SERUNJOGI HUZAIFA 2500726371 2025/HD05/26371U

Coffee Yield Prediction Using Climatic, Vegetation, and Production Data: Explanatory Report

1. Introduction

Coffee is one of Uganda's major cash crops. Its productivity is strongly influenced by climatic conditions, vegetation health, and production capacity. Accurately predicting coffee yield could realise benefits to farmers, policymakers, and value-chain stakeholders to make strategic plans about production, resource allocation, and climate adaptation.

In this project, a machine learning model is developed to predict annual coffee yield (kg/ha) across Uganda using historical data from **1961–2022**. The dataset integrates key variables which include rainfall, temperature, NDVI (vegetation index), bearing coffee trees.

The aim is to make an evaluation whether linear and nonlinear models can reliably estimate yield and also identify the most influential features.

2. Methodology

2.1 Dataset

The dataset consists of annual regional records containing:

- **Rainfall (mm)** – annual precipitation totals.
- **Temperature (°C)** – average annual temperature.
- **NDVI** – vegetation greenness as an indicator of crop health.
- **Bearing trees** – number of productive coffee trees.
- **Region** – one of the five main coffee-producing sub-regions.
- **Year** – from 2005 to 2022.

- **Yield (kg/ha)** – computed from production and area.

The yield data, areas with coffee plantations(hectares), bearing trees, incidence index was manually entered from UCDA yearly reports in combination with Faostat data about coffee output. Rainfall data was obtained from CHIRPS in relation to the Google earth engine. ISRIC SoilGrids for soil data. Fertilizer data was obtained from UCDA, World bank and FAOStat.

The final dataset used was a combination of historical climate data and production statistics, complemented by synthetic adjustments that maintain realistic regional and temporal patterns.

2.2 Data Preprocessing

- Missing values were checked and addressed where necessary.
- Numerical features were normalized where required.
- A train–test split was applied to evaluate model performance.

2.3 Exploratory Data Analysis (EDA)

- Data trends in rainfall, temperature, vegetation health, and yield were analyzed.
- Correlation analysis identified strong relationships, notably between yield and bearing trees, NDVI, and rainfall.
- Visualizations such as scatter plots, line charts, heatmaps helped reveal patterns and seasonal dynamics.

2.4 Model Selection

- Various modelling approaches were applied on the data including decision tree, random forest and XGBoost as shown in the Model Comparison notebook.
- The best performing model was Random Forest which gave meaningful metrics especially for the time-series score

2.5 Model Evaluation

- The models were evaluated against metrics like f1, accuracy, R Squared, MAE(Mean Absolute Error) and RMSE(Root Mean Square Error) as shown in the model comparison notebook.

- The Random forest model achieved overall best performance. This indicates that the dataset contains non-linear patterns and complex interactions between features.
- A sample Engine which makes predictions of the yield is included in a notebook.

---

3. Conclusion

The project successfully demonstrates that machine learning can be applied to predict coffee yield using production history, climatic, vegetation, and production-related variables. These insights support the potential of data-driven forecasting tools for agricultural planning and climate-smart decision-making in Uganda's coffee sector.

Future works could recommend finer spatial data, soil characteristics, satellite-derived vegetation metrics, and farm-level management practices to further improve accuracy in predictions.