Predicting School Attendance Rate in Sub-National African Countries (Using Educational Infrastructure)

Saturday, 11th November, 2023

TEAM NEURAL CRAFT

Our Team

Aberejo Habeeblah
Presenter 1

Caleb Balogun
Presenter 2

Project Lead:
- Tochukwu Collins

Assistant Project Lead:
- Qudus Bello

Query Analyst:
- Chukwuegbo Love

Other Active Members

Caleb Balogun	Model Development
Onyebuchi Mkpuluma	Visualization
Aberejo Habeeblah	Presenter

Problem Statement

- In subnational African countries, forecasting school attendance rates is essential for effective education planning and resource allocation.
- This project hope 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.
- The problem we are trying to solve is to determine attendance rate which can helps in educational infrastructure like building, equipments, instructors for the students.

Aim

- The aim of this project is to develop accurate machine learning models for predicting school attendance rates in subnational African countries, utilizing educational and socio-economic factors as key predictors.

Our Approach

- **Feature Selection:** Identify the most relevant features that impact school attendance rates within the selected regions from the data provided.
- **Model Development:** Create machine learning models that leverage the chosen features to predict school attendance rates with a high degree of accuracy.
- **Model Evaluation:** Assess the performance of the developed models through rigorous testing and validation using historical attendance data.

Dataset Description

- 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 dataset features provides data on 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.
- The dataset has a shape of (471, 68) which interpretes that there are 471 regions in 38 African countries with 68 columns.

Dataset Description

- Below is a flowcharts diagram explain how the data wrangling, processing and exploration is done.

- **Identification:** Getting our data
- **Cleansing:** Identifying data quality issues like null values and getting rid of them.
- **Transformation:** Making a new copy of our dataset into a usable format.
- **Loading:** Moving the cleaned unified data into algorithms for model building.

Codes addressing all of this can be found in the respective cells of the Colab file

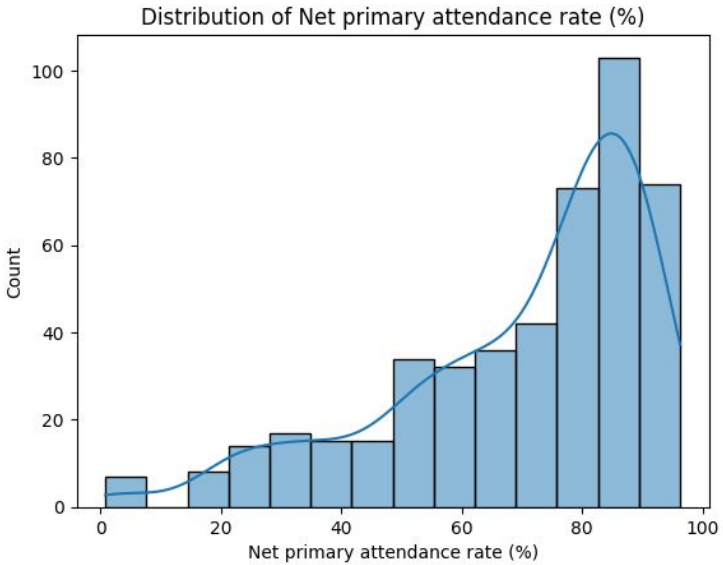
HAMOYE

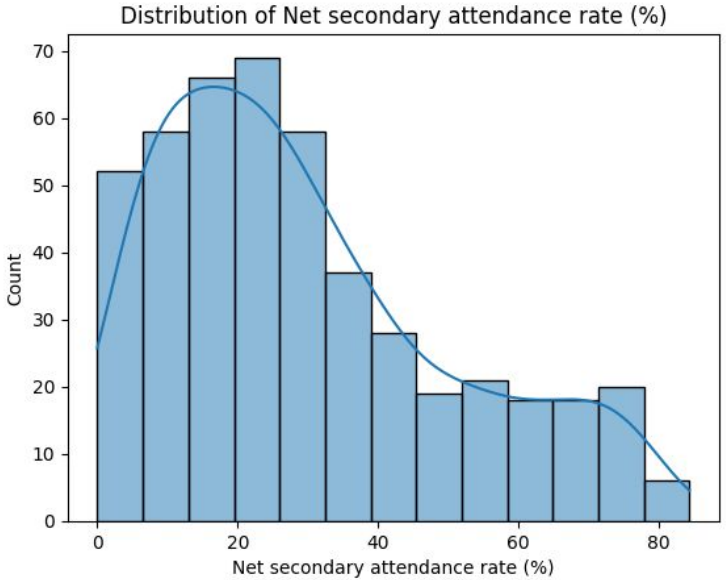
Dataset Description

- The cleaning of the dataset include filling column with less than 10 missing values with the mean of the column in the dataset. Such columns include Television (% of pop) sample size, Radio and/or Television in household (% of population), Access to improved water (% of households).
- We built a simple model to predict the values for columns that have missing values within the range of 11 to 23. Such columns include Electricity (% of hh) sample size, Radio in household (% of households), Radio
- Columns with missing values that's above 30 were totally dropped, this is because our dataset isn't that large to have missing column of about 124 as in some column like Literacy rate (15 & over), Literacy rate (15-24), Literacy rate (25-49).

Dataset Description

- The histogram below tell number of regions in relation to the attendance rate (both Primary and Secondary)

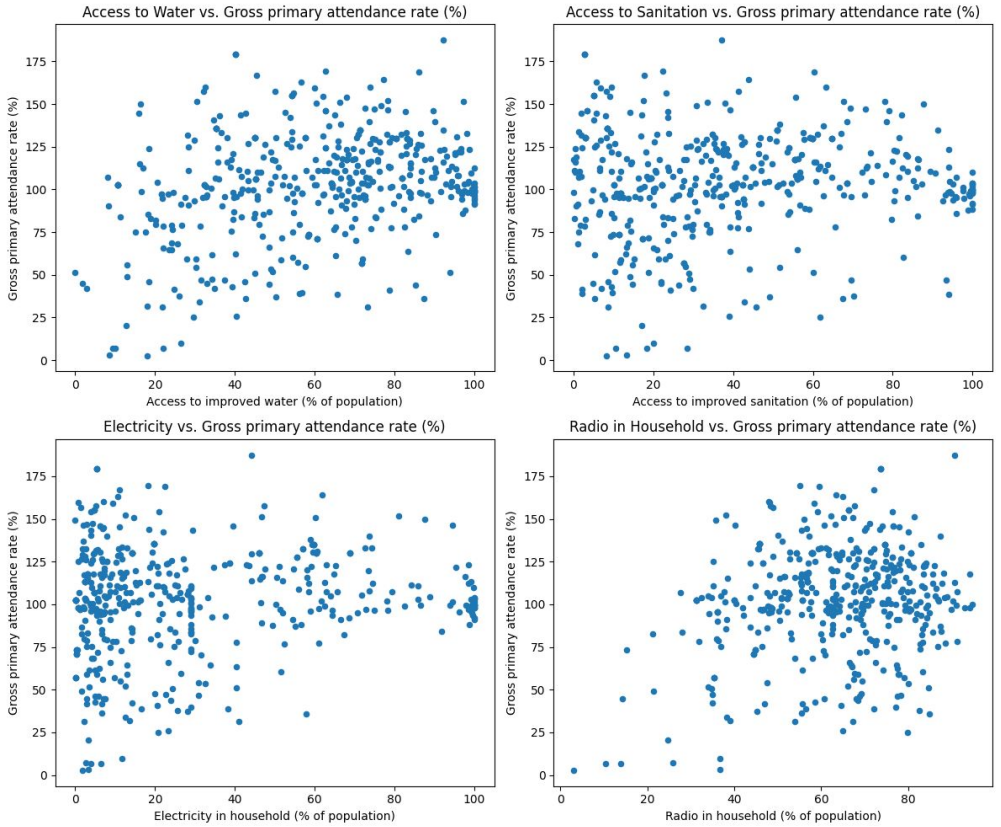




Dataset Description

- The scatterplot here plots the relation between some of our features to primary attendance rate

A detailed PowerBI report can be found via the link below, https://app.powerbi.com/view?r=eyJrIjoiY2JlODBjYTItMDlmMy00MDM4LTgzYTItN2NlNDVlMGZlYzczIiwidCI6ImRmODY3OWNkLWE4MGUtNDVkOC05OWFjLWM4M2VkN2ZmOTVhMCJ9



Dataset Description

After the EDA, we decide go build model with the features below

- Access to improved water (% of population)
- Access to improved sanitation (% of population)
- Electricity in household (% of population)
- Radio and/or television in household (% of households)

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)

Modeling: Result from evaluation metrics

Metric	Net Primary Attendance Rate (%)	Gross Primary Attendance Rate (%)	Net Secondary Attendance Rate (%)	Gross Secondary Attendance Rate (%)
Mean Squared Error (MSE)				
Support Vector Reg	586.3823933	948.300889	471.514031	922.1880847
K-Nearest Neighbo	361.3116137	670.906006	319.669848	637.2724139
XGBoost Regressic	138.3629945	238.0855609	65.16007364	259.3934687
Mean Absolute Error (MAE)				
Support Vector Reg	17.56923282	22.90956146	16.46599687	23.19936389
K-Nearest Neighbo	14.34091332	19.05497136	12.52406729	18.32289102
XGBoost Regressic	6.305303724	9.320854659	5.206873234	9.947774421
R-squared				
Support Vector Reg	-0.124980764	-0.002000152796	-0.09937158751	-0.1189699363
K-Nearest Neighbo	0.3068198842	0.2911027203	0.2546649194	0.2267426958
XGBoost Regressic	0.7345491456	0.7484324109	0.8480742271	0.6852556458

Conclusion

- Key Findings:
  XGBoost Regression Model is recommended for attendance rate prediction in educational institutions because it has the lower MSE Value, lower MAE value higher R2Score value as needed by a good model.
- Conclusions:
  The XGBoost Model successfully predict attendance rate with good accuracy.

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 XGBoost model can be considered ready for Real-world deployment of solution.
- We recommended that attention is given to regions with lower attendance rate

HAMOYE

Thank You