Machine Learning Model for Maize Yield Prediction in Marondera, Zimbabwe

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

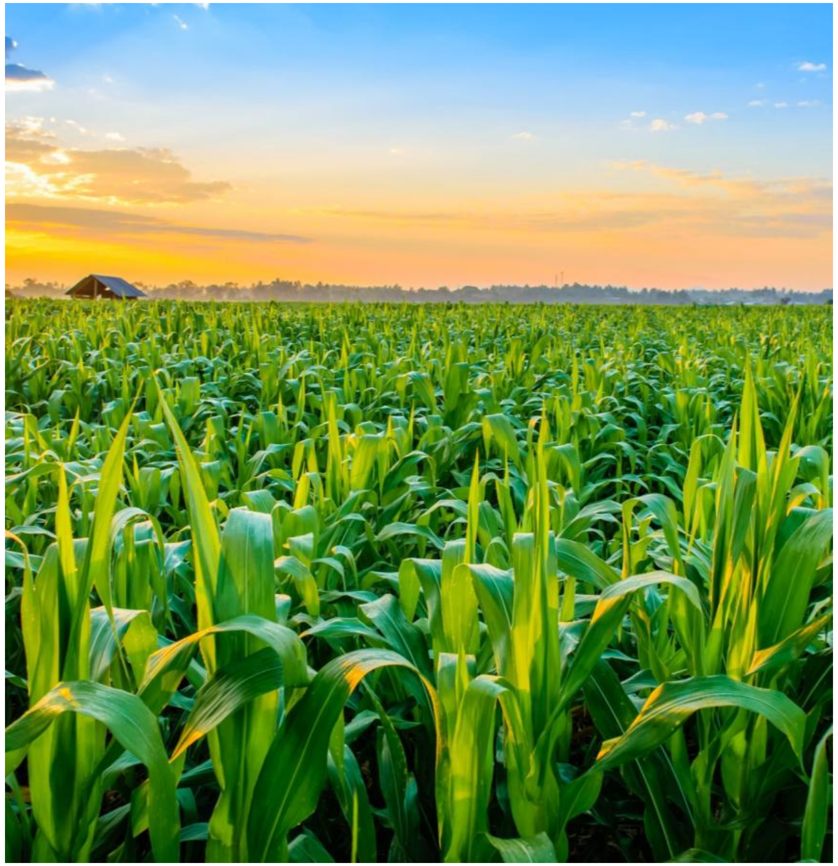
SOFTWARE ENGINEERING (HONS) BSEH BY THE ZIMBABWE OPEN UNIVERSITY

PREPARED BY: JAMES VASHIRI P1863122E

SUPERVISOR: MR MARUFU

1.0 Introduction

Agriculture is the backbone of Zimbabwe's economy, with maize being the country's staple crop. Smallholder farmers contribute significantly to maize production, but their yields are often affected by unpredictable weather patterns, poor soil fertility, and lack of access to timely agronomic advice. In the digital age, machine learning (ML) offers promising solutions to enhance agricultural productivity through predictive modeling and data-driven decision-making. This project focuses on developing a machine learning-powered maize yield prediction system for Marondera District, Zimbabwe, incorporating historical weather data and soil nutrient profiles to guide farmers toward optimal crop management.

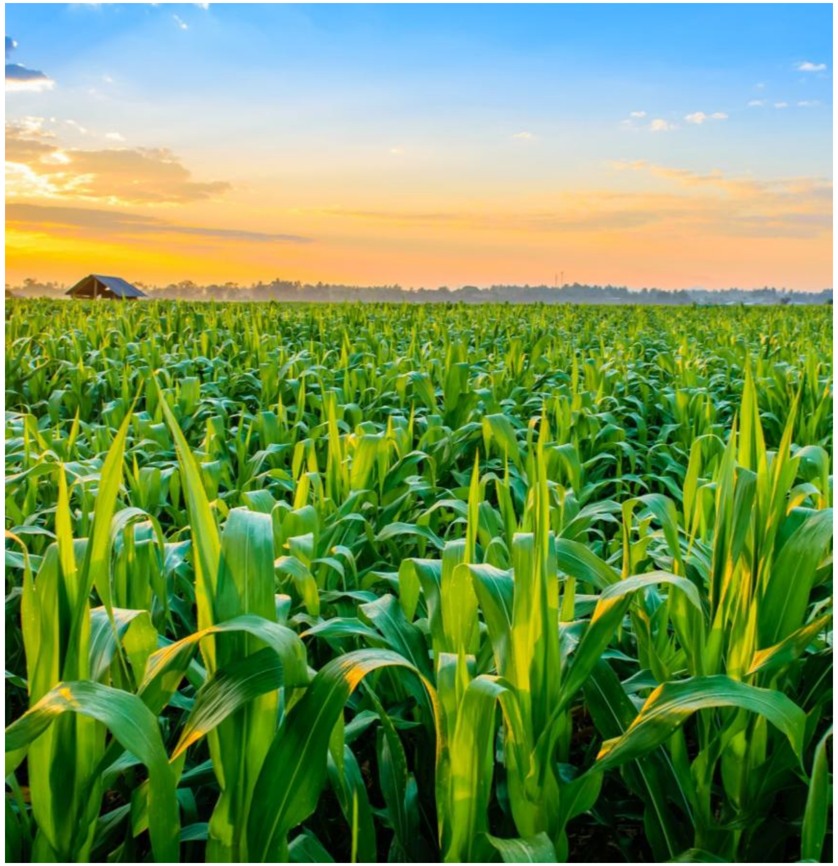


1.2 Background

Marondera District, located in Mashonaland East Province, has fertile soils and favorable climate conditions for maize farming. However, yield fluctuations are common due to inconsistent rainfall, climate change effects, and sub-optimal fertilizer use. Previous studies highlight the potential of ML models in crop yield forecasting by leveraging environmental data. Globally, countries like the USA use AI-driven farming systems extensively, but in Zimbabwe, adoption is still in its infancy. Addressing these gaps with a mobile-based solution tailored to Zimbabwean conditions could significantly improve yield predictability and input use efficiency.

1.3 Problem Statement

Despite Marondera’s strong agricultural potential, maize farmers continue to struggle with yield unpredictability, largely due to reliance on traditional farming practices and the absence of localized, data-driven agronomic support. The lack of accessible tools that integrate historical weather patterns and soil nutrient data limits farmers’ ability to make informed decisions. This results in suboptimal fertilizer application, inefficient irrigation, and frequent crop failures, undermining food security and economic stability.



1.4 Project Aim

The aim of this project is to develop a machine learning model that predicts maize yields in Marondera District.

1.5 Research Objectives

1. To clean agricultural and weather datasets into a unified, analysis-ready format.
2. To train maize yield prediction model using Random Forest regression.
3. To evaluate model performance using RMSE, MAE, and R^2 metrics.
4. To integrate Weather API into the maize yield prediction system.
5. To develop a Web app enabling farmers to access maize yield predictions.

Objective	Requirement(s)	Design Element(s)	Test Case(s)
1. To clean agricultural and weather datasets into a unified, analysis-ready format.	- Access to raw agricultural and weather CSV datasets. - Ability to handle missing data and normalize numerical features. - Merged dataset with consistent structure.	- Data preprocessing script (Python/Pandas). - Missing value imputation. - Feature normalization/scaling. - Final dataset saved as Final_Data.csv.	**TC1.1:** Verify that all datasets are loaded without error. **TC1.2:** Check that merged dataset contains no missing values. **TC1.3:** Confirm that numerical features are normalized between 0 and 1. **TC1.4:** Validate structure (column names, data types) of Final_Data.csv.
2. To train maize yield prediction model using Random Forest regression.	- Final_Data.csv as input. - Python with Scikit-Learn Random Forest Regressor. - Model trained and saved.	- Model training module (train_rf_model.py). - Random Forest Regressor model. - Model serialization using joblib (RF_Model.joblib).	**TC2.1:** Verify that the model trains without error. **TC2.2:** Check that RF_Model.joblib is created. **TC2.3:** Validate that model file is loadable and usable for predictions.

3. To evaluate model performance using RMSE, MAE, and R^2 metrics.	- Evaluation metrics: RMSE, MAE, R^2.
	- Test dataset split from training data.
	- Results output to console
		- Evaluation function in training script.
		- Sklearn metrics:
		    - mean_squared_error,
		    - mean_absolute_error,
		    - r2_score.
		- Print and/or save metrics results.
			TC3.1: Confirm that RMSE, MAE, R^2 are calculated and printed.
			TC3.2: Validate values are within acceptable range (as per benchmark).
			TC3.3: Check that poor performing models are flagged (optional).

4. To integrate Weather API into the maize yield prediction system.	- Access to Weather API (OpenWeatherMap). - Hardcoded location (Marondera district). - API key and internet access. - API response parsing.	- predict.py script or module. - API integration for fetching weather data. - Pre-processing API output for model prediction. - Output advisory messages.	**TC4.1:** Verify API call returns current weather data for Marondera. **TC4.2:** Check correct parsing and extraction of temperature, humidity, rain. **TC4.3:** Validate that prediction is made using API data..

1.6 Research Questions

1. How can agricultural and weather datasets be effectively cleaned and unified into a single, analysis-ready format for maize yield prediction?
2. How well does a Random Forest regression model predict maize yield using the cleaned agricultural and weather datasets?
3. How do RMSE, MAE, and R^2 metrics compare in assessing the accuracy and reliability of the maize yield prediction model?
4. How can real-time weather data from a Weather API be integrated effectively into the maize yield prediction system to improve prediction accuracy?
5. How can a web application be designed and developed to provide farmers with accessible, user-friendly maize yield predictions based on the trained model

1.7 Research Hypothesis

H0: A machine learning model cannot accurately predict maize yield based on environmental variables.

H1: A machine learning model can accurately predict maize yield based on environmental variables.

1.8 Significance of the Study

This study addresses critical gaps in agricultural technology adoption in Zimbabwe by providing a localized, data-driven solution for yield forecasting.

The outcomes can empower smallholder farmers to make evidence-based decisions, improve maize productivity, and optimize resource use (fertilizer, water).

Policymakers, agronomists, and stakeholders can also leverage the system for planning and extension services.

1.9 Scope

The project focuses on Marondera District and targets maize cultivation only.

It uses fixed soil nutrient profiles and historical weather data (2003–2022).

The web app is designed for all devices with English-language support.

1.10 Assumptions of the Research

- Farmers provide accurate field and planting details.
- Soil data remains relatively stable over the years.
- Weather data used (2003–2022) is reliable and representative of long-term trends.

1.11 Limitations

Limited to maize crop prediction (no multi-crop analysis).

Fixed soil data may not reflect micro-variations within the district.

The accuracy of yield prediction depends on data quality and completeness.

App adoption may be constrained by smartphone penetration and digital literacy.

1.12 Definition of Terms

Yield Prediction: Estimating future crop output based on input data.

Machine Learning: A subset of AI where models learn patterns from data to make predictions.

Soil Nutrient Profiles: Data on essential nutrients (for example, N, P, K) in soil.

Fertilizer Recommendation: Guidance on type and quantity of fertilizer based on soil and crop needs.

Irrigation Requirement: Additional water needed to meet crop water demands beyond rainfall.