Crop Production and Yield Prediction Project

Presentation

Our Team

Kennedy Wamwati

Ray Onsongo

Abigael Musyoka

Marylyne Ingwe

Elizabeth Kiluu

Pauline Kimenzu

MaryBennah Kuloba

Business Understanding

Agricultural production in Kenya is highly sensitive to climatic variability, land-use dynamics, and long-term structural transformation. Reliable prediction of crop production is therefore critical for food security planning, early warning systems, and evidence-based agricultural policy formulation.

This project evaluates the performance of baseline statistical models and ensemble machine learning methods in predicting crop production in Kenya, with particular attention to methodological rigor, temporal structure, and avoidance of target leakage.

Business Understanding

Agriculture plays a central role in Kenya's economy and food security, making it essential to understand how the production of crops and livestock products has changed over time.

This project seeks to analyze historical agricultural production data to identify long-term trends, variations and key contributors to national output.

By examining production quantities across different years and products, the analysis aims to answer questions such as which agricultural products have experienced sustained growth or decline and how production patterns have evolved over time.

The insights generated from this analysis are relevant to policymakers, agricultural planners, development organizations, and agribusiness stakeholders who rely on data-driven decision-making.

Data Sources

The dataset is sourced from FAOSTAT and contains annual observations of crops and livestock products in Kenya from 1961 to 2021.

Data Analysis Scope

This project focuses exclusively on *crop production*, retaining only the following elements:

1. Total Production (tonnes), which is the primary prediction target , with yield modeled indirectly through leakage-safe transformations.
2. Area harvested (hectares)
3. Yield (hectograms per hectare)

Data Cleaning

- ✓ Selection of Relevant Variables
- ✓ Handling Missing and Inconsistent Values
- ✓ Standardization
- ✓ Renaming of Variables
- ✓ Filtering based on data availability and reliability
- ✓ Construction of time-ordered panel structures
- ✓ Outlier assessment and diagnostic

- ✓ Retained crop-related elements only
- ✓ Enforced consistent units:
- ✓ Production → tonnes
- ✓ Area harvested → hectares
- ✓ Yield → hg/ha
- ✓ Converted data types and removed invalid or missing records
- ✓ Normalized categorical text field

Feature Engineering

Panel Construction

The dataset was reshaped from long to wide format, producing one row per:

1. Crop (item)
2. Year (Time-Series)

Core variables:

1. production_t
2. area_harvested_ha
3. yield_hg_per_ha

Missing yields were derived when both production and harvested area were available:

Yields were converted to tonnes per hectare.

Time-Based Feature Engineering

The following features were created:

1. Lagged variables (1-3 years)
2. Moving averages (3-year and 5-year)
3. Year-over-year growth rates
4. Normalized time trend

Outlier indicators

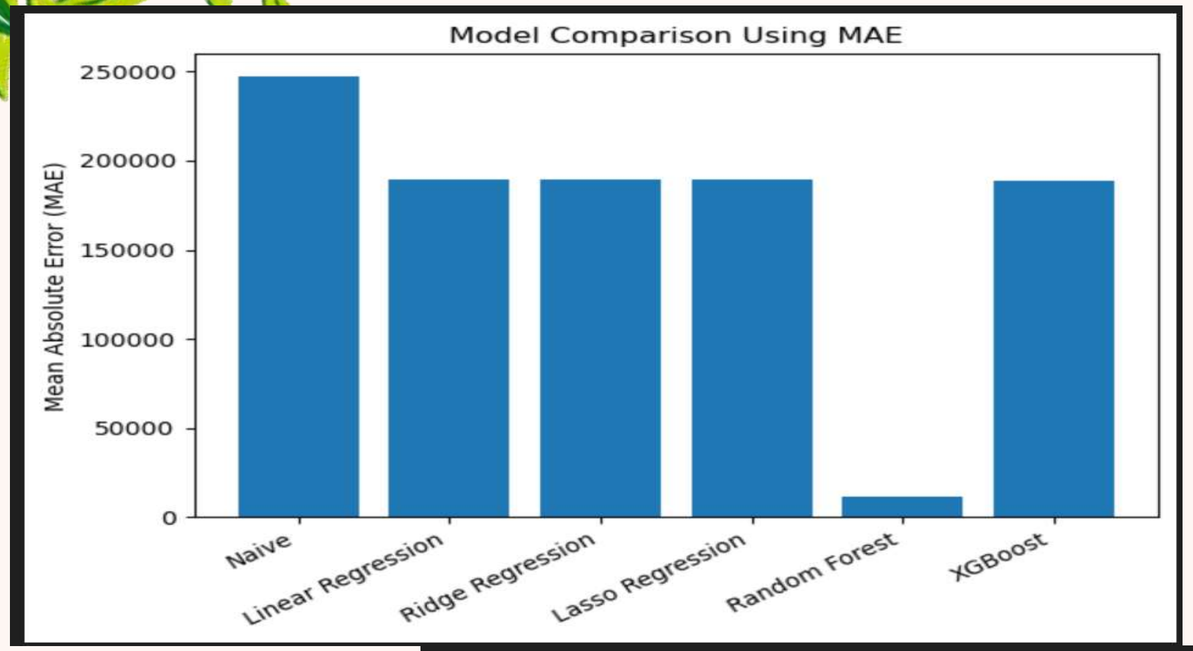
1. One-hot encoded flag classes The final modeling dataset contains 5,764 rows across 139 crops.

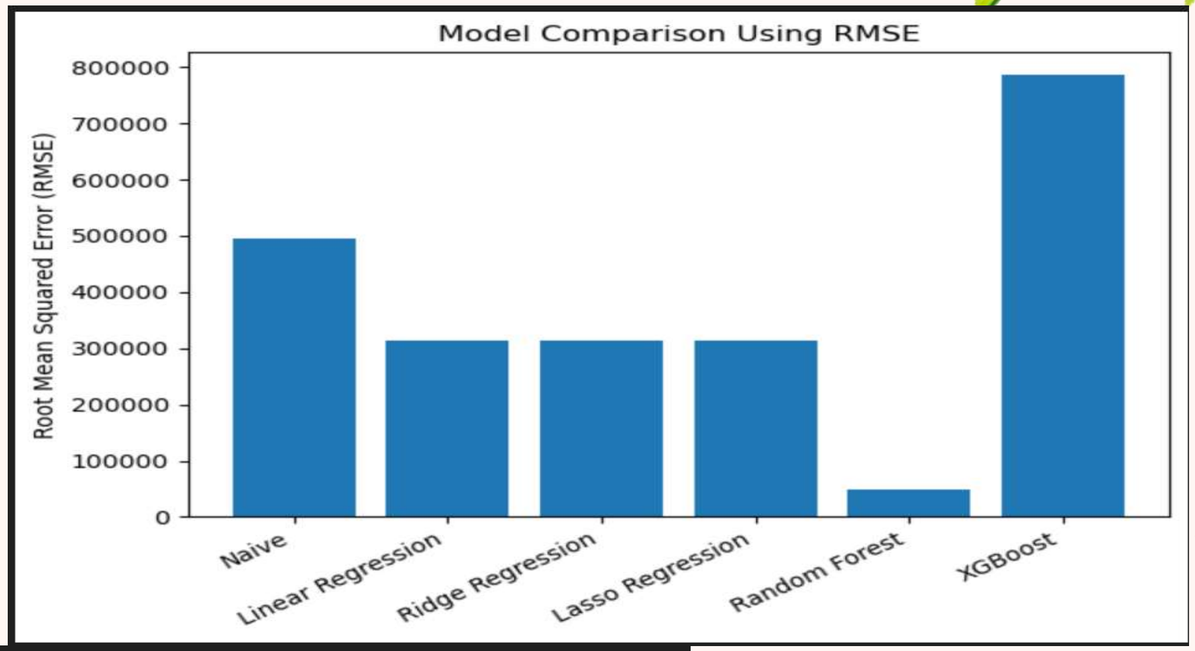
Modelling Approach

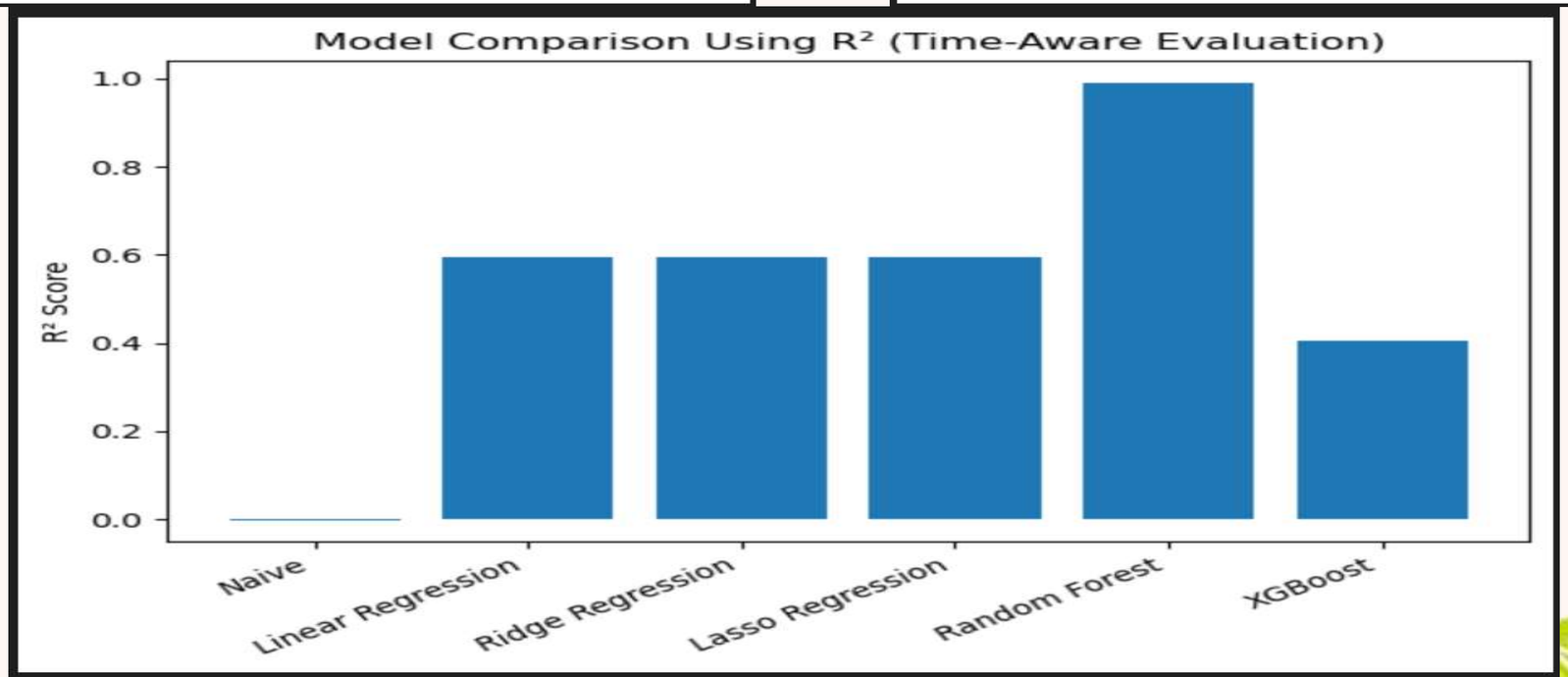
Models compared

	Rank	Model	MAE	RMSE	R2
0	1	Random Forest	11210.038975	48392.215783	0.990443
1	2	Ridge Regression	189360.031075	314454.438140	0.596480
2	3	Lasso Regression	189411.192101	314467.899576	0.596446
3	4	Linear Regression	189411.197333	314467.900679	0.596446
4	5	XGBoost	188986.483479	786760.208399	0.406632
5	6	Naive	247424.677709	495445.720711	-0.001709

Model Performance







Final Model Selection

Random Forest clearly outperforms all other models across all evaluation metrics.

- It achieves the lowest MAE and RMSE and an R^2 value close to 1.0, indicating excellent predictive performance.
- This suggests that the dataset contains strong non-linear patterns and interactions that are effectively captured by ensemble tree-based models.

Thus Random Forest becomes our final model due to its superior performance across all evaluation metrics and its ability to capture complex patterns in the data.

Practical Deployment

- We first confirmed that the trained models exist and that the project has the correct training data and features ready to use.
- We then verified the data-cleaning step used by the API, making sure the same inputs used in training are processed correctly for live predictions.
- Next, we tested the prediction service with sample requests to confirm it returns valid results and handles missing fields safely.
- After that, we packaged the model and API into a Docker container so it can run consistently on any computer.
- Finally, we documented how to run it locally and how to deploy it to cloud platforms (AWS, Google, Azure, or a private server), and provided a deployment checklist, environment settings, and API documentation so teams can monitor and use the service reliably.

Conclusion

This study evaluated multiple predictive models using MAE, RMSE, and R^2 to identify the most suitable approach for the given dataset.

- The naive baseline performed poorly, confirming the need for more advanced methods.

- While linear, ridge, and lasso regression models achieved moderate performance but were limited by their inability to capture complex, non-linear relationships.

- XGBoost showed mixed results, with some improvement in average error but weaker overall generalization.

- In contrast, the Random Forest model consistently outperformed all other models, achieving the lowest error values and the highest R^2, indicating excellent predictive accuracy.

- These results demonstrate that the underlying data exhibits strong non-linear patterns and that ensemble tree-based methods, particularly Random Forest, are the most appropriate choice for this prediction task.

Thank You