**Capstone Project Assignment 3**

**Data-Driven Crop Yield Prediction: Integrating Environmental Factors**

**Group 15**

**Group Members:**

1. **Lielina Akele…………………….@Lielina19**
2. Martha Egigu—------------- @marthaoo
3. Melat Kebede—------------ @MelatKebedeAbraham
4. Fikremariam Yalew—----- @Fikiremariyam
5. Sintayehu Mandefro—--- @Sintayehu-M

**Date: April 18, 2025**

**Introduction**

Ethiopia’s economy relies heavily on agriculture, it contributes over 30 percent to the GDP and over 70 percent of the labor force. This important sector generally faces challenges because of its dependence on unpredictable environmental variables such as rainfall, temperature, and humidity. These uncertainties not only affect productivity, but also pose a lot of challenges in resource allocation, market planning, and policy formulation.

Therefore, the prediction of crop yield on the basis of multiple environmental inputs is highly significant in addressing these challenges. A forecasting solution enables government planners to develop food security strategies, while allowing farmers to make necessary decisions concerning collecting season/time, labor force needed and resource utilization in general. This will reduce loss and maximize profits. Import and export decisions can be made easily by policymakers when the results of crop yield prediction are accurate. Seed companies can also evaluate the performance of new seeds in different environments. The motivation behind this project is the need to modernize Ethiopian agriculture through data, especially as the demand for efficient farming practices keeps increasing.

Integrating machine learning into yield prediction could bring about a complete transformation in Ethiopia's agricultural sector. Traditional agricultural methods alongside locally derived knowledge remain important yet insufficient to adapt farming practices when climate conditions change. Machine learning provides a robust, scalable, and data-driven approach to address existing challenges. Models in machine learning use historical and real-time environmental data to predict crop yield with a very high degree of accuracy. These models consider several inputs in the form of temperature, rainfall, humidity, and precipitation, thus giving a holistic understanding of the factors that affect productivity(yield). The significance of using a machine learning based solution is that it offers reliable yield forecast to help farmers and government make better decisions, offer policymakers insights for agricultural development & disaster preparedness and improve national food security by optimizing resource planning and distribution.

## ****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.

# 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.

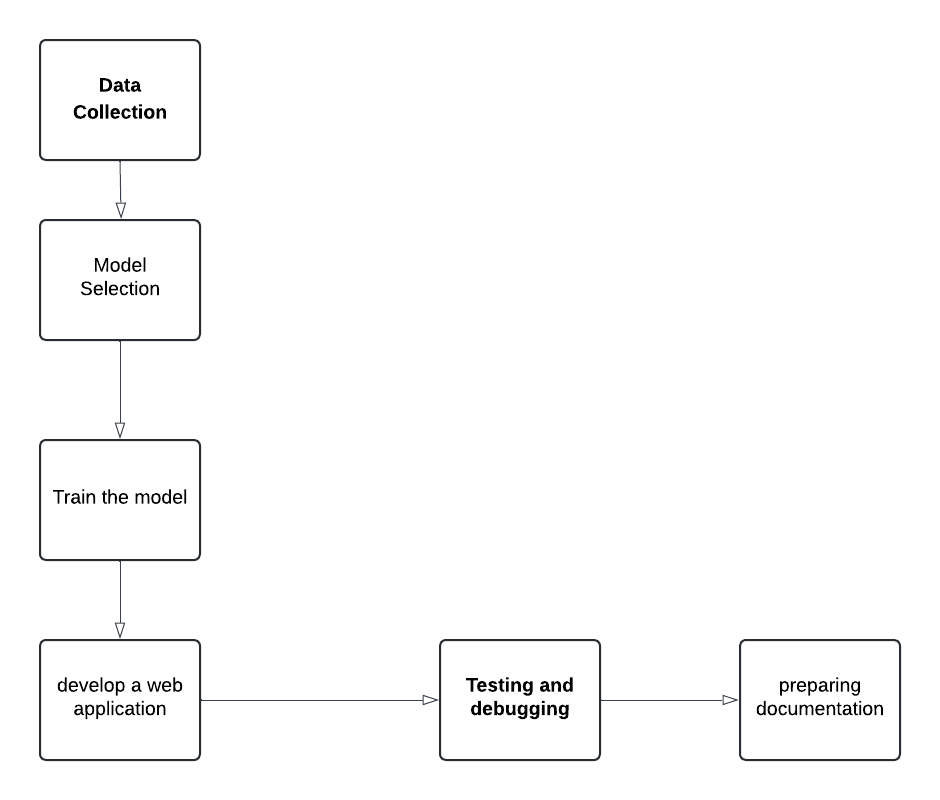
## Phase 2. Machine Learning Model Development

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. Testing

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

**Architecture Design Diagram**

# 

# **Data Sources**

# The data we plan to use for this project is from kaggle. (https://www.kaggle.com/datasets/shaikasif89/wheat-yeild/data). This dataset provides detailed information related to wheat production in India, intended for predictive modeling and analysis of wheat yield. This dataset captures various environmental, meteorological, and agronomic factors essential for accurately forecasting wheat yield.The dataset includes approximately 200,000 rows of data, with each row representing a unique data record related to wheat production. Since the data is a mock data the pre processing step might be a lot for this data set.

**Literature Review**

Recent research has demonstrated the effectiveness of machine learning, particularly neural network-based models, in predicting agricultural yield by analyzing complex environmental data. One study utilized three Self-Organizing Map models—Counter-Propagation Artificial Neural Network (CPANN), Supervised Kohonen Network (SKN), and XY-fusion Network (XYF)—to associate precision agriculture data with yield productivity classes. These models employed supervised learning techniques to process data collected using an on-line visible and near-infrared (vis–NIR) spectroscopy sensor, focusing on physicochemical soil parameters. The findings highlight the capability of machine learning to manage high-dimensional data for accurate yield estimation. Building upon this approach, our project extends the application by incorporating additional environmental features—such as temperature, rainfall, and NDVI—into classical models like Random Forest and Deep Neural Networks, tailored specifically to Ethiopian agricultural contexts for greater accessibility and impact.

**Technology Stack**

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

# Timelines

**Detailed timeline**

|  |  |  |
| --- | --- | --- |
| Milestones | **Tasks** | **Timeline** |
| **Data Processing** | Collect environmental and agricultural data from different sources. | April |
| Preprocessing, cleaning, and vi datasets. | April |
| Split data into training, testing, and validation sets. | April 25 |
| **ML Model Development** | Check different model and select the appropriate one | May, Week 1 |
| **Model Selection** | Begin training selected models and review initial results. | May, Week 2-3 |
| **Initial Model Training** | Fine-tune model parameters to improve accuracy. | May, Week 2-3 |
| **Hyperparameter Tuning** | Validate model on testing set and analyze accuracy. | May, Week 3-4 |
| **Testing** | Testing the Model | May, Week 3 |
| **Phase 5: Deliverables & Documentation** | Compile final project documentation and usage instructions. | May, Week 4 |
| **Documentation** | Prepare and rehearse project presentation. | May, Week 4 |
| **Presentation Preparation** | Finalize the web app, check all functionality and documentation. | May, Week 4 |

**Challenges and Mitigations**

One of the main anticipated challenges is **data quality**, as the dataset currently in use is a mock dataset that may not accurately reflect real-world environmental and yield data. This could result in weak correlations during data visualization and ineffective model training due to the randomness and lack of real patterns in the data.

Another challenge is **model performance**, which could be significantly impacted by the limitations of the mock dataset. Since machine learning models rely heavily on the quality and relevance of input data, we may experience poor accuracy and overfitting. To address this, we will conduct rigorous validation using different subsets and cross-validation techniques, and continually refine the model using real-world data as it becomes available.

Finally, **technical constraints** such as limited computational resources may hinder efficient training and testing of complex models. To overcome this, we will leverage cloud-based platforms such as Google Colab, which provides free access to GPUs, and optimize model architecture to reduce computational overhead.

****Ethical Considerations****

While our project does not directly handle personal or sensitive user data, ethical considerations still play a crucial role. **Data privacy** is respected by exclusively using publicly available datasets ensuring no violation of user rights. However, **bias** may arise if the model is trained on data that does not represent the full range of environmental conditions across different Ethiopian regions. This could lead to inaccurate predictions for certain areas.

**References**

1. **https://www.sciencedirect.com/science/article/abs/pii/S0168169915003671**