The aim of this project is to develop a machine learning model that accurately predicts student performance, enabling early identification of students at risk and supporting timely educational interventions

Objectives:

- To identify key factors influencing student performance using machine learning techniques and data analysis
- To develop a predictive model that can accurately forecast students' academic outcomes, aiding in targeted interventions for at-risk students.

Significance of the study

- Improved Educational Outcomes: The project provides insights into factors affecting student performance, enabling schools to implement targeted strategies to improve academic success.
- Early Identification of At-Risk Students: The predictive model helps identify students who may need additional support, allowing for timely interventions to prevent academic failure.

Methodology

- **Data :** The dataset comprised information of all the activities within a public school. The generation of the datasets were primarily driven based on substantial facts with regards to the education policy that governs secondary schools in Nigeria, Lagos to be precise. The variables were created in accordance to the exact rules that governs public schools in Nigeria. The number of observations and fidelity of the data, and the fact that it is a good representative sample shows its usefulness.
- Train and Test Data: The models were trained using 80% of the available data, with the remaining 20% reserved for evaluating and testing algorithm performance.

Methodology : Feature Engineering

- Feature engineering is the process of creating, transforming, or selecting variables (features) from raw data to improve the performance and accuracy of a machine learning model. Some of the features engineered are discussed below:
- We applied extensive feature engineering / data preporecising by creating a pipeline that will aid the training and deployment, and also manke monitoring easy

Methodology : Machine Learning Algorithms

Machine Learning algorithms : The selection of machine learning models in this project was guided by their practical benefits and solid mathematical foundations. The ensemble models used are random forest, xgboost, catboost,gradient boost, Lightgbm, SVM, Logistic regression and Multilayer perceptron.

- Random Forest: An ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

$$\hat{y} = \text{majority vote}(T_1(X), T_2(X), \ldots, T_n(X)) \tag{1}$$

Methodology : Some other used Machine Learning Algorithms

- Gradient Boosting (GB): A boosting algorithm that builds models sequentially, where each subsequent model attempts to correct the errors made by the previous one.
- Extra Trees (ET): Similar to Random Forest, but the splits are made randomly, not based on information gain or Gini index, which helps in reducing variance and overfitting.

Optimization Techniques Used

We employed Grid Search as our primary hyperparameter optimization technique. Grid Search is a robust method that systematically searches through a predefined set of hyperparameters for each model. It evaluates all possible combinations of the hyperparameters, ensuring that we find the best configuration to maximize model performance.

How Grid Search Works: Grid Search exhaustively tests different combinations of hyperparameters by iterating over each parameter and evaluating the model's performance using a selected evaluation metric (such as accuracy, precision, or F1-score). Once all combinations are evaluated, the best-performing set of hyperparameters is selected based on the metric scores.

This approach, although computationally expensive, ensures that we do not miss any potential combinations that may lead to the optimal performance for our model.

Methodology : Evaluation metrics

- R-squared (R^2)**: R-squared indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values.

- Mean Squared Error (MSE)**: MSE measures the average of the squares of the errorsâthat is, the average squared difference between the estimated values and the actual values. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of observations.

Results:

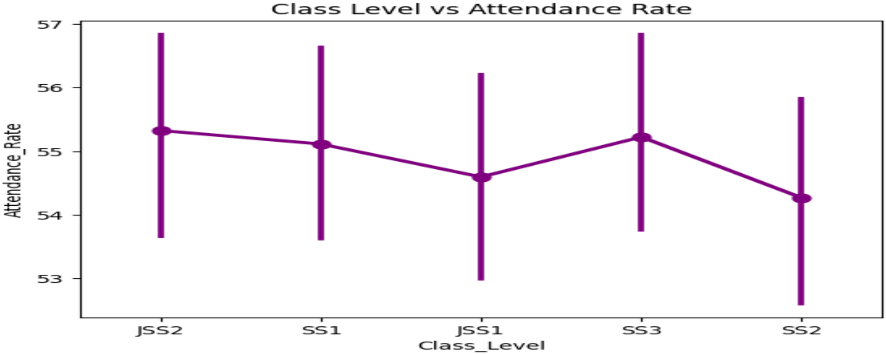


Figure: Class level by attendance

Results:

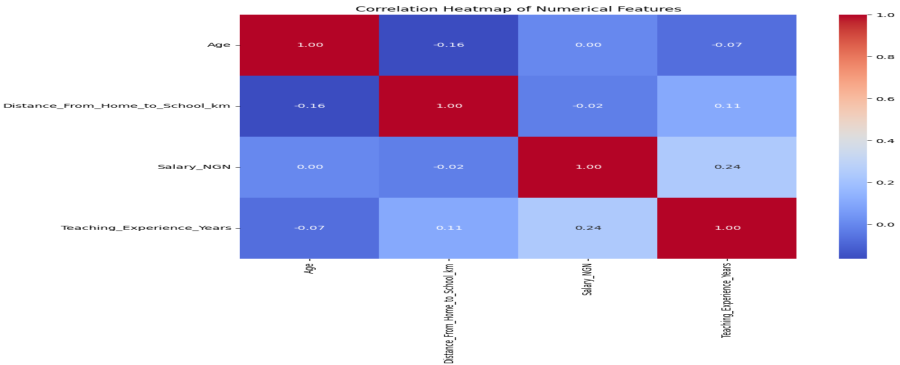
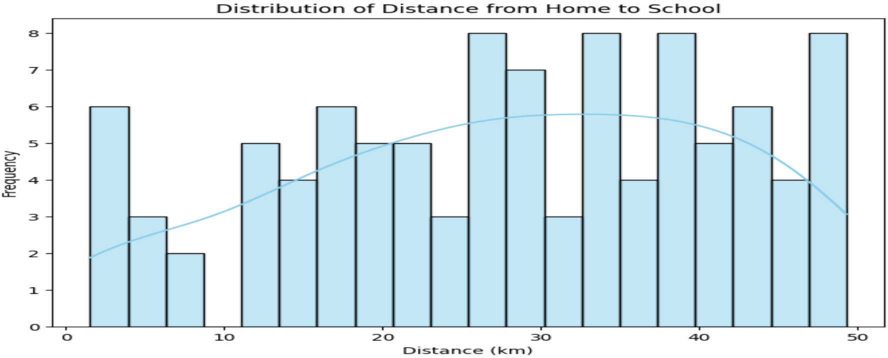


Figure: Correlation map

Results:



Distribution plot of distance from home to school

Results:

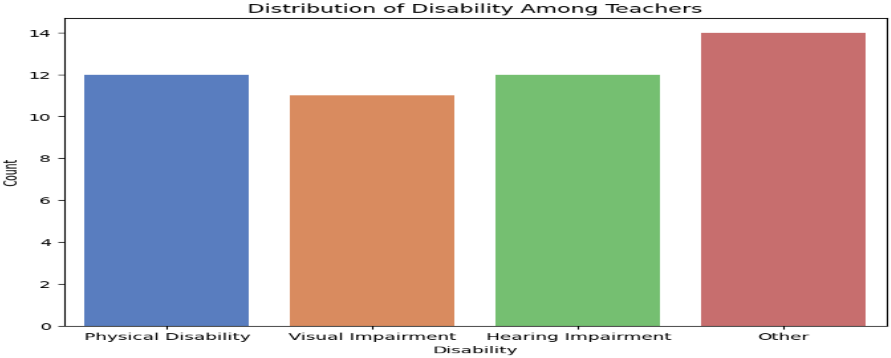


Figure: Distrbution of disability among teachers

Results:

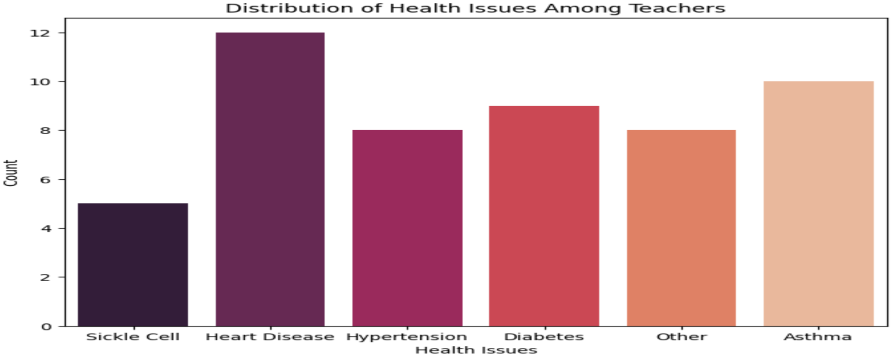


Figure: Distribution of teachers health issues

Results:

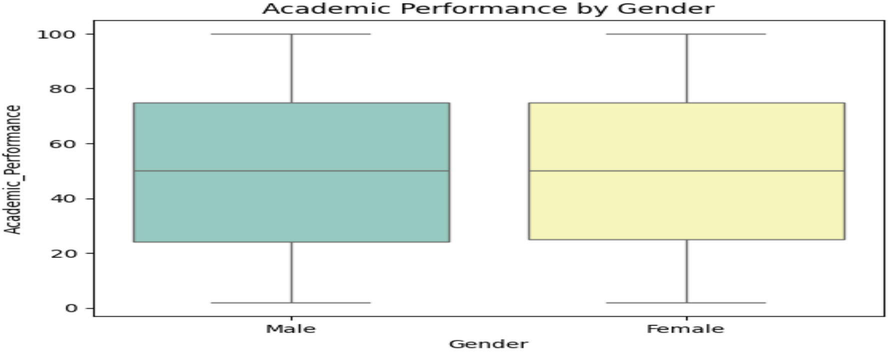


Figure: Academic performance by gender

Results:

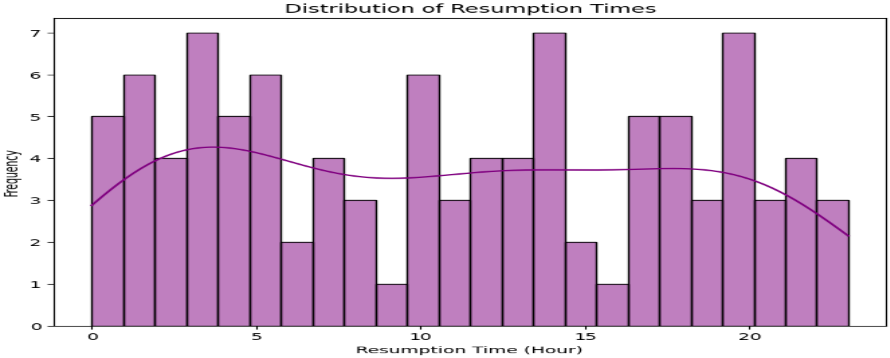


Figure: Distribution of resumption time

Results:

Figure: Salary by Education Level

Results:

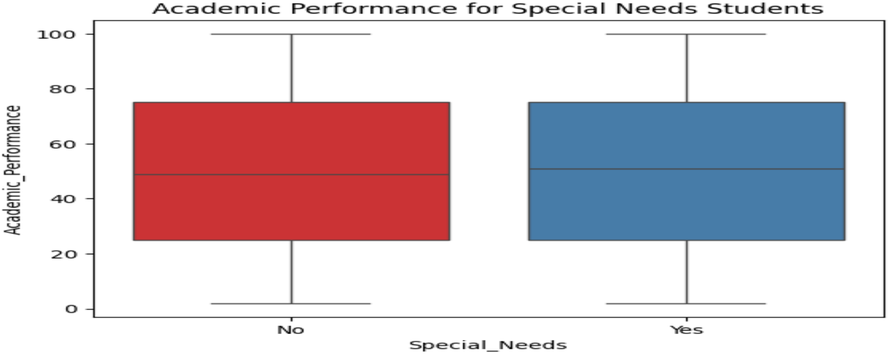


Figure: Academic Performance for Special Needs Students

Results:

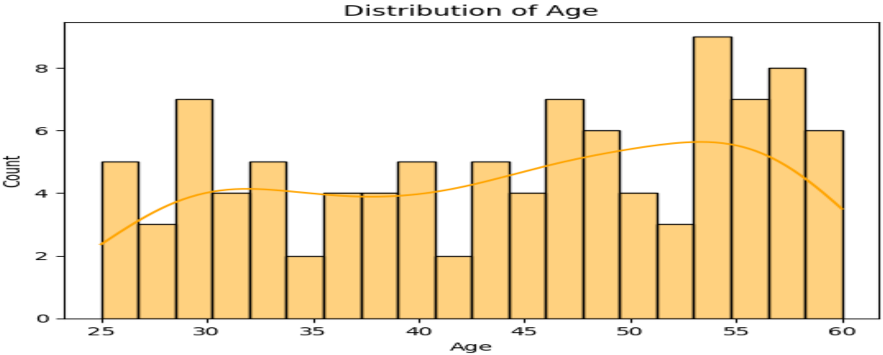


Figure: Teachers Age Distribution

Results:

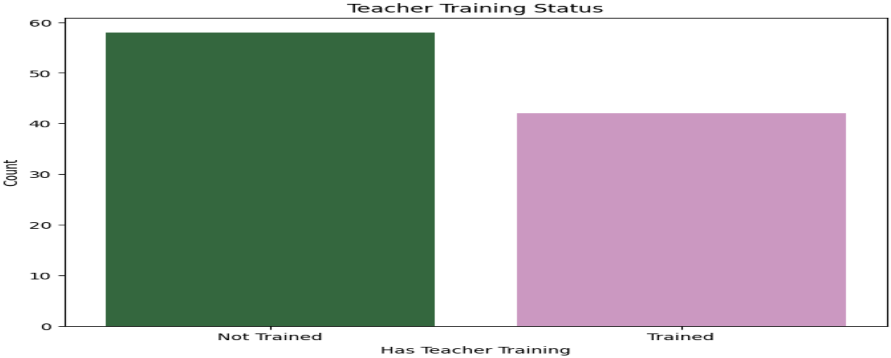


Figure: Teachers Training Status

Results:

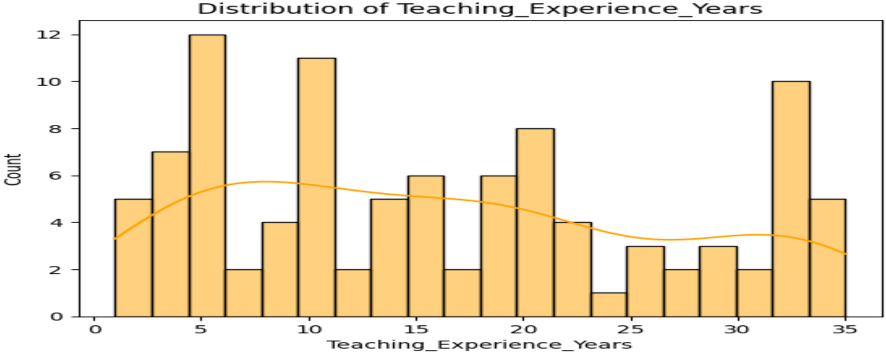


Figure: Distribution of teaching experience

Discussion of results

Table: Performance of Models Based on MSE

Model	**Mean Squared Error (MSE)**
Random Forest	629.91
Gradient Boosting	789.50
AdaBoost	827.00
Extra Trees	549.76
Stacking	411.96

Conclusion

- This project aimed to predict student performance using various regression models, including Random Forest, Gradient Boosting, AdaBoost, Extra Trees, and a Stacking Regressor. The Stacking Regressor achieved the best performance with the lowest Mean Squared Error (MSE) of 411.96, demonstrating the effectiveness of ensemble methods in enhancing predictive accuracy.
  Extra Trees also performed well with an MSE of 549.76, while Gradient Boosting and AdaBoost showed higher MSE values (789.50 and 827.00, respectively). These results highlight the importance of model selection and the potential of combining algorithms to improve predictions.
  This project emphasizes the value of accurate student performance prediction, enabling educators to identify at-risk students and implement timely interventions. Future work can focus on optimizing hyperparameters further and exploring additional features to enhance model performance.

Recommendation

- Implement Data-Driven Interventions: Schools should utilize predictive analytics to identify at-risk students early. By analyzing performance data, educators can tailor interventions, such as tutoring or mentoring programs, to address individual learning needs effectively.
- Adopt Ensemble Modeling Approaches: Schools should consider adopting ensemble modeling techniques, like the Stacking Regressor, for their internal performance evaluations. These models provide more accurate predictions and can guide strategic decisions regarding resource allocation and curriculum development.
- Invest in Continuous Training for Educators: Providing professional development for teachers on data interpretation and utilization of predictive tools will empower them to make informed decisions. Educators equipped with analytical skills can better support student learning and foster a culture of continuous improvement within the school.


Thank you