# FERTILIZER OPTIMIZER DATA REPORT

MORINGA DSF-FT 11 – HYBRID

MODEL KOMBAT

**PRESENTED BY:**

Rodgers Ndemo

Petina Leni

Kennedy Kariuki

Betsy Gitje

Michael Gatero

Raniel Musyoki

# TABLE OF CONTENTS

# FORECASTING FERTILIZER EFFICIENCY AND AGRICULTURAL PRODUCTIVITY

## 1.0 Business Understanding

### 1.1 Business Context

Agriculture forms the backbone of East Africa's economy, employing over 60% of the workforce and contributing significantly to the GDP. Despite investments in fertilizer programs, productivity remains low. This disparity raises questions about the efficiency and strategic application of fertilizers in the region.

### 1.2 Business Problem

Fertilizer use is rising across East Africa, but crop yield improvements are inconsistent and vary widely between countries. Policymakers, NGOs, and agritech investors are concerned that fertilizer is being applied inefficiently, without data-driven guidance. Without clear insights into the fertilizer productivity relationship and future needs, policies may misallocate resources, and farmers may suffer from suboptimal yields. There is need to optimize fertilizer use to sustainably boost agricultural output using historical data.

### 1.3 Project Objectives

1.3.1 Main Objectives

- To develop a data-driven framework that forecasts fertilizer usage and supports sustainable agricultural productivity across East Africa, empowering stakeholders with insights that guide better policies, investments, and resource allocation.

1.3.2 Specific Objectives

1. To analyze historical fertilizer consumption trends across East African countries from 1960 to 2023.
2. To investigate the relationship between fertilizer usage and agricultural productivity indicators.
3. To develop time-series models for forecasting future fertilizer demand up to 2035.

4. To build machine learning models that predict productivity outcomes based on fertilizer use and other variables.
5. To cluster countries with similar fertilizer efficiency patterns for targeted policy recommendations.
6. To generate insights and actionable strategies that support sustainable agriculture and food security beyond the scope of the project.

## 1.4 Success Metrics

1. **Accuracy of Forecasting Models**:
   - Achieve a minimum accuracy of X% (e.g., >85%) in time-series forecasting models for fertilizer demand by 2035.
   - Low Mean Absolute Percentage Error (MAPE) or Root Mean Square Error (RMSE) for predictions.
2. **Correlation Strength**:
   - Strong positive correlation (R^2 value > Y, e.g., >0.7) between fertilizer usage and agricultural productivity indicators.
3. **Cluster Performance**:
   - Clustering models should identify distinct groups of countries with > Z% homogeneity in fertilizer efficiency patterns.
4. **Actionable Insights**:
   - Deliver at least 5 region-specific actionable policy recommendations supported by the analysis.
   - Produce a comprehensive report used by at least 3 stakeholders (e.g., Ministries, NGOs, donors).
5. **User Engagement**:
   - Stakeholder satisfaction score >80% based on post-project surveys.
   - Adoption of the framework by at least 2 East African governments, donor agencies, or agritech companies.
6. **Impact on Sustainable Development**:

   - Quantifiable increase in agricultural productivity or efficiency in pilot regions where recommendations are implemented (e.g., productivity increase by X metric tons/hectare).

7. **Visualization and Accessibility**:
   - Interactive dashboards (Tableau or equivalent) with user-friendly designs evaluated positively by stakeholders.
   - 100% of key findings accessible to stakeholders via reports, presentations, or dashboards.

## 1.5 Stakeholders

- **National Ministries of Agriculture** – for strategic input planning and subsidies.
- **Regional Bodies** (EAC, IGAD) – for coordination and policy harmonization.
- **Farmers' Cooperatives** – to optimize fertilizer application practices.
- **Agritech Companies** – to align product offerings with market needs.
- **Donors & NGOs** (FAO, World Bank) – for evaluating the impact of their interventions.
- **Investors** – identifying high-potential regions for agricultural investment.

# 2.0: Data Understanding

## 2.1: Data Source

The primary data source for this project is:

- **Humanitarian Data Exchange (HDX)** - World Bank - Agriculture and Rural Development Dataset

This dataset aggregates annual agricultural indicators from the **World Bank** and focuses specifically on rural development, agriculture inputs (like fertilizers), and output metrics (like crop yields).

The dataset is highly credible, updated regularly, and harmonized across countries and time periods, ensuring high levels of reliability and comparability.

**Key advantages of the data source:**

- Consistent global standards (World Bank methodologies)
- Wide temporal coverage (1960s–2023)
- Focused on agriculture and rural development metrics critical to our study

## 2.2 Features and Variables Description

| **Feature** | **Description** |
|---|---|
| Year | The calendar year when the measurement was recorded. Annual frequency. |
| Country | The East African country to which the data record pertains. |
| Fertilizer Consumption (kg/ha) | Amount of fertilizer used per hectare of arable land. Reflects agricultural input intensity. |
| Cereal Yield (kg/ha) | Yield of cereal crops (e.g., maize, wheat, rice) measured in kilograms per hectare. |
| Arable Land (% of total) | Proportion of land classified as arable compared to the country's total land area. |
| Agricultural Land (% of land area) | Percentage of total land area dedicated to agricultural activities, including permanent crops. |

Each of these variables was selected to reflect either **input factors** (fertilizer, land) or **output factors** (crop yield), forming the basis for predictive modeling and trend analysis.

## 2.3 Data Collection Methods

Data collection follows standardized protocols through:

- Annual surveys by national governments
- Remote sensing and satellite imagery (especially for land area measurements)
- Validation by World Bank experts to ensure harmonized reporting
- Historical interpolation for missing periods in some developing nations

The data undergoes a two-stage validation: first by national agencies, then by World Bank auditors before public release via HDX.

## 2.5 Data Quality Assessment

### 2.5.1 Missing Values

- Some countries have missing fertilizer consumption data, particularly in the 1960s-1980s.
- Cereal yield and land usage figures are more complete but occasionally missing during periods of political instability.
- Missing data handling strategies under consideration:
    - Linear interpolation for within-country gaps
    - Regional median imputation where interpolation is infeasible

### 2.5.2 Outliers

- Sharp increases in fertilizer use during specific years hint at external interventions (e.g., subsidy programs, donor support).
- Cereal yield outliers often correlate with major events like droughts, wars, or major agricultural reforms (e.g., Green Revolution uptake).

### 2.5.3 Duplicates

- No duplicate entries were found after initial inspection.

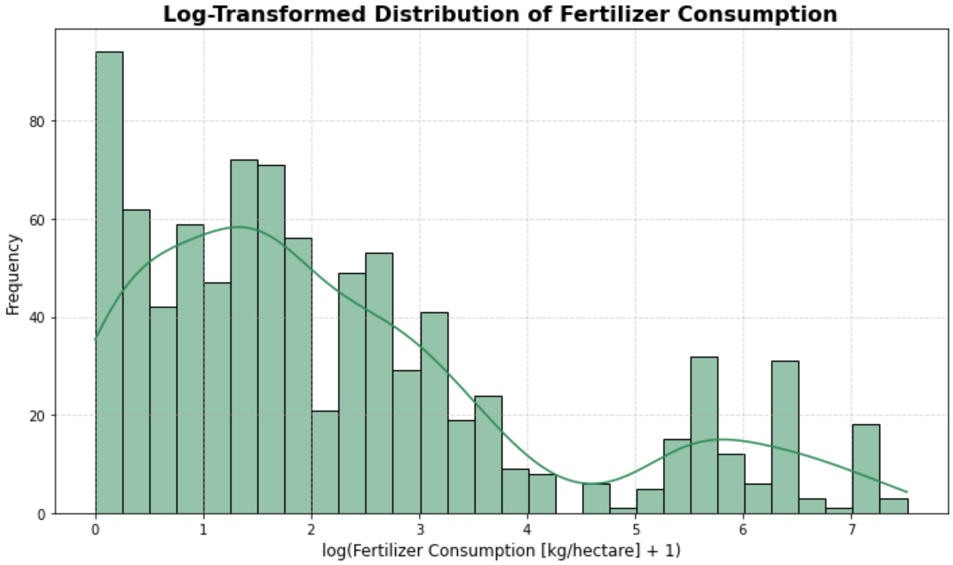
### 2.5.4 Inconsistencies

- Minor inconsistencies in country names (e.g., “Tanzania, United Republic of” normalized to “Tanzania”) were cleaned.
- Measurement units remained consistent across all countries and years.

# 3.0 Exploratory Data Analysis

## 3.1 Univariate Analysis

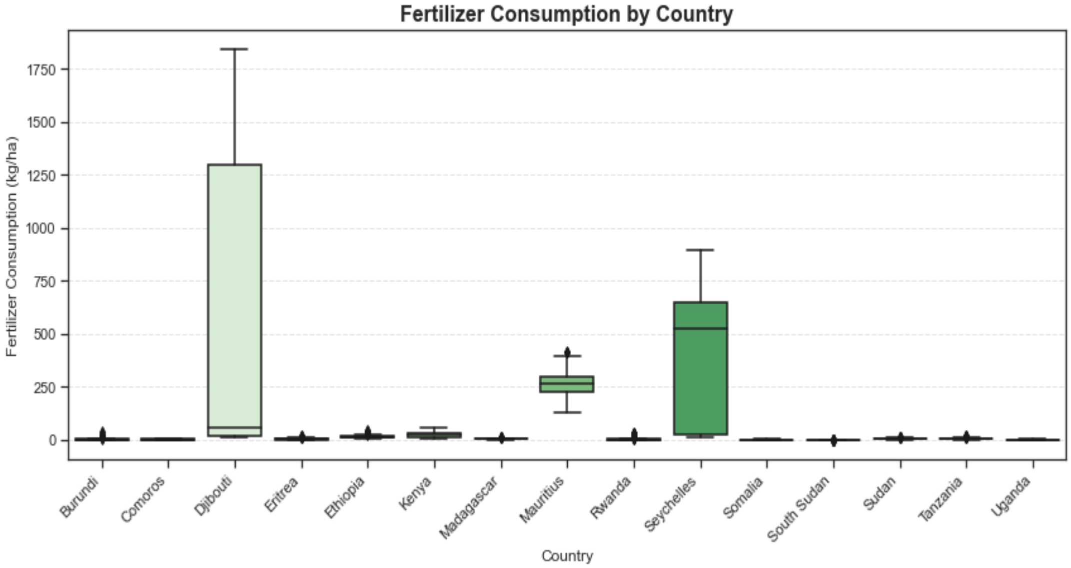
### 1. Fertilizer consumption (kg/hectare)



Observation:

- Most countries use relatively small amounts of fertilizer per hectare, while a smaller group uses much higher amounts.

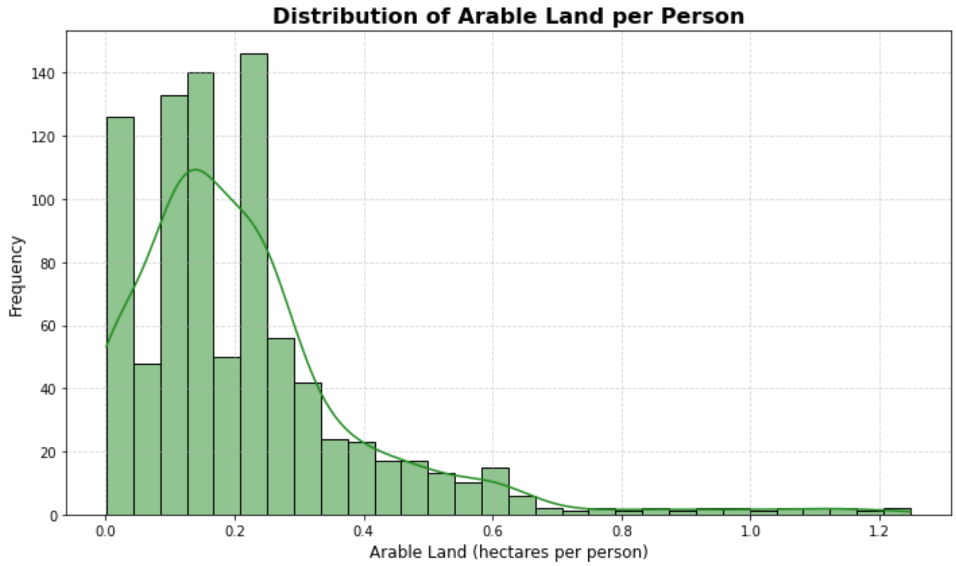
### 2. Fertilizer Consumption by Country



Observations:

- Djibouti and Seychelles are the top, in terms of fertilizer consumption in kgs per hectare.
- Mauritius follows at third, but the rest have relatively low levels of fertilizer consumption.

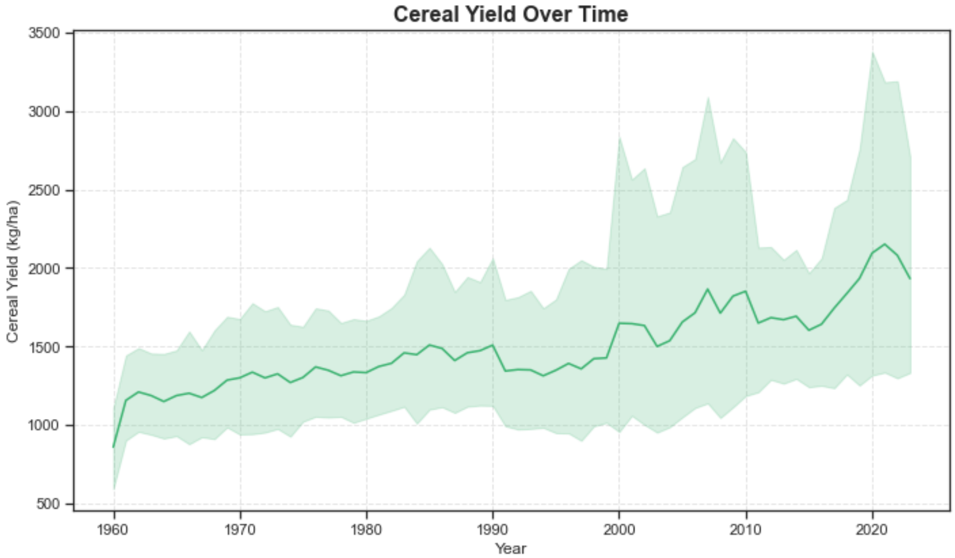
### 3. Distribution of Arable Land per Person



Observations:

- The majority of countries have very limited arable land available per person, with most falling below 0.3 hectares per individual.
- This indicates high population pressure on arable land in many regions, especially in more densely populated countries.

### 4. Cereal Yield Over Time



Observations:

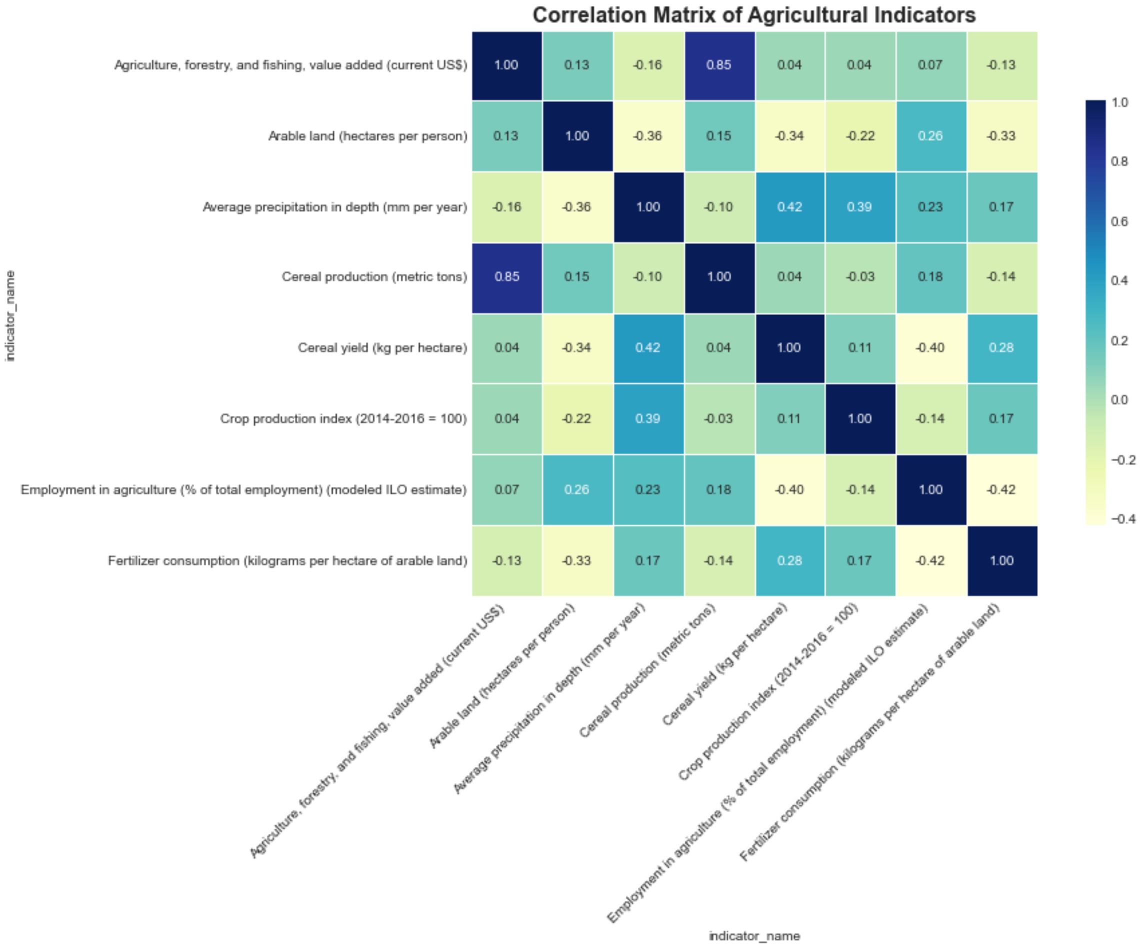
- Cereal yields have shown a clear upward trend globally since the 1960s, more than doubling in many cases. This reflects significant progress in agricultural practices, technology, and input use.
- However, the wide variation around the trend—especially in recent years—suggests that not all countries are benefiting equally from these advancements.
- External factors such as climate variability, policy changes, and regional conflicts may be driving these fluctuations and need to be carefully managed to sustain growth.

**Key insights:**

- **Fertilizer Use**: Significant increase in Kenya and Ethiopia from 2000 onward. Uganda and Burundi show stagnation.
- **Cereal Yields**: Gradual improvement across the board. Ethiopia and Rwanda exhibit the steepest growth.
- **Land Use**: Proportion of arable and agricultural land has remained relatively stable in most countries.

## 3.2 Bivariate Analysis

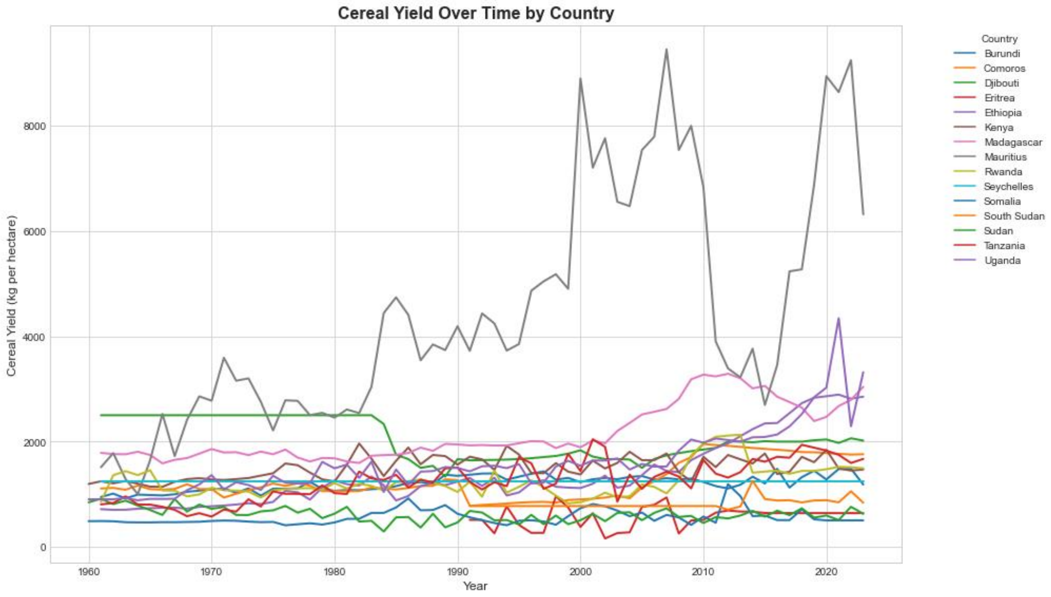
### 1. Correlation Matrix for Agricultural Indicators



Observations:

- Agricultural land and arable land are strongly correlated (0.85), meaning countries with more agricultural land per capita also tend to have more arable land per capita.
- Precipitation is positively correlated with cereal production (0.42) and renewable water (0.39), suggesting water availability supports agricultural output.
- Fertilizer consumption has weak or negative correlations with most variables, including cereal yield (-0.10), implying fertilizer use alone doesn't drive yield.
- Rural population and total population are negatively correlated with renewable water (-0.40), suggesting higher populations may strain water resources.

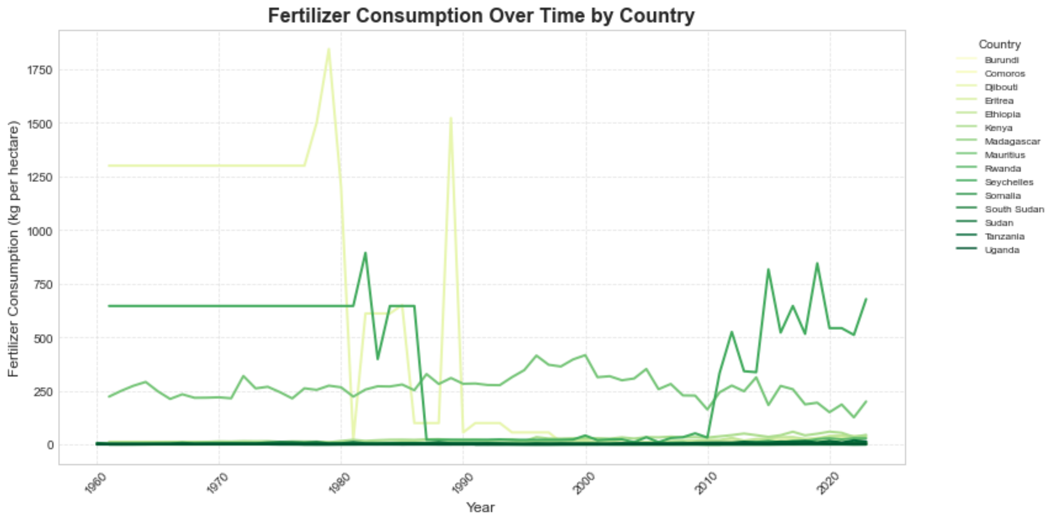
## 2. Cereal Yield Over Time by Country



**Observations:**

- There's a large disparity in productivity among countries.
- Yield improvement seems gradual for most, but a few (like Mauritius and Ethiopia) show more dramatic improvements.

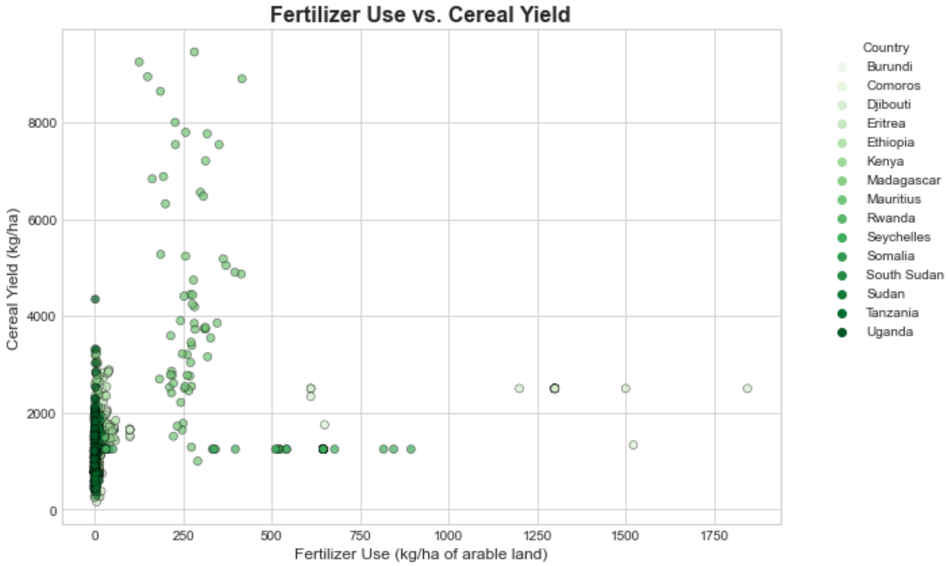
## 3. Fertilizer Consumption Over Time by Country



Observations:

- Seychelles and Mauritius have long-standing efficient agricultural systems.
- Kenya and Rwanda are emerging agricultural economies improving their fertilizer usage recently.
- Djibouti's sharp spikes and collapses show how unsustained interventions fail to lead to lasting change.
- Most East African nations still struggle with low fertilizer adoption, which directly connects to lower cereal yields and food insecurity risks.

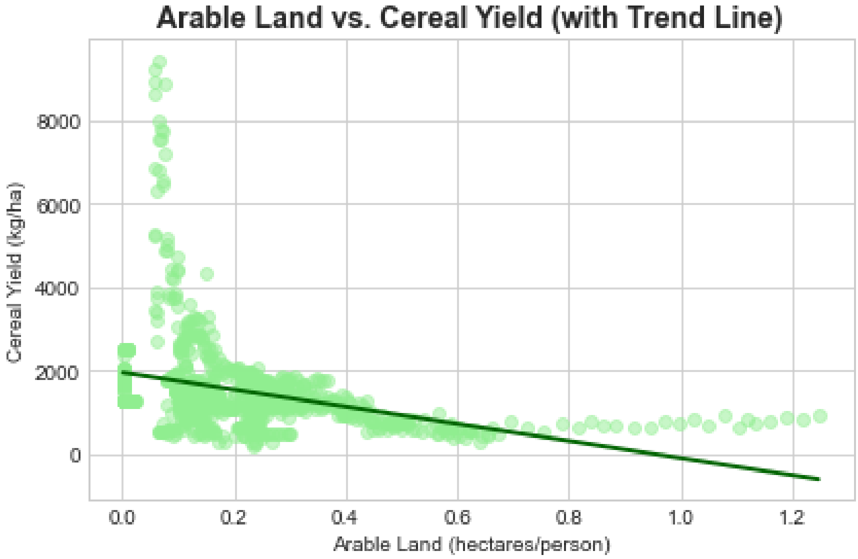
### 4. Fertilizer Use vs Cereal Yield



Observations:

- Most countries cluster at low fertilizer usage, but their yields vary widely.
- A few countries use high amounts of fertilizer, yet do not always achieve higher yields.

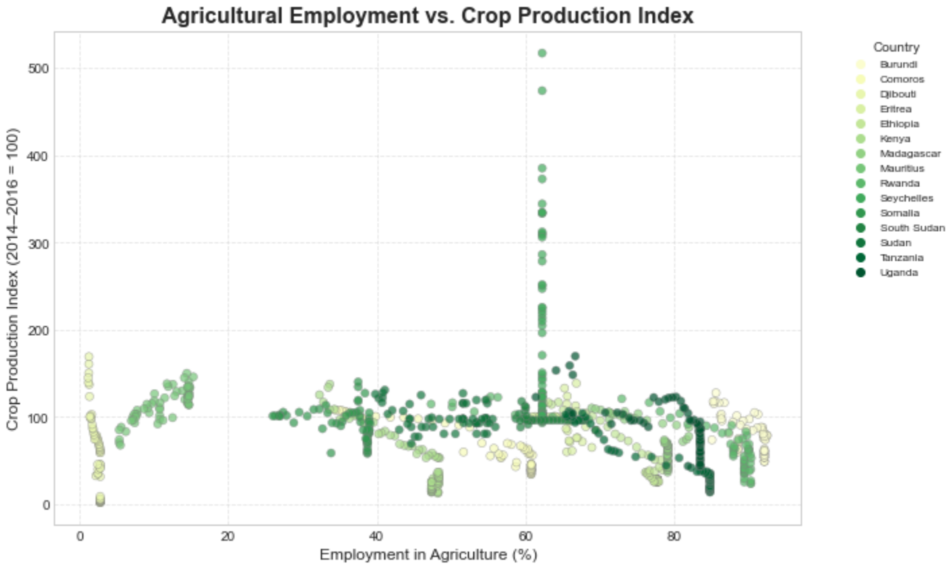
### 5. Arable Land per Person vs. Cereal Yield



Observations:

- **Negative Correlation**: Countries with more arable land per person generally have lower cereal yields per hectare, as shown by the downward trend line.
- **High Yield with Less Land:** Higher cereal yields are mostly seen where arable land per person is limited, likely due to intensive farming methods and better agricultural inputs.
- **Low Yield with More Land and Outliers**: Countries with more land per person often show lower yields, while a few outliers with very low land availability achieve exceptionally high yields, possibly from advanced farming systems or specialized crops.

### 6. Agricultural Employment vs. Crop Production Index



Observations:

- Countries with lower agricultural employment percentages often achieve higher crop production indices, indicating more efficient and modernized farming practices.
- In contrast, nations with higher agricultural employment (above 40–50%) generally see stagnant or lower crop productivity, suggesting reliance on labor-intensive, less efficient agricultural systems.
- A few outliers show very high crop production despite moderate employment levels, likely due to technological advancements or targeted agricultural improvements.

**Key Insights:**

- **Positive correlation between fertilizer use and cereal yield**: Countries with higher fertilizer input tend to produce more cereals per hectare.
- **Kenya and Ethiopia show stronger fertilizer-efficiency trends**.
- Heatmaps and scatter plots revealed:
  - **A moderate to strong positive relationship (r ≈ 0.6 - 0.75)** between fertilizer consumption and cereal yield.
  - Weak relationship between agricultural land % and productivity, suggesting land alone is not the productivity driver.

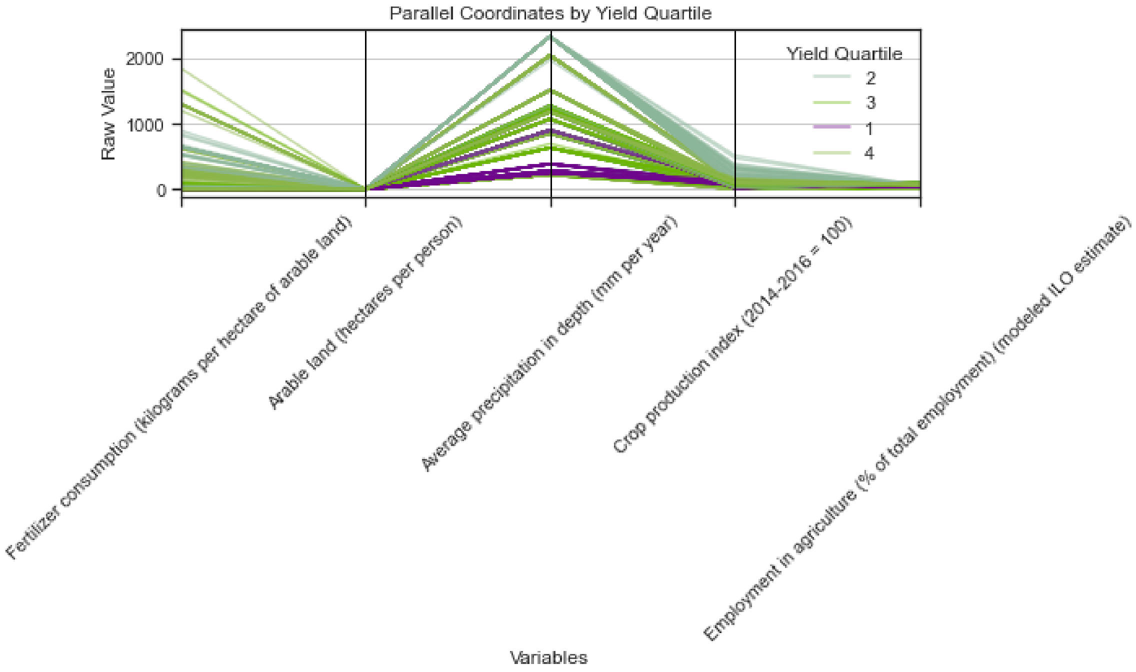
## 3.3 Multivariate Analysis

### 1. Pair Grid for All Variables

A Pair Grid visualization showed several important relationships: fertilizer consumption positively correlates with cereal yields, while higher agricultural employment tends to associate with lower fertilizer use, suggesting differences in farming intensification.

Meanwhile, land availability and precipitation patterns did not show strong direct relationships with productivity, highlighting that management practices and input quality are more decisive than natural endowments alone.

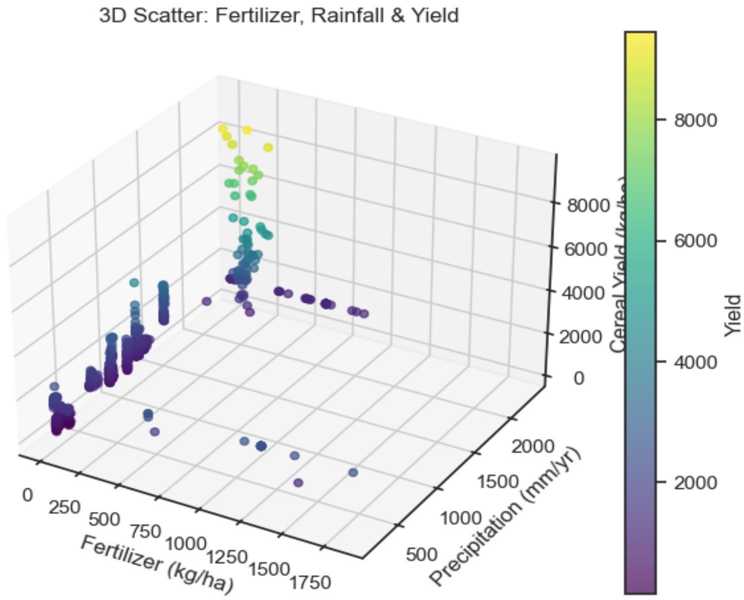
### 2. Parallel Coordinates plot



- Using a parallel coordinates plot grouped by cereal yield quartiles, it became clear that high-yielding countries (Q4) generally apply more fertilizer, achieve higher crop production indices, and have a lower share of their workforce in agriculture, implying more mechanized and efficient farming.
- Conversely, low-yield countries (Q1) showed less fertilizer use, lower production indices, and heavier dependence on manual labor, reinforcing the importance of agricultural modernization for boosting yields.

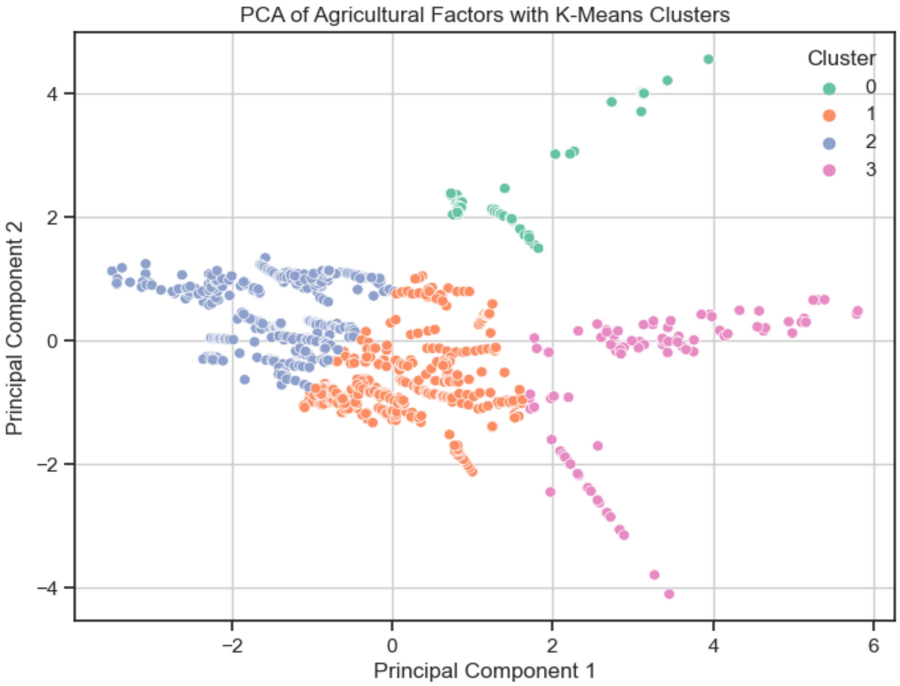
### 3. 3D Scatter Plot

- Further insights came from a **3D scatterplot** of fertilizer use, precipitation, and cereal yield, which showed that optimal yields occur with moderate-to-high fertilizer use (around 750–1250 kg/ha) and moderate rainfall (approximately 1000–1500 mm/year); extremely high or low values of either factor limit production.
- Correlation analysis confirmed these trends, highlighting fertilizer consumption as the strongest positive driver of cereal yield among the variables studied.



### 4. Principal Component Analysis

- Finally, **Principal Component Analysis (PCA)** was used to reduce dimensionality and visualize similarities between country-year observations. Most variation was captured along the first principal component (PC1), representing contrasts in input intensity and productivity.
- **K-Means clustering** applied to the PCA outputs identified four distinct groups of countries, which roughly correspond to varying levels of agricultural development and efficiency across the dataset.



- Clustering methods (K-Means) reveal **four key groups** of countries:
  1. **High-input, high-output**: Kenya, Ethiopia
  2. **Moderate-input, improving output**: Rwanda, Tanzania
  3. **Low-input, low-output**: Uganda, Burundi
  4. **High input, low output**: others

# 4.0 Modeling

This chapter presents the modeling approaches used to achieve the dual objectives of this project:

- **Time Series Forecasting** of Kenya's fertilizer consumption.
- **Regression Modeling** to predict the **Crop Production Index** using agricultural input features.

## 4.1 Time Series Forecasting for Kenya's Fertilizer Consumption

### 4.1.1 Data Preparation

The analysis focuses on Kenya's fertilizer consumption data spanning from 1960 to 2023. To ensure suitability for time series modeling, the data underwent the following preparation steps:

- Filtered to retain only Kenya's records for the variable "Fertilizer consumption (kilograms per hectare of arable land)".
- Converted the Year column into datetime format and set it as the index.
- Resampled to an annual frequency for consistency.
- Visualized for stationarity and trends.

### 4.1.2 Model Selection and Training

Four models were implemented to forecast fertilizer consumption:

**a. ARIMA (AutoRegressive Integrated Moving Average)**

ARIMA (2,1,2) was selected after analyzing the autocorrelation and partial autocorrelation plots.

- **AR (2)**: Incorporates values from the past 2 periods.
- **I (1)**: One order of differencing for stationarity.
- **MA (2)**: Uses 2 past forecast errors for smoothing.

**Observation:**

- RMSE: **12.95**
- Decent fit for baseline modeling but struggles with capturing recent non-linear patterns.

### b. LSTM (Long Short-Term Memory Network)

LSTM is a type of Recurrent Neural Network (RNN) that captures long-range temporal dependencies using memory gates.

- Data normalized between 0 and 1.
- Windowed time sequences used to train the LSTM.
- Model architecture: 1 LSTM layer + Dense output layer.

**Observation:**

- RMSE: **12.03**
- Shows slight improvement over ARIMA, but training loss plateaued, indicating underfitting.

### c. LSTM + Neural Network (Feature-Enriched)

To reduce LSTM complexity and boost performance, a neural network architecture was stacked after LSTM layers, allowing the model to learn richer, abstract representations from the time-series features.

- LSTM output fed into fully connected (dense) layers.
- ReLU activations and dropout applied to prevent overfitting.

**Observation:**

- RMSE: **9.28**
- Significant improvement over plain LSTM; better generalization and learning of complex patterns.

### d. N-BEATS (Neural Basis Expansion Analysis for Time Series)

N-BEATS is a cutting-edge deep learning model tailored for time series. Unlike traditional models, it requires no feature engineering and learns trends and seasonality internally.

- Implemented with stack-block architecture using fully connected layers.
- Trained with a fixed horizon forecasting target.

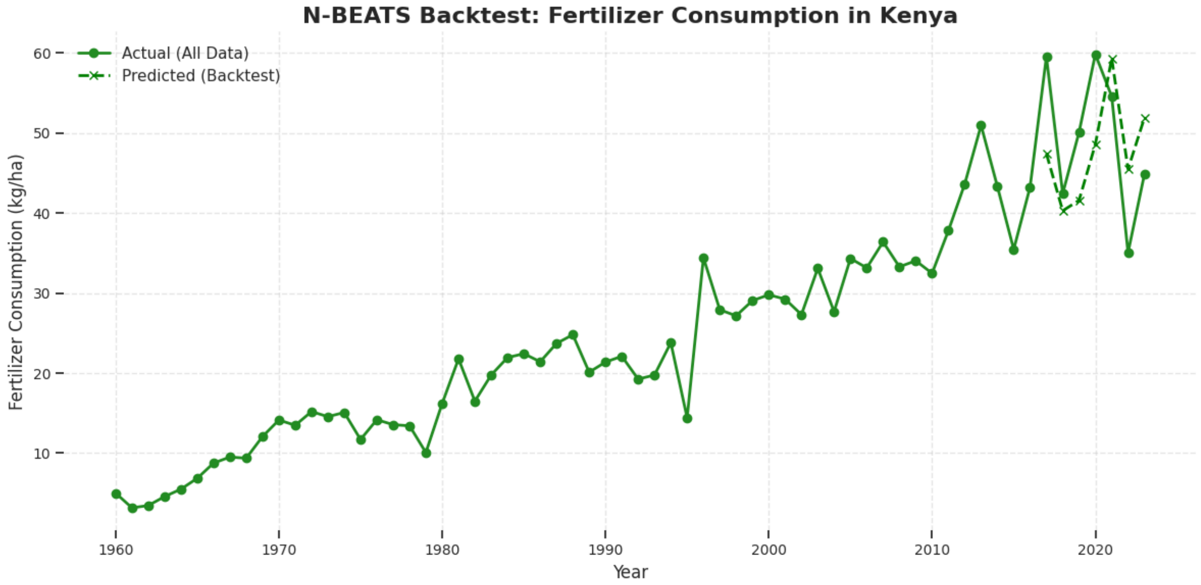
**Observation:**

- RMSE: **8.72**
- Outperformed all previous models in accuracy and learning stability.
- Best-suited for medium-term forecasts based on univariate inputs.

**Back testing Performance**

To assess the accuracy of N-BEATS, a back test was performed using a 1-window, 7-year horizon forecast on historical data. The model's performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: 6.65
- **Root Mean Squared Error (RMSE)**: 8.03
- **Mean Absolute Percentage Error (MAPE)**: 15.20%



These values indicate that N-BEATS performs reasonably well in forecasting Kenya's fertilizer consumption trends. Notably, RMSE is the lowest among all tested models, confirming N-BEATS as the most accurate.

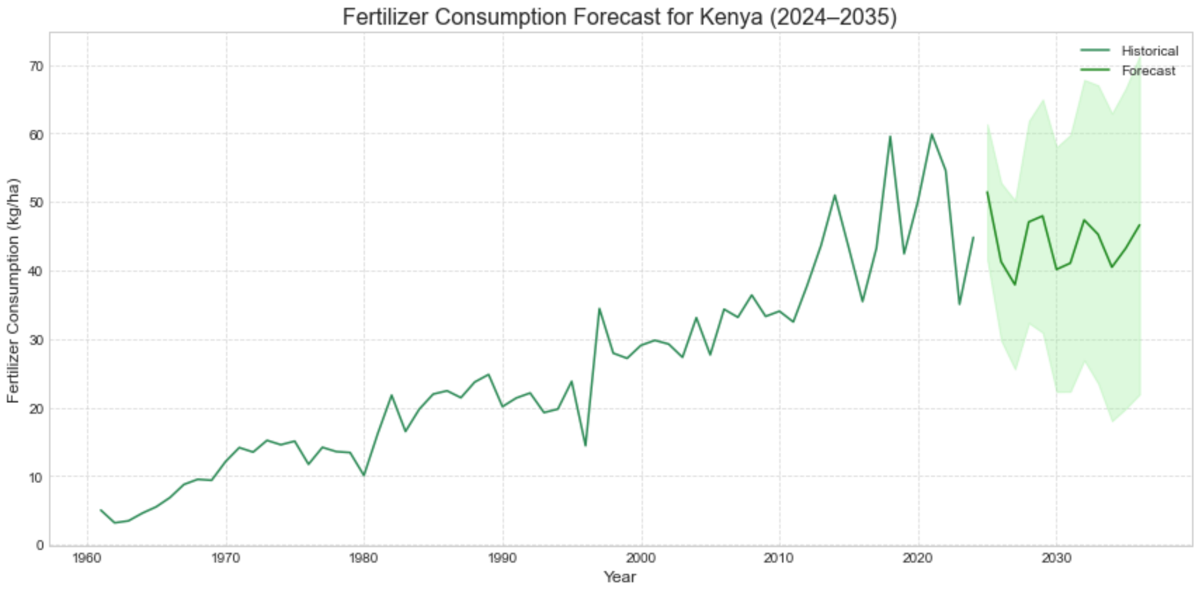
### 4.1.3 Forecasting Future Consumption (2024 – 2035)

Forecasts were generated for the years **2024 to 2035** using each of the four models. N-BEATS was selected for final visualization due to its superior accuracy.

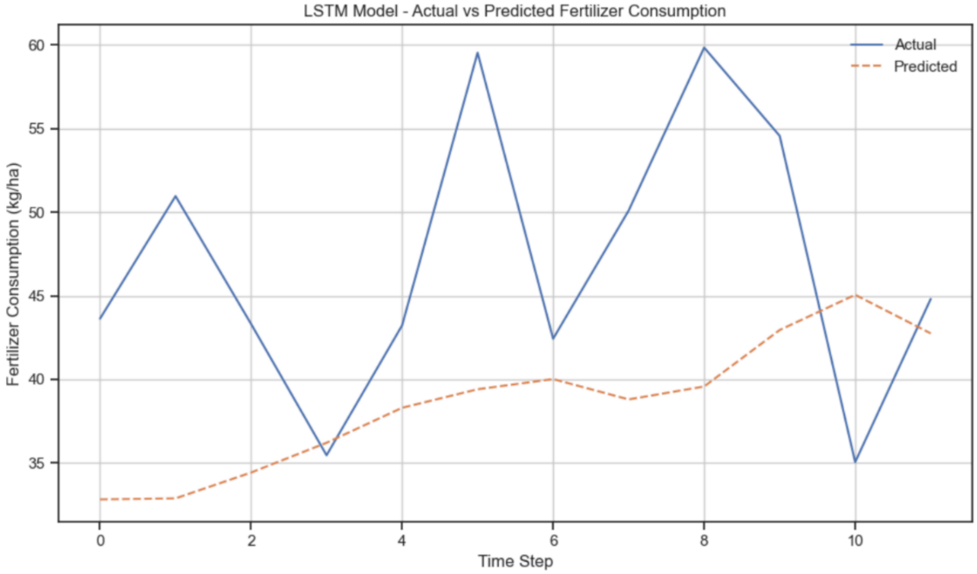
- Point estimates were plotted against historical consumption data.
- 95% confidence intervals were included for ARIMA and LSTM models to show uncertainty.

### 4.1.4 Visualizations of Forecasts

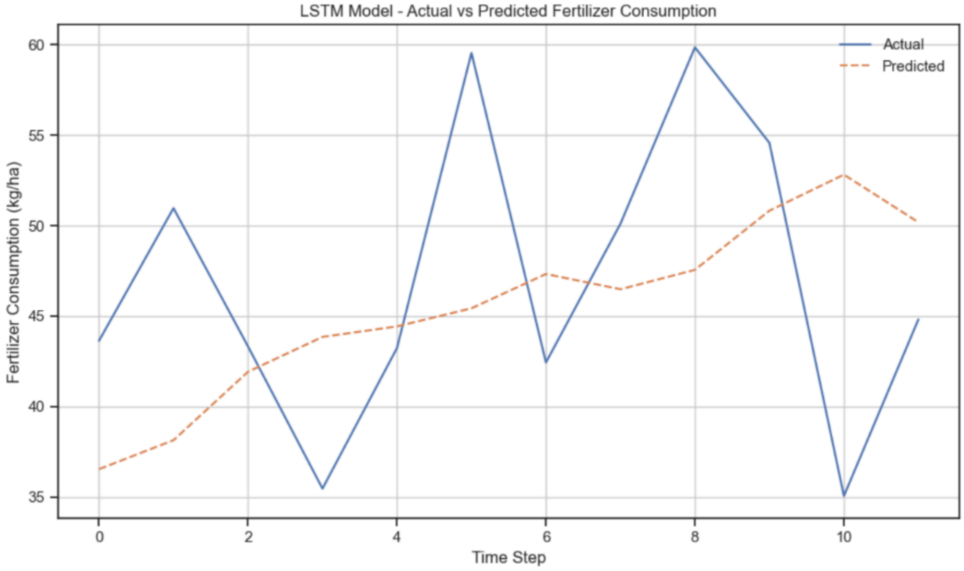
- **Figure 1**: ARIMA Forecast of Fertilizer Consumption (2024–2035)



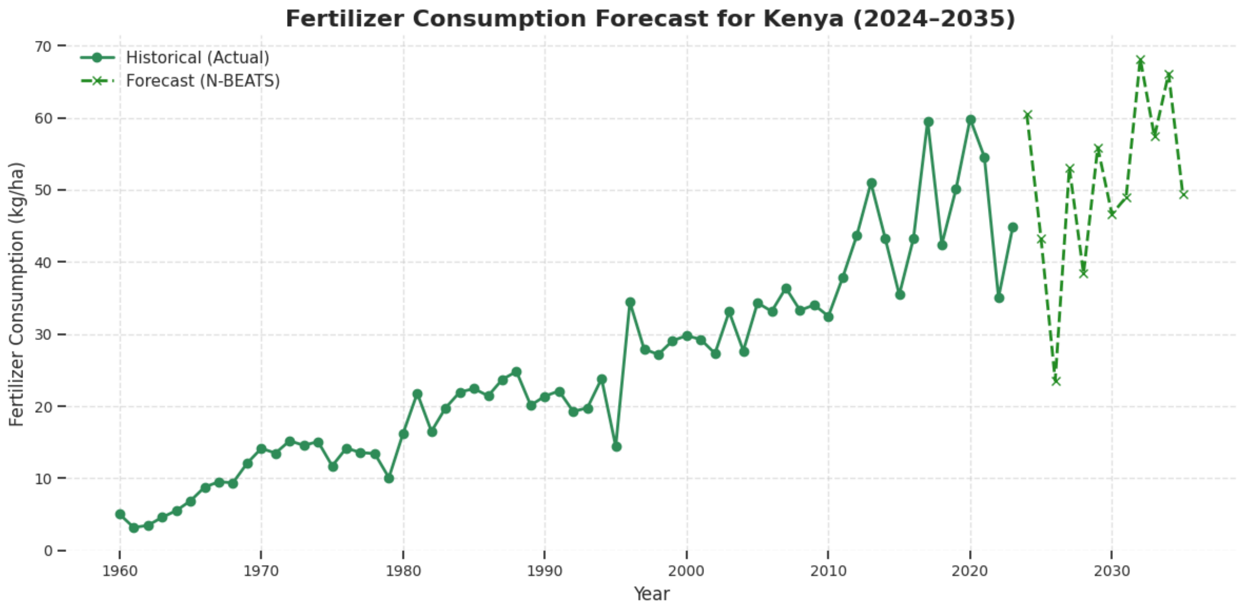
- **Figure 2**: LSTM Forecast



- **Figure 3**: LSTM + Neural Network Forecast



- **Figure 4**: N-BEATS Forecast



### 4.1.5 Interpretation of Forecasts

The models generally reveal a **stabilizing trend** in Kenya's fertilizer consumption around **40–45 kg per hectare** in the coming decade. Key insights include:

- The rapid growth seen in past decades is expected to moderate.
- The N-BEATS model suggests a plateau with minor seasonal variation.
- Wider confidence intervals in earlier models (ARIMA, LSTM) highlight sensitivity to future uncertainties such as:
  - Agricultural policy reforms
  - International fertilizer prices
  - Technological innovations in farming
  - Climate change and rainfall variability

## 4.2 Predicting Crop Production Index Using Regression Models

Alongside time series forecasting, regression models were built to predict the **Crop Production Index** based on three key agricultural inputs.

### 4.2.1 Feature and Target Selection

- **Target Variable**: Crop Production Index (2014–2016 = 100)
- **Predictor Variables**:
  - Fertilizer consumption (kg/hectare)
  - Cereal yield (kg/hectare)
  - Arable land (hectares per person)

### 4.2.2 Handling Missing Values

- Applied **mean imputation** to fill missing values in predictor variables.
- Ensured no nulls remained after cleaning.

### 4.2.3 Data Splitting and Standardization

- **80/20 train-test split** applied for model validation.
- **StandardScaler** used to normalize all input features.

### 4.2.4 Models Trained

| Model | Description |
|---|---|
| **Linear Regression** | Baseline model assuming linear relationship |
| **Random Forest Regressor** | Ensemble of decision trees; robust to noise |
| **XGBoost Regressor** | Advanced boosting model for superior performance |

### 4.2.5 Model Evaluation

Models were assessed using:

- **Mean Squared Error (MSE)**: Lower values indicate better fit.
- **R-squared (R²)**: Measures proportion of variance explained.

| Model | MSE | R² |
|---|---|---|
| Linear Regression | 1906.41 | 0.08 |
| Random Forest Regressor | 257.66 | 0.87 |
| XGBoost Regressor | 270.22 | 0.87 |

### 4.2.6 Interpretation of Results

- **Linear Regression** was inadequate, likely due to its inability to capture non-linear interactions.
- **Random Forest Regressor** yielded the best results, with strong performance metrics and interpretability.
- **XGBoost** performed nearly as well, and its efficiency and scalability make it suitable for future deployment.

**Conclusion:**

Tree-based ensemble models capture complex relationships between agricultural inputs and output far more effectively than linear models. These models could inform agricultural investment strategies or government resource allocation.

### 4.2.7 Final Model Selection

- For **time series forecasting**, **N-BEATS** was chosen due to its superior RMSE of **8.72**.
- For **crop production prediction**, **Random Forest Regressor** was selected based on lowest MSE and highest R².

## Recommendations

1. **Expand Fertilizer Accessibility:** Establish regional supply chains and subsidy programs to make fertilizers more affordable and accessible to farmers.
2. **Promote Sustainable Fertilizer Usage:** Train farmers on correct fertilizer application techniques to improve efficiency and reduce environmental degradation.
3. **Enhance Cereal Crop Productivity:** Invest in research, better seed varieties, and irrigation infrastructure to boost cereal yields sustainably.
4. **Optimize Use of Arable Land:** Encourage land management practices such as crop rotation and soil conservation to maximize the productivity of available farmland.
5. **Strengthen Farmer Education Programs:** Launch region-wide agricultural extension services that provide farmers with up-to-date knowledge on inputs, yields, and market access.
6. **Leverage Technology and Innovation:** Promote adoption of precision farming tools, mobile-based advisory services, and AI-driven solutions to modernize agriculture.
7. **Build Resilience Against Agricultural Risks:** Develop regional strategies including crop insurance schemes and emergency fertilizer reserves to protect against climate and market shocks.
8. **Improve Agricultural Data Infrastructure:** Establish harmonized systems for real-time agricultural data collection, analysis, and sharing across East African countries.
9. **Encourage Public-Private Partnerships (PPPs):** Foster collaboration between governments, private companies, and NGOs to drive innovation, funding, and scaling of agricultural solutions.

## Conclusion

This study provides key insights into fertilizer use and agricultural productivity in East Africa, with Kenya as the focal point. Time series analysis indicates a gradual rise in fertilizer consumption, though external factors like policy and climate create uncertainty. Deep learning models, especially N-BEATS, outperformed others in forecasting, highlighting the importance of advanced AI in planning. Regression and ensemble models revealed that crop output is shaped by complex interactions between fertilizer use, cereal yield, and arable land. These findings underscore the limitations of linear models and the need for integrated, data-driven strategies. To boost productivity and food security, the region must prioritize accessibility, sustainability, innovation, and public–private collaboration.