Capstone project Proposal:Machine Learning-Based Wheat Yield Prediction (የስንዴ ምርት ትንበያ)Using Environmental Factors

GROUP 15

Name	Github Username
1. Lielina Akele	@Lielina19
2. Martha Egigu	@marthaoo
3. Melat Kebede	@MelatKebedeAbraham
4. Fikremariam Yalew	@Fikiremariyam
5. Sintayehu Mandefro	@Sintayehu-M

Abstract

This research focuses on developing a machine learning-based system to predict wheat yield using environmental factors such as temperature, precipitation and cloud cover. By analyzing historical meteorological data, the system will provide accurate yield forecasts to assist all actors involved in the agricultural sector including farmers in planning and decision-making. The project addresses the challenges of climate variability and the lack of accessible tools for yield prediction, which are critical for optimizing agricultural productivity. The integration of machine learning models will enable precise yield estimations, helping important stakeholders and farmers improve resource management and adapt to changing environmental conditions. This project aims to empower the agricultural sector and farmers with data-driven insights, ultimately enhancing wheat production.

Table of Contents

Introduction

Ethiopia's economy relies heavily on agriculture, contributing over 30 percent to the GDP and employing more than 70 percent of the labor force. This crucial sector faces significant challenges due to its dependence on unpredictable environmental variables such as rainfall, temperature, and humidity. These uncertainties not only affect productivity but also create difficulties in resource allocation, market planning, and policy formulation.

Therefore, predicting crop yields based on multiple environmental inputs is essential to addressing these challenges. A reliable forecasting solution enables government planners to develop food security strategies while helping farmers make informed decisions about harvesting schedules, labor needs, and resource utilization. Accurate crop yield predictions also assist policymakers in making better import and export decisions and allow seed companies to evaluate the performance of new crop varieties in different environments.

The motivation behind this project is the need to modernize Ethiopian agriculture through data-driven insights, especially as the demand for efficient farming practices continues to grow. By integrating machine learning into yield prediction, this project supports **Sustainable Development Goal (SDG) 2: Zero Hunger**, which aims to ensure food security, improve nutrition, and promote sustainable agriculture. Traditional agricultural methods and local knowledge remain important but are often insufficient in adapting to climate variability. Machine learning provides a scalable, data-driven approach that enhances agricultural decision-making. By analyzing historical and real-time environmental data, machine learning models can predict crop yields with high accuracy, helping farmers and government agencies optimize resource management, prepare for climate uncertainties, and improve national food security.

Real-World Applications and Practical Benefits

This yield forecasting project has numerous practical applications and real-world value, including:

1. **Planning for Farmers:** Yield forecasts can be used by farmers to plan the use of resources like the manpower and labor that will be needed during harvesting.
2. **Planning for the Government**: Yield forecasts enable the government to plan effectively for food security and market regulation, ensuring a stable agricultural economy.
3. **Market Insights:** By predicting crop yields, supply chain stakeholders can better plan for market demand, pricing, and storage requirements.
4. **Policy Formulation:** Accurate yield forecasts become a fundamental tool for government officials to create agricultural policies and intervention programs.

This project utilizes the capability of machine learning to study and predict agricultural yields depending on various environmental factors and thus address some of the critical challenges in Ethiopian agriculture. It holds the promise of improving productivity, enhancing food security, and modernizing farming practices.

Problem Statement

Agriculture is the backbone of Ethiopia's economy and sustains the livelihood of millions of individuals. Nonetheless, the country struggles to make reliable predictions of crop yields, hence making agricultural planning, resource distribution, and food security management complex. Existing methods of crop yield prediction often rely on previous trends and basic assumptions about weather patterns, without considering complex environmental factors and climatic shifts.

The main issue is the inability to accurately predict crop yields in real-time, especially under changing weather conditions unreliable predictions inhibit farmers and government agencies from making informed decisions regarding crop production, resource allocation, and disaster response. Not only this but also without accurate predictions, farmers and government officials struggle to plan for potential food shortages, droughts, or surplus yields, thereby leading to inefficiencies and economic losses.

There are some gaps that this project aims to fill but there are also some existing limitations and challenges. For instance, lack of accessibility and quality of data. Ethiopian agricultural data are scattered and not real-time which in turn will limit the accuracy of current forecasting methods. The dataset from the Ethiopian Agriculture Research Institute contain no meteorological data. Even the dataset they have is very small which makes it inadequate for a machine learning project.

Another gap is that traditional methods of yield prediction used by farmers are based on basic past trends or generalized methods, which fail to incorporate a wide range of environmental variables like temperature, precipitation and humidity. Despite the fact that agriculture is a backbone sector in Ethiopia’s economy, there has been limited research and development of AI-based solutions for crop yield prediction in Ethiopia. The lack of local context in AI application for agriculture can be used as an explanation for inefficient agricultural policies.

Objectives

General Objective:

The General objective of this research project is to develop a machine learning model that will accurately predict agricultural yields using different environmental parameters, such as temperature, rainfall, humidity, and precipitation.

Specific objectives:

1. Training a predictive model that uses historical and real-time data to predict crop yields with high accuracy.
2. Provide practical advice to farmers, policymakers, and agricultural stakeholders on how to optimize resource and labor utilization and marketing strategies.
3. Develop a model that considers Ethiopia's particular climatic trends and environmental variability so that it can be applied in any rainfed agriculture.
4. Develop a solution specifically designed for the Ethiopian agricultural environment, filling gaps in the use of machine learning for yield prediction in the country.

Methodology

To make the project viable and impactful, the scope is defined with the following boundaries and areas of focus:

- ✔ ***Focus on Rainfed Agriculture:*** The project is about rainfed farming systems, which are the prevailing agriculture sector in Ethiopia. Rainfed agriculture is a type of farming that relies on rainfall for water. As key features of the data it could contain humidity, precipitation intensity, precipitation Probability, cloud cover and dew point, which are crucial in the context of rain-fed agriculture.
- ✔ ***Target Crops:*** The project is dealing mainly with wheat which is critical for food security in Ethiopia.
- ✔ ***Data Utilization:*** Historical weather data, crop yield data, and other relevant datasets will be used in the model. Data cleaning and preprocessing are included in the methodology.

Machine Learning Techniques: Advanced machine learning techniques, deep neural networks (DNN) or random forest regression will be used by the project. Our methodology involves collecting and analyzing secondary data, additionally inputs from local farmers also might be used.

Given the anticipated impact of proper yield prediction, our primary variables are environmental data like temperature, rainfall. Although other factors like fertilizer usage and herbicide application can affect yield, this project specifically focuses on environmental variables as the key predictor.

Phase 1. Data Collection

We will use secondary data sources, primarily existing datasets having features like:

- ✔ Temperature: Daily maximum and minimum apparent temperatures, along with dew point temperatures, to account for climate variability.
- ✔ Humidity and Precipitation: Metrics capturing atmospheric moisture and precipitation intensity/probability, which are crucial for understanding water availability and stress levels.
- ✔ Cloud Cover and Visibility: Information on cloud cover and visibility conditions, which can indirectly impact photosynthesis and temperature regulation.
- ✔ Wind Speed and Direction: Data on wind conditions, which affect evapotranspiration rates and possibly pollination.
- ✔ Soil and Vegetative Health: Indices such as NDVI (Normalized Difference Vegetation Index) provide insights into crop health and growth stages.
- ✔ Day in Season: The day within the growing season, which helps in tracking growth phases and yield potential.
- ✔ Yield: The primary target variable representing wheat yield, which can be used for model training in predictive analytic.

The dataset will be split into training, testing, and validation sets to ensure suitability for machine learning model development. We will assess the model's performance based on these splits to ensure generalizability and accuracy.

Phase 2. Machine Learning Model Development

The machine learning phase involves selecting suitable models for yield prediction. We will experiment with several machine learning models to identify the one best suited to accurately predict yield. We will evaluate each model's performance in capturing both the overall trends and complex patterns in the data, selecting the one that provides a reliable prediction.

Then after choosing the suitable one we will train the model and tune it to improve predictions and after that we will test and check validation for accuracy.

Phase 3. Product Development

After the completion of model training, we will develop a web application to provide government officials especially the officials working in the agricultural sector, policy makers and farmers with real-time output or yield predictions. The application will include:

- ✔ **User Interface Design:** A user-friendly interface where agricultural experts working in the higher offices of the ministry of agriculture to experts working on Kebele level and farmers can view what the output yield will be.
- ✔ **Model Integration:** integration of the machine learning model to provide dynamic, data-driven recommendations.

Phase 4. Testing

✔ After training and integrating the model, the application will go through a testing and debugging phase to ensure usability and reliability.

Phase 5. Deliverable and Documentation

Finally, we will prepare the documentation for our project:

✔ ***Web Application:*** A platform for farmers or agricultural experts to view recommended wheat planting dates.
✔ ***Documentation:*** Detailed project documentation covering data sources, model development, software tools, and usage instructions.
✔ ***Presentation:*** A comprehensive presentation of the methodology, findings, and application functionality.

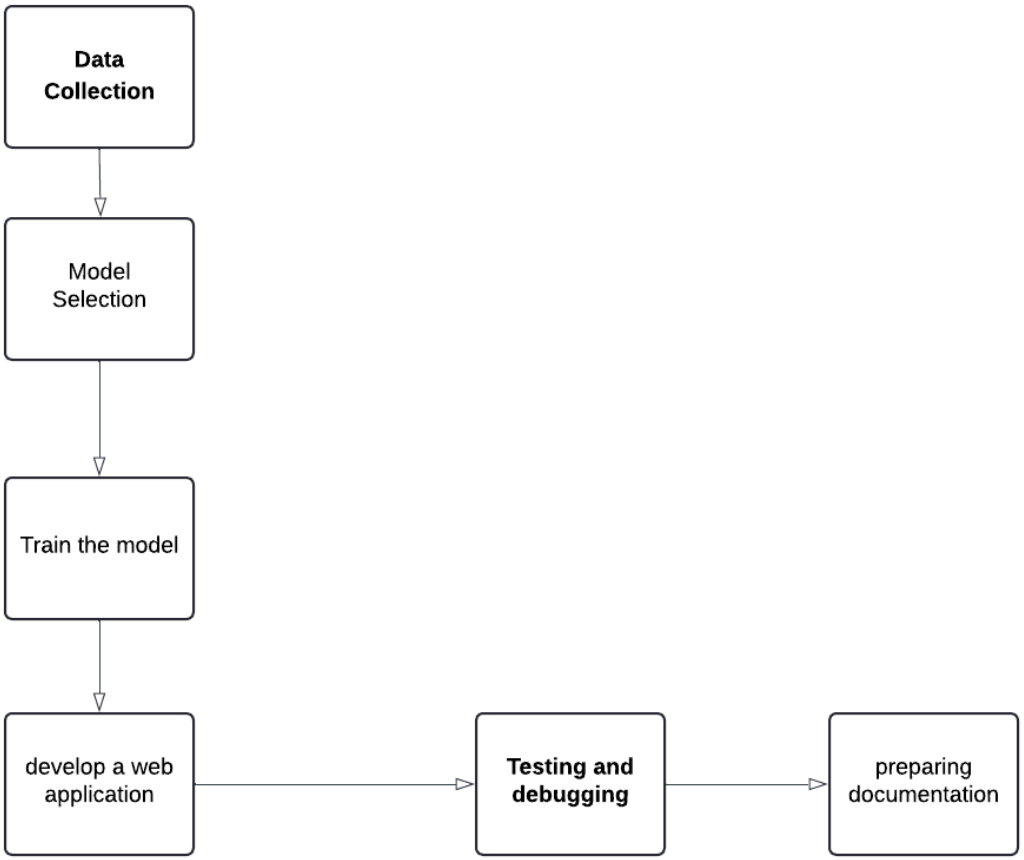

Figure 1- A flow chart for Methodology

Required Resources and Estimated Budget

Data Collection and API Access

- **API for Weather Data**: Basic API plans for weather data are often free but may require higher tiers for frequent requests or advanced data features.
    - **Cost Estimate:** The estimated cost for API access is 6,000 Birr, but this can be reduced by opting for a simpler solution that does not require advanced data features, potentially bringing the cost down to 0 Birr.
- **Data collection** : Datasets available online
    - **Cost Estimate**: 0 Birr.

Technology Infrastructure

- **Computing Hardware**: Existing computers are sufficient for data preprocessing and model training. Cloud services. Google colab have free plan for small machine learning projects
    - **Cost Estimate**: 0 Birr

Machine Learning Development Tools

- **Programming Language and Libraries:**
    - **Python**: Free and open-source.
    - **Jupyter Notebook**: Free for local use; also available on free cloud platforms like Google Colab.
    - **Libraries**: Free open-source libraries like Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn, TensorFlow, and PyTorch.
    - **Cost Estimate**: 0 Birr (all tools are free).

Human Resources

- **Project Team**:
    - Team members will contribute time and effort . Therefore there is no additional cost for labor.
    - **Cost Estimate**: 0 Birr (time invested by team members).

Total cost of the project = 6,000 birr.

Further Studies for Broader Scope

Our project specifically focuses on wheat to maintain a manageable research scope. After completing analysis and testing, we plan to extend this approach to other crops, including potatoes, beans and more, which are vital to agriculture in our country. Additionally, while this project emphasizes on enviromental factors such as temperature, rainfall etc to make the prediction, we recognize the importance of various other parameters influencing yield. Factors such as farming practice, fertilizer usage are critical in achieving higher yield, and these will be considered in future expansions of this project.

Timelines

Detailed timeline

Phase	Tasks	Timeline
Phase1: **Data Processing**	up to April 1	
	Collect environmental and agricultural data from different sources.	March 19-March 23
	Preprocessing, cleaning, and vi datasets.	To Be Determined (TBD))
	Split data into training, testing, and validation sets.	To Be Determined (TBD))
Phase 2: **ML Model Development**		
Model Selection	Check different model and select the appropriate oneTo Be Determined (TBD)	To Be Determined (TBD))
Initial Model Training	Begin training selected models and review initial results.	To Be Determined (TBD)
Hyperparameter Tuning	Fine-tune model parameters to improve accuracy.	To Be Determined (TBD)
Model Validation & Testing	Validate model on testing set and analyze accuracy.	To Be Determined (TBD)
Phase 3: **Product Development**		
UI Design	Design a user-friendly interface for farmers and experts.	To Be Determined (TBD)
Backend & Model Integration	Integrate model into backend for real-time recommendations.	To Be Determined (TBD)

Front-end Completion	Finalize the responsive front-end with interactive features.	To Be Determined (TBD)
Phase 4: **Testing**	Test entire application, debug, and ensure usability.	To Be Determined (TBD)
Phase 5: Deliverables & Documentation		
Documentation	Compile final project documentation and usage instructions.	To Be Determined (TBD)
Presentation Preparation	Prepare and rehearse project presentation.	To Be Determined (TBD)
Final Launch	Finalize the web app, check all functionality and documentation.	To Be Determined (TBD)