Project Title: Machine Learning for Attendance Rate Prediction(Educational infrastructure)

Date: November 9, 2023.

Team Members

Tochukwu Collins, Qudus Bello, Caleb Balogun, Chukwuegbo Love, Onyebuchi Mkpuluma, Habeeblah Aberejo

Executive Summary:

This project focuses on enhancing educational infrastructure with the prediction of attendance rates in educational institutions using machine learning techniques.
(Key findings from exploratory data analysis).

The project demonstrates the effectiveness of XGBoost Regression in predicting attendance rates.

Introduction

Accurate school attendance forecasting is crucial in sub-national African countries with limited educational resources. Various factors affect attendance rates, including educational infrastructure, socio-economic conditions, and cultural norms.

A number of factors can influence school attendance rates, including:
Educational infrastructure: The availability and quality of schools, classrooms, and other educational resources can play a major role in determining whether children are able to attend school regularly.

Socio-economic conditions: Poverty, unemployment, and other socio-economic challenges can make it difficult for families to send their children to school.

Cultural factors: In some communities, there may be cultural norms or traditions that discourage children from attending school.

Attendance forecasting is essential for effective education planning and resource allocation, and the project's machine learning models have the potential to significantly enhance educational outcomes in sub-national African countries.

Problem Statement

In subnational African countries, forecasting school attendance rates is essential for effective education planning and resource allocation.

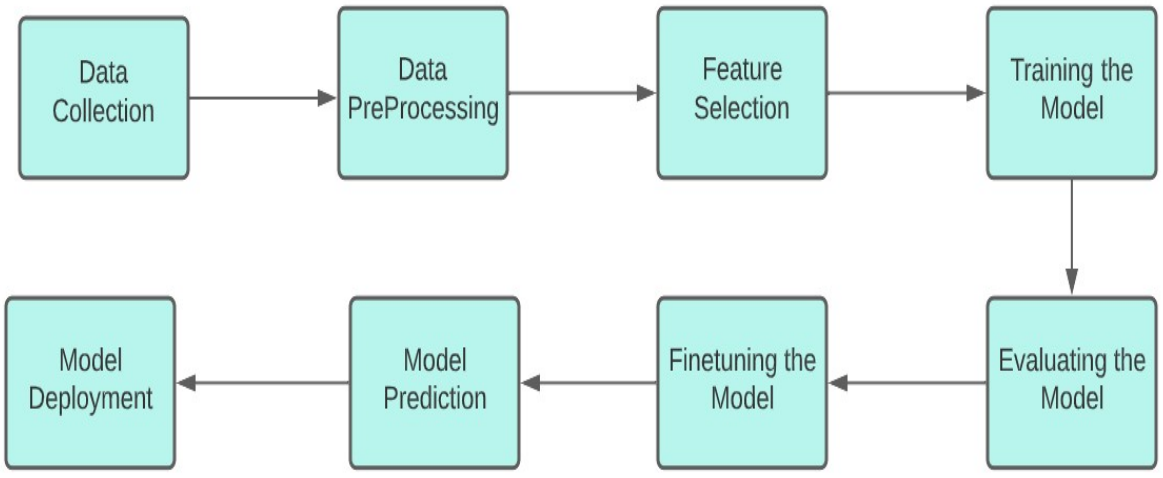
This project hopes to address this issue by developing machine learning models that predict attendance rates based on a comprehensive analysis of educational and socio-economic factors.

If this problem is solved, improving attendance predictions will empower education policymakers to enhance resource allocation and educational outcomes in these regions.

Objectives

The primary objectives are to improve attendance rate prediction accuracy and provide insights for educational institutions.

Data Collection and Preprocessing
Flow Process
The steps taken are illustrated with the flowchart below:

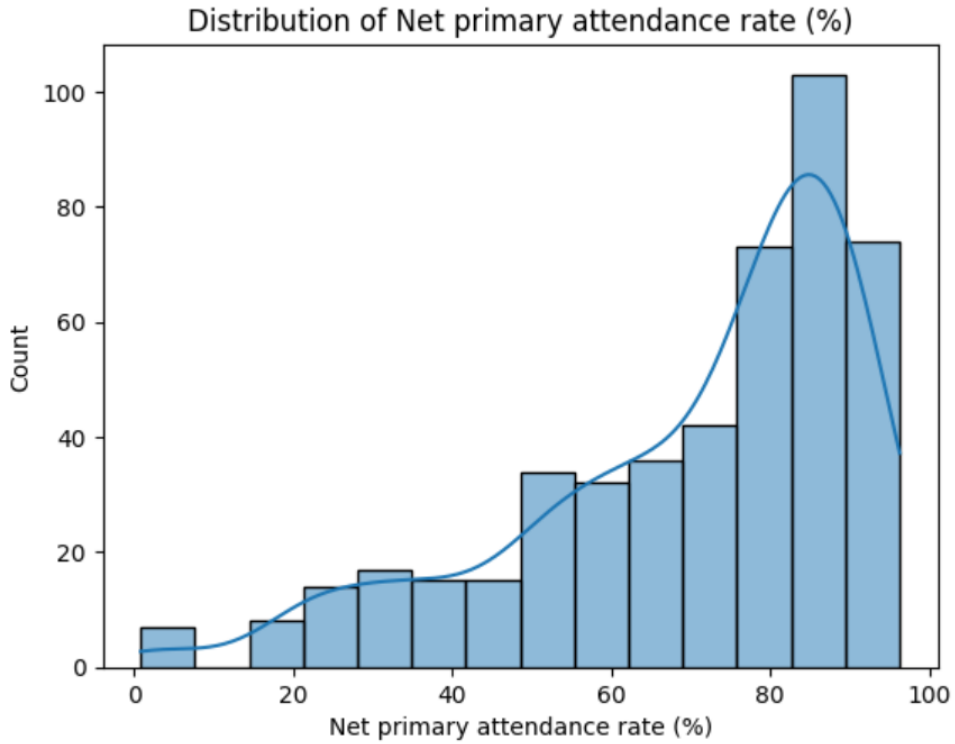


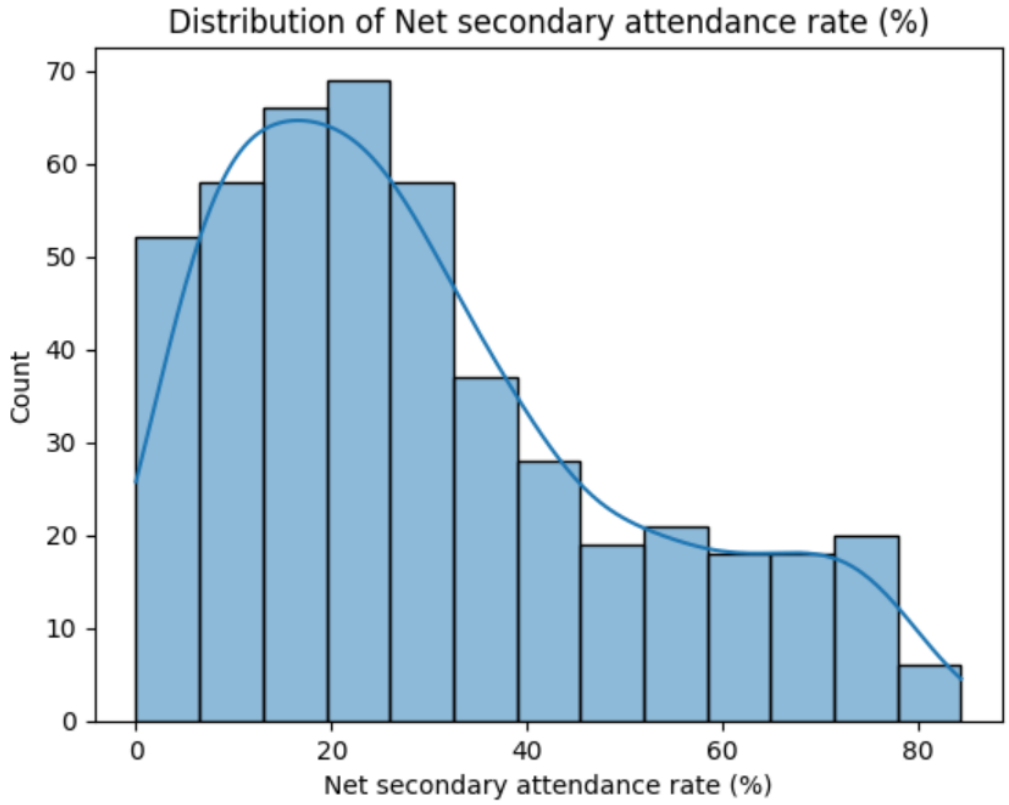
Data Collection
The dataset, compiled by Climate Change and African Political Stability (CCAPS), whose focus is analyzing how climate change, conflict, governance, and aid intersect to impact African and international security.
The link to the dataset is stated below:
[https://www.strausscenter.org/wp-content/uploads/Subnational_African_Data_August_2013.zip]

Data PreProcessing
The main objective of data cleaning was to address missing and null values, which sometimes required dimensionality reduction by eliminating columns with insufficient data.
For our project we dropped the columns with high null values, and used the mean values to fill rows with minimal null values.

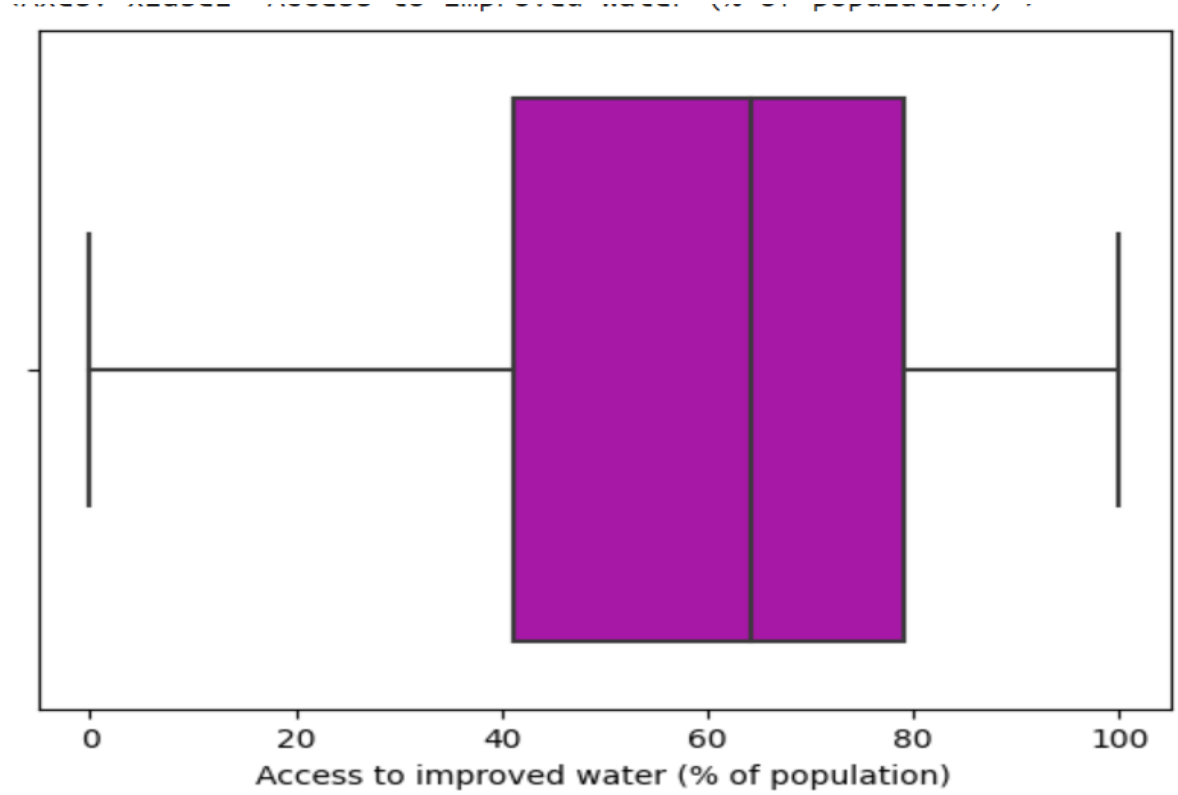
Feature Selection
The dataset key features in our dataset includes literacy rates, primary and secondary school attendance rates, access to improved water and sanitation, household access to electricity, and household ownership of radio and television at the subnational level, specifically the first administrative district level.
The dataset contains values between the year 2003 to 2011 in regions within Africa.

Exploratory Data Analysis
During this phase, our objective was to uncover patterns, relationships, anomalies, and test assumptions within our data using graphical representations and simple statistical summaries.
The following procedures were used to prepare the data:
Data collection: The data consisted of 470 rows and 60 columns.
Data visualization: Here, information was displayed using histograms, box plots and other visualaids to facilitate clear and simple interpretation.


Distribution of Net primary attendance rate (%)
Count
Net primary attendance rate (%)
0
20
40
60
80
100


Distribution of Net secondary attendance rate (%)
Count
Net secondary attendance rate (%)
0
10
20
30
40
50
60
70
0
20
40
60
80

From our histograms above, we found that the higher the level of education, the lesser the rate of attendance. This is shown in the skewness to the right of the Net primary attendance rate and vice versa for the Net secondary attendance rate.



Access to improved water is one of the infrastructural facilities. It's median access is about 65%, the Q1 is 40% while the Q3 is 80%. The plot shows that majority of the subnational african countries have between 40 to 80% access to improved water.

Modeling
Machine Learning Models: We utilized ML models like:
Support Vector Regression
K-Nearest Neighbors Regression
XGBoost Regression.

Evaluation Metrics:
Model performance was evaluated using evaluation metrics like
Mean Squared Error (MSE)
Mean Absolute Error (MAE)
R-squared (R2_Score)
Summary

Training the model after thorough data preprocessing techniques, the model was evaluated with evaluation metrics which include, R2Score, MAE, MSE which gave general comment on the model and the model is can be considered ready for Real-world deployment of solution.

Conclusions

The XGBOOST model successfully improves attendance rate prediction accuracy.
Improvement of model performance
-Further experiments with additional features and models and hyperparameter tuning for the selected models
- Enhancements in data preprocessing techniques.
- Real-world deployment of solutions.