;

KCA UNIVERSITY

BACHELORS IN DATASCIENCE

BDS 4105 :FINAL YEAR PROJECT 1.

21/04428 NJOROGE HALDLINE MUKAMI

SUPERVISOR: GLADYS MANGE

PROJECT TITLE : MAIZE CROP YIELD PREDICTION AND ADVISORY SYSTEM .

Declaration of Originality

I declare that this project proposal is my original work and has not been submitted to any other institution or examination body. All sources used or quoted have been duly acknowledged.

Signed: _____

Date: _____

Table of Contents

1. Introduction

Agriculture plays a major role in Kenya's economy, with maize being the country's main staple food. However, maize production has become increasingly unstable due to unpredictable rainfall patterns, poor soil fertility, and limited access to modern farming technologies. This instability directly affects food security and the income of many smallholder farmers who depend on maize as their main source of livelihood.

Recent advances in data science and machine learning have made it possible to use data-driven approaches to solve agricultural problems. By analyzing satellite data, weather patterns, soil information, and past yields, it is possible to predict future crop production accurately. Such predictive systems can help farmers make informed decisions on planting, irrigation, and fertilizer use, reducing the risk of low yields.

This project, titled **"Maize Crop Yield Prediction and Advisory System (CYPAS)"**, aims to develop a functional web-based system that predicts maize yield in Kenya using machine learning models. The system will integrate multiple data sources and provide users with insights and recommendations . This approach will support both farmers and agricultural planners in making data-informed decisions to improve productivity and stability in maize production.

2. Background

Maize is a major food crop in Kenya and is grown by most small-scale farmers across the country. Despite its importance, maize yields have remained low and inconsistent over the years. Factors such as erratic rainfall, soil degradation, and the effects of climate change have made it difficult for farmers to achieve stable production. Many farmers also rely on traditional knowledge rather than modern data-driven tools, limiting their ability to plan and respond effectively to environmental changes.

The increasing availability of open agricultural data and affordable computational tools provides a new opportunity to improve agricultural planning. Datasets such as **satellite imagery**, **weather data**, and **soil maps** can now be accessed freely and used to build predictive models. Machine learning techniques like **Random Forest**,

XGBoost, and **Neural Networks** can analyze patterns within this data to predict crop yields with good accuracy.

Developing a maize yield prediction system will bridge the gap between data science and practical agriculture. By combining prediction with simple advisory outputs, farmers and decision-makers can receive timely and localized insights. Integrating this model into a **platform** makes it easily accessible and usable by a wide range of users, even with limited technical skills. Ultimately, this study seeks to demonstrate how data-driven solutions can promote sustainable agriculture and improve food security in Kenya.

3. Problem Statement

Maize production in Kenya has remained unstable over the past decade, despite being the country's most important food crop. This instability is largely caused by unpredictable weather patterns, poor soil fertility, pest infestations, and inadequate use of modern farming technologies. As a result, farmers experience frequent yield fluctuations, leading to food shortages, increased import dependency, and reduced household income.

Current agricultural monitoring systems in Kenya mostly focus on data collection and reporting rather than prediction and decision support. Many farmers still rely on traditional experience and rainfall patterns that are no longer consistent due to climate change. This makes it difficult to plan for inputs such as fertilizer, seed varieties, and irrigation effectively.

The lack of accessible, data-driven predictive tools limits the ability of both farmers and policymakers to respond to these challenges in time. There is, therefore, a need for a system that can accurately predict maize yields using real data—such as weather, soil, and past production trends—and present these insights through an easy-to-use digital platform.

This project seeks to address this problem by developing a **machine learning–based Crop Yield Prediction and Advisory System (CYPAS)** that provides accurate yield forecasts and practical recommendations. The system will empower farmers,

researchers, and agricultural planners to make informed, timely, and data-supported decisions aimed at improving maize production and ensuring food security.

4. Proposed Solution

High-level description:
The proposed system — Crop Yield Prediction and Advisory System (CYPAS) — is a data-driven application that predicts maize yields for Kenyan counties and delivers simple, actionable advisories to users via a web interface. The system combines historical yield records, satellite-derived vegetation indices, weather variables, and soil characteristics to produce localized yield forecasts and farm-management recommendations.

Major system functionality:

Data Integration: Ingest and fuse multi-source data (FAO / national yield statistics, NASA POWER or CHIRPS weather data, MODIS/Sentinel NDVI, SoilGrids).

Preprocessing & Feature Engineering: Clean and align datasets spatially and temporally; derive seasonal aggregates and growth-period features.

Predictive Models: Train and validate regression models (baseline: Linear Regression / Random Forest; advanced: XGBoost or LSTM for temporal patterns).

Explainability: Provide feature-importance outputs (e.g., SHAP or permutation importance) to make predictions interpretable.

User Interface: Deploy a Streamlit web app where users can: select a county, view predicted yield, inspect trend charts, and download results.

Advisory Module: Offer concise recommendations (planting window suggestions, irrigation reminders, fertilizer timing) based on model output and recent weather forecasts.

Visualization & Mapping: Display interactive charts and county-level maps (Plotly / Folium) showing predicted vs actual yields and risk indicators.

Reporting: Allow export of results and summary reports as CSV/PDF for offline use.

Optional Live Data: Integrate NASA POWER or similar API to fetch near-term weather for dynamic (on-demand) predictions.

Why this solution:
This approach moves beyond static reporting to a proactive decision-support tool. By combining predictive models with a simple advisory layer and an accessible UI, CYPAS delivers timely, localized information that farmers and planners can act upon immediately. The modular design also allows future extension (other crops, finer spatial resolution, mobile integration).

Constraints & considerations:

Predictions will initially be at **county** level (scalable later to sub-county/farm level).

Model accuracy depends on data quality and temporal coverage; missing or sparse local yield records may limit performance in some counties.

The advisory outputs will be conservative and phrased as recommendations (not prescriptive actions), to respect on-ground agronomic nuance.

5. Project Objectives

5.1 General Objective

To design and develop a **machine learning–based Crop Yield Prediction and Advisory System (CYPAS)** that predicts maize yields in Kenya using multi-source data and provides users with accurate forecasts and practical farming recommendations through a Streamlit web platform.

5.2 Specific Objectives

To collect and preprocess multi-source agricultural datasets — including weather, soil, and historical yield data — for use in maize yield prediction.

To build and evaluate machine learning models (such as Random Forest and XGBoost) that can accurately predict maize yield based on climatic and environmental variables.

To design and develop a Streamlit-based web application that allows users to input data, view predictions, and access yield insights interactively.

To integrate an advisory module that provides recommendations on planting periods, irrigation, and fertilizer application based on predicted yield and environmental conditions.

To test and deploy the system for public access, evaluating its usability, accuracy, and potential to support data-driven agricultural decision-making.

6. Literature Review

Several studies have explored the use of data science and machine learning techniques to improve agricultural productivity and crop yield forecasting. According to the Food and Agriculture Organization (FAO, 2023), accurate yield estimation is a critical factor in ensuring food security and sustainable agricultural planning. Traditional methods of yield estimation, which rely on manual surveys and field sampling, are often time-consuming, costly, and prone to human error. These limitations have driven researchers and institutions to adopt data-driven approaches that use weather, soil, and remote sensing data to model and predict crop yields more efficiently.

Recent studies have shown that machine learning algorithms such as **Random Forest**, **Support Vector Regression (SVR)**, and **XGBoost** can capture complex non-linear relationships between climatic variables and crop yields. For example, a study by Shastry et al. (2022) demonstrated that Random Forest models can achieve high accuracy in maize yield prediction when trained on multi-temporal satellite and weather data. Similarly, Ghosh et al. (2021) applied XGBoost to predict rice yields in India using rainfall, temperature, and soil features, achieving an R^2 of over 0.8. These

findings highlight the growing reliability of machine learning in agricultural forecasting, especially when diverse datasets are combined.

In Kenya, research on machine learning–based yield prediction is still emerging. Most existing systems focus on weather reporting or advisory messaging but lack predictive modeling and integration of spatial data. Projects by the Kenya Meteorological Department and the Ministry of Agriculture primarily emphasize climate monitoring rather than yield forecasting. Therefore, there remains a gap in systems that combine **historical yield, soil, and climatic data** to generate **localized and actionable insights** for farmers.

This project seeks to bridge that gap by developing a **Crop Yield Prediction and Advisory System (CYPAS)** that integrates machine learning, remote sensing, and user-friendly visualization. The proposed system will provide accurate maize yield predictions and accessible recommendations through an interface, contributing to Kenya's move toward data-driven agriculture and sustainable food production.

7. Methodology

7.1 Research Methodology

This study will adopt a **quantitative and exploratory research approach** aimed at identifying the relationship between environmental factors and maize yield in Kenya. The research will rely primarily on **secondary data sources**, including open agricultural datasets and public APIs, rather than field-based primary data collection.

Data will be obtained from credible and reliable sources such as the **Food and Agriculture Organization (FAO)** for historical crop yield data, **NASA POWER API** for temperature and rainfall data, and **SoilGrids/OpenLandMap** for soil composition data. The data will be cleaned, standardized, and merged to create a unified dataset for analysis.

The analysis will involve:

- **Exploratory Data Analysis (EDA)** to identify patterns, correlations, and outliers.

Feature engineering to derive meaningful predictors such as average rainfall, mean NDVI, and growing season temperature.

Model testing and validation using statistical performance metrics such as R^2 (Coefficient of Determination), RMSE (Root Mean Square Error), and MAE (Mean Absolute Error).

The study's results will then be interpreted and presented using visualizations such as graphs, charts, and heatmaps to support decision-making and system design.

7.2 Development Methodology

The system development process will follow the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, which is widely used in data science projects for its flexibility and structured workflow. The CRISP-DM approach includes six iterative stages:

Business Understanding – Defining the project goals, identifying the agricultural problem, and determining success criteria.

Data Understanding – Gathering data from multiple sources (FAO, NASA, SoilGrids) and assessing data quality.

Data Preparation – Cleaning, merging, and transforming data into a format suitable for machine learning modeling.

Modeling – Developing and training machine learning models (e.g., Random Forest, XGBoost) to predict maize yield.

Evaluation – Measuring the model's performance using evaluation metrics to ensure accuracy and reliability.

Deployment – Integrating the trained model into a **web application** to provide an interactive platform for yield prediction and advisory functions.

The system will be built using **Python** for model development, **Scikit-learn** and **XGBoost** for predictive modeling, and web interface. The system will also integrate **Plotly** and **Folium** for data visualization and mapping.

The entire methodology ensures that both the research and system development components are systematically executed, resulting in a reliable, functional, and user-friendly predictive system.

8.Budget and Resources

Resources Required

To successfully design, develop, and deploy the **Crop Yield Prediction and Advisory System (CYPAS)**, a combination of **hardware, software, data, and human resources** will be required. These resources ensure that all stages — from data collection to model deployment — are effectively executed.

Resource Type	Description	Estimated Cost (KES)
Hardware	Personal computer/laptop (8GB RAM, 500GB storage, Intel i5 or equivalent) for model training and testing	Already available
Additional Hardware	External hard drive or flash storage (for data backup and large dataset handling)	2,000
Software Tools	Python (Open-source), Streamlit (Free), Power BI (Free academic license), Jupyter Notebook	Free
Data Sources	FAO datasets, NASA POWER API, SoilGrids — all open access	Free
Internet & Utilities	Internet connectivity for data download, model testing, and deployment	3,000
Documentation & Printing	Printing of reports, proposal, and final documentation	1,500
Contingency	Miscellaneous costs during development and testing	1,500
Total Estimated Cost		**KES 8,000**

Total Estimated Budget: Minimal (project relies on open-source tools

Data Resources

The project will use open and publicly available datasets, ensuring cost-effectiveness and reproducibility:

FAOSTAT: Historical maize yield data by country and year.

NASA POWER API: Weather variables (temperature, rainfall, solar radiation).

SoilGrids/OpenLandMap: Soil characteristics including moisture, organic content, and pH.

Kenya Open Data Portal: County-level agricultural and demographic data.

Labour Cost Breakdown.

Task No.	Task Name	Duration	Estimated Hours	Rate (KES)	Subtotal (KES)
1	Proposal Writing	2 weeks	20 hours	600	12,000
2	SRS (Software Requirements Specification)	2 weeks	15 hours	600	9,000
3	SDS (Software Design Specification)	2 weeks	20 hours	600	12,000
4	Testing & Evaluation	2 weeks	25 hours	600	15,000
5	System Documentation	1 week	10 hours	600	6,000
6	User Manual	1 week	8 hours	600	4,800
7	Final Report Submission	1 week	12 hours	600	7,200
	TOTAL	**11 weeks**	**110 hours**		**KES 66,000**

Human Resources

Role	Responsibility
Project Developer (Mukami Haldline)	System design, data collection, model development, Streamlit app creation, documentation, and testing
Supervisor	Academic and technical guidance throughout the project lifecycle
Peer Reviewer/Colleague	System testing and feedback collection during evaluation stage

9.Project Schedule

Task No.	Description	Task No. of hrs	Planned Start Date	Planned Completion Date	Deliverables
1	Proposal Writing	80 hrs (≈2 weeks)	01 Oct 2025	20 Oct 2025	Completed project proposal
2	Software Requirement Specification (SRS)	80 hrs (≈2 weeks)	21 Oct 2025	30 Nov 2025	Approved SRS document
3	Software Design Specification (SDS)	80 hrs (≈2 weeks)	01 Dec 2025	15 Jan 2026	Completed SDS document
4	Testing and Evaluation	80 hrs (≈2 weeks)	16 Jan 2026	20 Feb 2026	Tested and validated system
5	System Documentation	40 hrs (≈1 week)	11 Mar 2026	20 Mar 2026	Final system documentation
6	User Manual	40 hrs (≈1 week)	21 Mar 2026	31 Mar 2026	User manual ready for deployment
7	Final Report Submission	40 hrs (≈1 week)	25 Mar 2026	31 Mar 2026	Final project report submitted

10.Conclusion

The proposed **Crop Yield Prediction and Advisory System (CYPAS)** seeks to address the persistent problem of unstable maize yields in Kenya by using data-driven methods. Through the integration of weather, soil, and historical yield data, the system will apply machine learning models to generate accurate predictions and practical recommendations for farmers and agricultural planners.

By presenting insights through a user-friendly **Streamlit web interface**, CYPAS will make complex data accessible to non-technical users and support timely decision-making. The system will contribute to improved planning, resource allocation, and food security while demonstrating how modern data science can be applied to solve real-world agricultural challenges in Kenya.

Upon completion, this project will serve as both a functional predictive tool and a foundation for future research in smart agriculture and AI-driven decision support systems.

11. References

Food and Agriculture Organization of the United Nations. (2023). *FAOSTAT: Crops and livestock products*. Food and Agriculture Organization. https://www.fao.org/faostat/en/#data/QCL

Ghosh, S., Saha, A., & Mandal, R. (2021). *Predicting crop yield using XGBoost algorithm: A data-driven approach*. *International Journal of Computer Applications, 183*(45), 1–6. https://doi.org/10.5120/ijca2021921442

ISRIC – World Soil Information. (2024). *SoilGrids: Global gridded soil information system*. https://soilgrids.org/

Kenya Meteorological Department. (2023). *Climate data and annual reports*. Government of Kenya. https://meteo.go.ke/

NASA POWER Project. (2024). *NASA Prediction of Worldwide Energy Resources (POWER) Data Access Viewer*. National Aeronautics and Space Administration. https://power.larc.nasa.gov/

Shastry, V., Kumar, M., & Patel, R. (2022). *Crop yield prediction using machine learning and remote sensing data*. *Journal of Agricultural Informatics, 13*(2), 55–67. https://doi.org/10.17700/jai.2022.13.2.722

World Bank. (2023). *Climate-smart agriculture in Africa: Enhancing productivity and resilience*. The World Bank Group. https://www.worldbank.org/