Project Title: Predicting Access to clean water using Machine Learning

**1 Overview:**

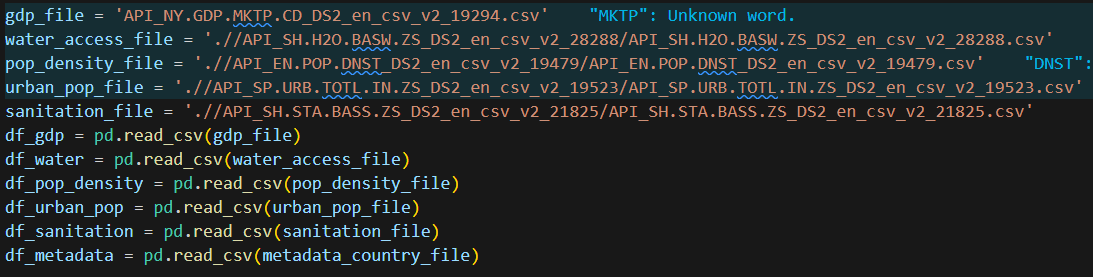
**Purpose:**

The data preparation and feature engineering phase ensures raw data is transformed into a clean, structured format suitable for machine learning models. This phase is critical because:

* **Quality assurance**: Removes noise, inconsistencies, and biases in the data
* **Relevance**: Identifies features most predictive of the target (clean water access).
* **Model performance**: Well-engineered features improve accuracy, interpretability, and generalizability.

**2 Data Collection:**

* World Bank Open data
* <https://washdata.org/data>



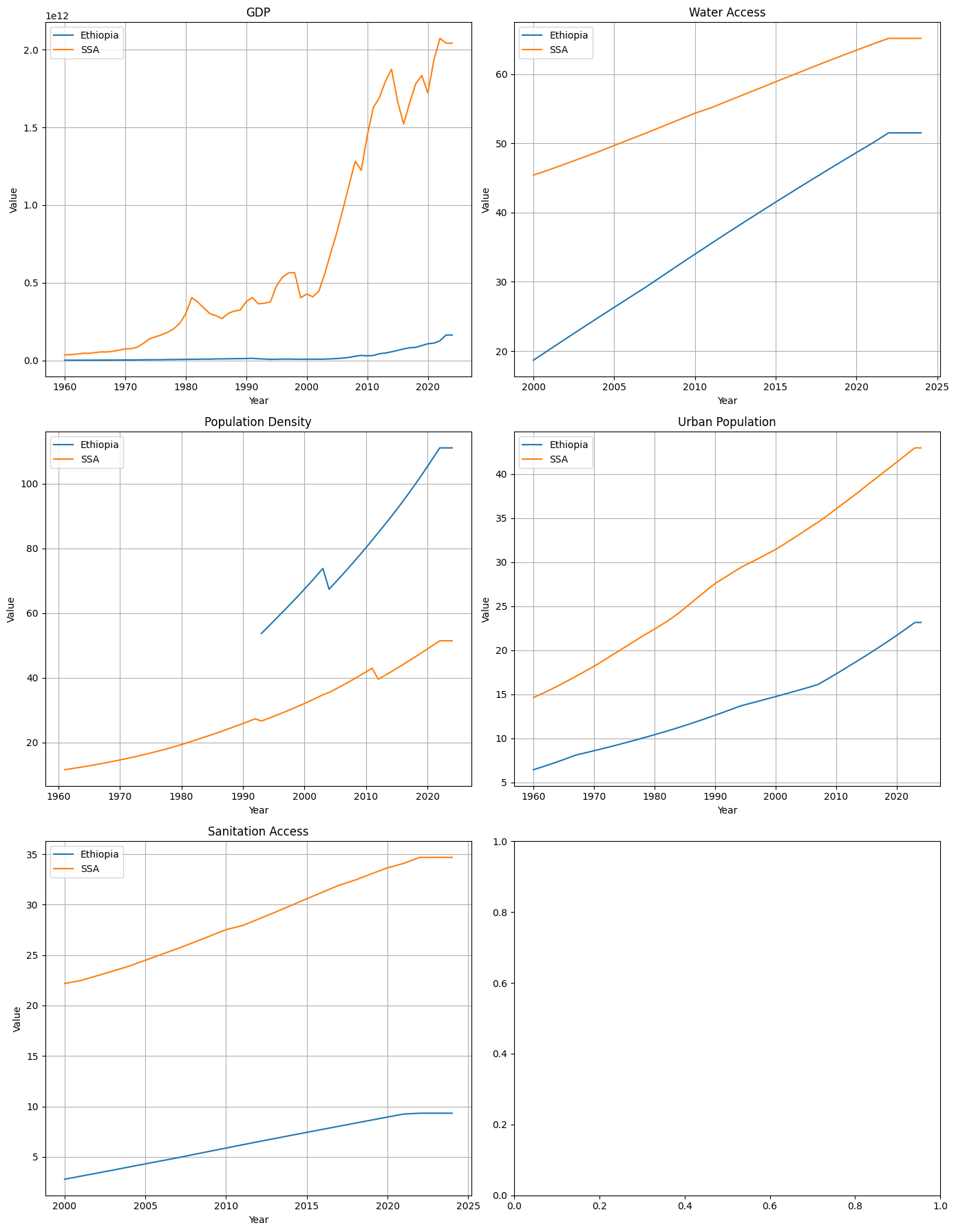
3 Data Cleaning:

We cleaned each dataset by:

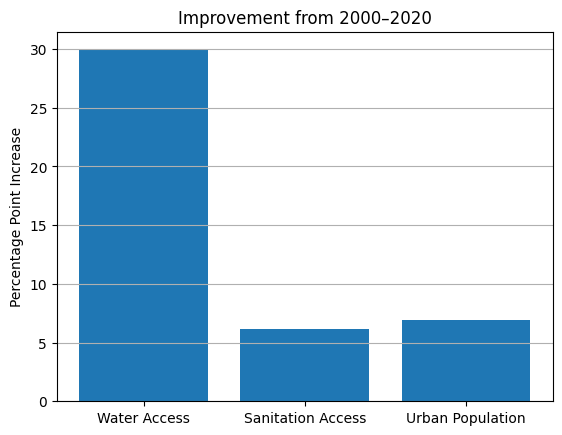
* Converting year columns to numeric
* Replacing invalid values with NaN
* Applying forward fill to inter

4 Exploratory Data Analysis (EDA)

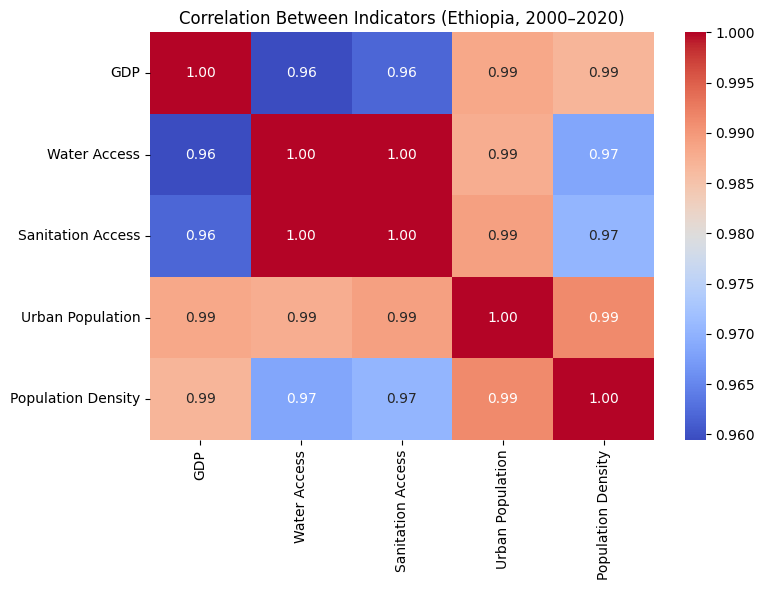
Visualis trends Over time for each indicator, comparing Ethiopia vs SSA.



Improvement From 2000-2020



Correlation Between Indicators (Ethiopia, 2000-2020)



**6. Data Transformation**

Techniques Applied:

**1 Scaling :**

Standardized numerical features (e.g., GDP, population density) using StandardScaler() for gradient-based models.

Log-transformed skewed features (e.g., GDP):

df["Log\_GDP"] = np.log1p(df["GDP"])

**2 Encoding :**

One-hot encoding for Region (from metadata) to avoid ordinal bias.

Target encoding for income groups (low/medium/high) to preserve order.

**3 Normalization :**

Min-Max scaled features like urban population (%) to bound values [0,1].

**4 Train-Test Split :**

Stratified split by year and region to prevent temporal leakage and regional bias.

**Model Evaluation**

**1. Model Selection**

Selected Model: Random Forest Regressor

We chose the Random Forest Regressor because:

It performs well on non-linear relationships

Handles missing values and noisy data robustly

Provides feature importance insights

Requires minimal preprocessing (no scaling needed)

**✅ Strengths:**

High accuracy on mixed data types

Good for small-to-medium datasets

Reduces overfitting through averaging multiple trees

**❌ Weaknesses:**

Slower than linear models

Not ideal for very high-dimensional sparse data

Harder to interpret than simple regression

**2. Model Training**

We trained the Random Forest Regressor using indicators from all available countries , focusing on Ethiopia during inference.

Steps:

Merged all indicators (GDP, Urban Population, Sanitation Access, etc.)

Filtered rows where target variable (Water Access) was available

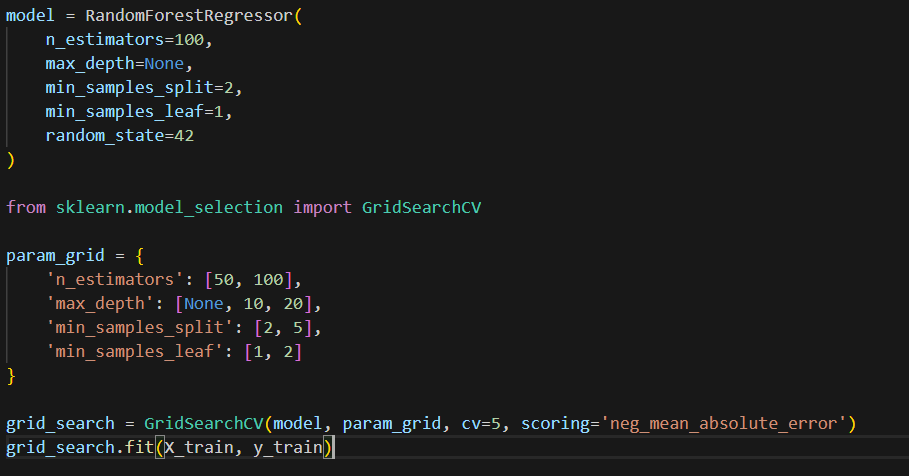
Created long format: each row = country + year + features + water access

Split into train/test by year

Trained model on global data

Applied to Ethiopia to forecast missing/future values

Hyperparameters (Default Values):



**3. Model Evaluation**

We evaluated the model using:

Mean Absolute Error (MAE) – average error in prediction

R² Score – how well the model fits test data

Residual Plots – to check for bias

Feature Importance Visualization – which factors most influence water access