📚 Predicting Literacy to Bridge the Education Gap: A Machine Learning Approach for Ethiopia

# Authors: Group 13\_FTL Ethiopia

# 1. Introduction – Why Literacy? Why Now?

In a world increasingly driven by digital literacy and knowledge economies, the ability to read and write remains a fundamental human right and a cornerstone for individual and national development. Yet, for countries like Ethiopia, disparities in literacy continue to challenge progress.

As of recent years, national reports from the Ministry of Education (MOE) and demographic insights from Humanitarian Data Exchange (HDX) reveal that significant gaps persist in literacy rates across regions, genders, and age groups. These disparities hinder not only education outcomes but broader development indicators tied to poverty, health, and employment.

## 🔗 Link to SDGs

This project aligns directly with SDG 4: Quality Education, and indirectly supports SDG 1: No Poverty and SDG 10: Reduced Inequalities. By harnessing machine learning to uncover trends and forecast literacy rates, we aim to help policymakers, educators, and NGOs allocate resources effectively and equitably.

## 🎯 Purpose

The core goal of this blog is to walk you through the creation of a predictive model for literacy rates in Ethiopia. We explore how historical data on population and school enrollment can be transformed into actionable insights using machine learning.

# 2. Body – Understanding the Problem and Building the Solution

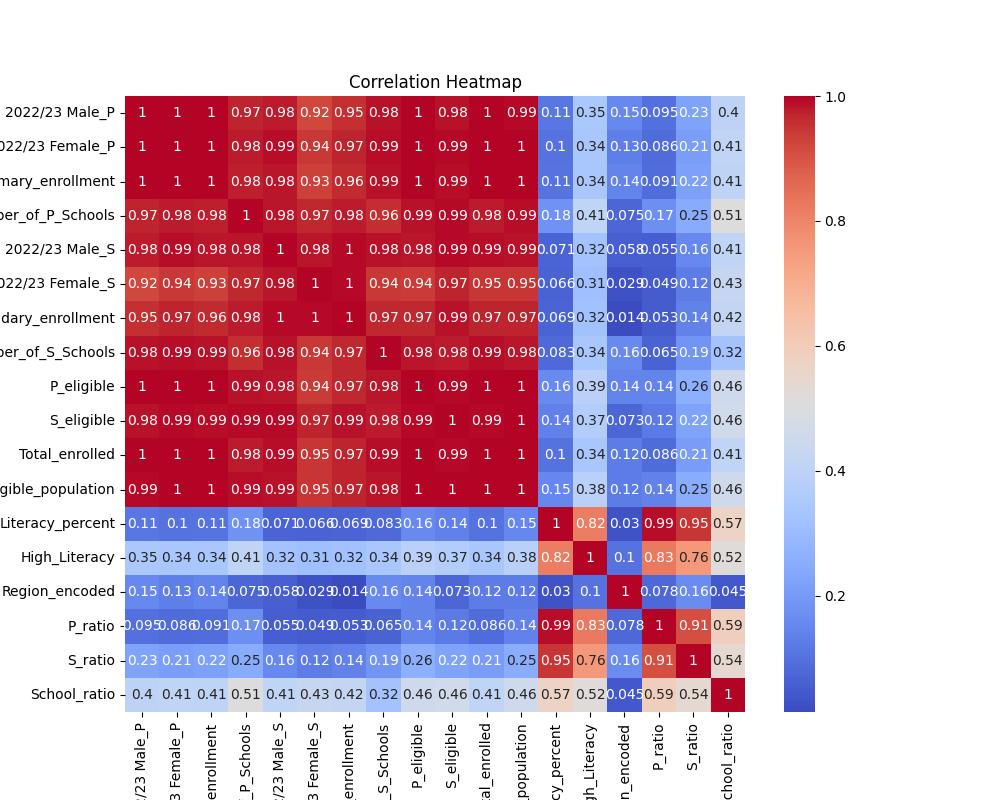
## a. The Literacy Challenge in Ethiopia

While Ethiopia has made substantial progress in school enrollment, particularly at the primary level, literacy levels still lag behind. Regions such as Afar, Somali, and Gambella face persistent challenges. This discrepancy points to deeper structural issues beyond access, including quality of education and sociocultural factors.  
  
In response to this, we asked:  
> Can we predict literacy rates using historical trends in enrollment and population data to inform better interventions?

## b. The Process

### i. Methodology

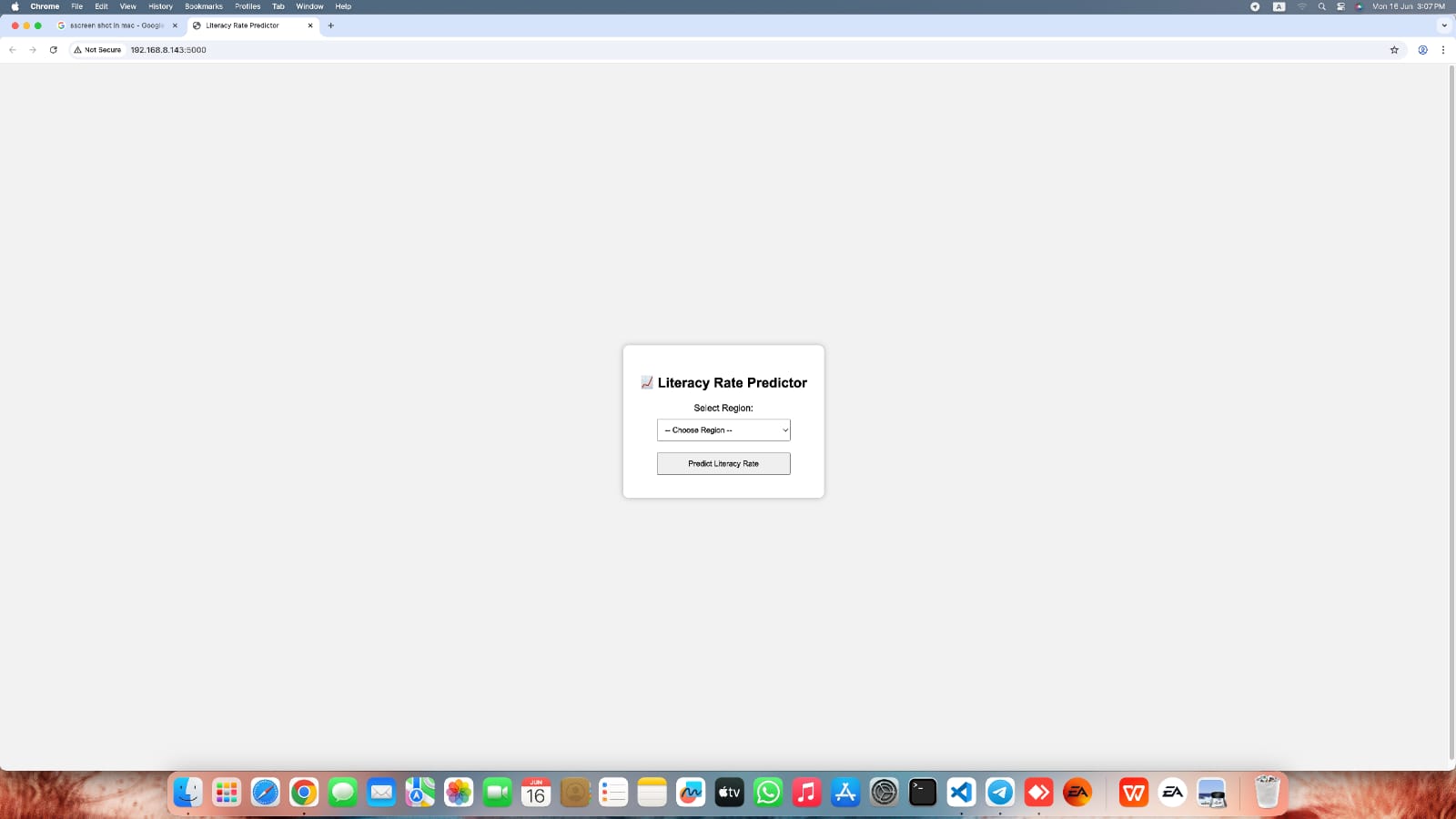
We built our model using two main datasets:  
- MOE PDF Reports (converted to CSV): Enrollment rates in primary and secondary schools by region and year.  
- HDX Population Dataset: Yearly population estimates by region.  
  
After preprocessing and feature engineering, we formulated the problem as a regression task:  
- Input features: Year, Region, Primary Enrollment, Secondary Enrollment, Total Population.  
- Target: Literacy Rate.  
  
We tested several algorithms:  
- Linear Regression  
- Random Forest Regressor  
- XGBoost Regressor  
  
We split the data into 80/20 train-test sets, used GridSearchCV for hyperparameter tuning, and scaled numeric features.



### ii. Implementation

Key steps of the pipeline:

from sklearn.linear\_model import LinearRegression  
from sklearn.model\_selection import train\_test\_split  
from sklearn.metrics import mean\_squared\_error  
  
# Example training  
X = df[['Year', 'PrimaryEnrollment', 'SecondaryEnrollment', 'Population']]  
y = df['LiteracyRate']  
  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2)  
model = LinearRegression()  
model.fit(X\_train, y\_train)  
  
predictions = model.predict(X\_test)



## Feature engineering

1. Literacy rate

2. Primary to secondary school ratio

3. Secondary to primary school ratio

## c. Results and Findings

Model performance comparison:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE | MAE | R² Score |
| Linear Regression | 3.42 | 2.19 | 0.97 |
| XGBoost | 14.16 | 11.42 | 0.55 |
| Random Forest | 14.89 | 11.26 | 0.50 |

# 3. Discussion

While Linear Regression worked well on this dataset, some challenges arose:  
- Data quality: MOE reports are in PDF format and often required cleaning and manual parsing.  
- Limited features: We had no access to variables like teacher-student ratio, infrastructure, or socioeconomic factors.  
- Regional disparities: Outlier regions require more contextual modeling than a one-size-fits-all solution.  
  
Future iterations could incorporate:  
- NLP for automatic extraction from PDFs  
- Deep learning for more nuanced regional modeling  
- Interactive dashboards for policymakers

# 4. Conclusion – Why This Matters

This project shows that even with limited data, machine learning can play a powerful role in addressing educational inequality. We have demonstrated that it’s possible to model and forecast literacy with high accuracy using accessible datasets and simple algorithms.

## 🛠️ Recommendations

- For developers: Expand the model with geospatial, economic, or health-related data.  
- For governments: Use such models to target literacy interventions regionally.  
- For NGOs: Leverage these insights to align project funding with underserved populations.  
  
Ultimately, this is about transforming data into decisions—and decisions into meaningful change.

# 5. References

- Ministry of Education, Ethiopia. “Annual Education Abstracts” (PDF Reports), various years.  
- Humanitarian Data Exchange (HDX). “Ethiopia Population by Region and Year.” https://data.humdata.org  
- SDG Knowledge Platform. “Sustainable Development Goal 4.” https://sdgs.un.org/goals/goal4