Tribhuvan University

Institute of Science and Technology

A Final Year Project Report

On

"WHEAT YIELD PREDICTION SYSTEM"

Submitted To:

Department of Computer Science and Information Technology,
AMBITION COLLEGE
Mid-Baneshwor, Kathmandu

In the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Information Technology

Submitted by:

Dip Kishor Regmi (TU Roll No.: 26268/077)
Surakshya Adhikari (TU Roll No.: 26296/077)
Prajwal Baral (TU Roll No.: 26280/077)

Under the supervision of

Mr. Yubraj Dahal

February, 2025

ACKNOWLEDGMENT

This report entitled website review of **“WHEAT YIELD PREDICTION SYSTEM”** has been the result of the research dedication and effort poured for the fulfillment of a certain portion of external evaluation for the Bachelors Computer Science and Information Technology degree. This output could not have been displayed without the involvement and support of few individuals. Hence, we are really thankful for their help and would like to acknowledge all of them with our sincere gratitude.

We would like to convey our thankful greeting toward Bsc.CSIT department and our project Supervisor **Mr.Yubraj Dahal** for guiding us with advice and support to prepare the report along with providing the responsibility of preparing the report. Similarly, we are thankful to our well-wisher for giving us necessary feedback, comments and ideas when required. We would like to thank everyone revolving around our project with pleasure.

Lastly, we would like to express our sincere thanks to all our friends and seniors who helped us directly and indirectly during this project.

Thanking you,

Dip Kishor Regmi (TU Roll No.: 26268/077)
Surakshya Adhikari (TU Roll No.: 26296/077)
Prajwal Baral (TU Roll No.: 26280/077)

ABSTRACT

The Wheat Yield Prediction System is an advanced analytical platform designed to optimize agricultural productivity through precise yield forecasting. Leveraging a custom built random forest regression model, the system predicts wheat yield by analyzing a comprehensive dataset of agro-meteorological and soil parameters, including Rainfall, AvgTemp, Relative Humidity, SoilTemp, Sand, PHLevel, Phosphorus, Potassium, Clay, and ProductionArea. The system has been meticulously crafted without the use of prebuilt machine learning libraries, demonstrating a ground-up approach to model implementation. By incorporating robust techniques for handling missing data and ensuring accurate parameter analysis, the system delivers reliable predictions tailored to various climatic and soil conditions. Its scalable design allows for integration with real-time data sources, empowering farmers and agricultural stakeholders with actionable insights to make informed decisions. Developed using Python and adhering to an agile methodology, the project underscores the potential of AI-driven solutions in addressing critical challenges in agriculture. By minimizing uncertainty and enhancing planning accuracy, the Wheat Yield Prediction System presents a cost-effective and innovative tool for advancing sustainable farming practices and boosting agricultural efficiency on a global scale.

Keywords: *Random Forest, Prediction, Sustainable, Machine Learning*

TABLE OF CONTENTS

LIST OF FIGURE

LIST OF TABLE

ABBREVIATION

API	:	Application programming interface
AWS	:	Amazon Web Services
CP-ANN	:	Counter-Propagation Artificial Neural Networks
DL	:	Deep learning
FYM	:	Farmyard manure
GP	:	Genomic prediction
IIS	:	Internet Information Services
ML	:	Machine learning
PM	:	Poultry manure
RCBD	:	Randomized Complete Block Design
SKN	:	Supervised Kohonen Networks
SS	:	Sewage sludge
SQL	:	Structured Query Language

CHAPTER 1:
INTRODUCTION

1.1 Introduction

Agriculture is the backbone of Nepal's economy, with wheat being one of the most significant cereal crops cultivated across the country. However, wheat production is highly dependent on several environmental and agronomic factors such as soil quality, climatic conditions, and fertilizer usage. The unpredictability of these factors often leads to challenges in forecasting wheat yield, making it difficult for farmers and policymakers to make informed decisions. Our project**,** Wheat Yield Prediction System, is designed to address this challenge by providing an accurate and data-driven approach to predict wheat production across Nepal’s 77 districts. By analyzing key factors that influence yield, the system aims to help farmers optimize their resources, reduce potential losses, and improve overall agricultural productivity. This system will not only benefit individual farmers but also assist governmental and agricultural organizations in planning and policymaking for better food security and sustainability.

The Wheat Yield Prediction System works by utilizing machine learning techniques to analyze historical data related to wheat production. The system collects and processes data such as temperature, rainfall, soil type, and fertilizer application to predict future wheat yields. The Random Forest algorithm**,** a powerful ensemble learning method, is used to train the model on historical agricultural datasets. By identifying patterns and relationships among various parameters, the system can provide accurate yield predictions for different regions. Users can input specific environmental and agricultural parameters into the system, and the model will generate an estimated wheat yield based on past trends and available data. This predictive capability allows farmers to make better decisions regarding crop management, while policymakers can use the insights to implement data-driven agricultural policies and improve food security strategies.

The implementation of this project involves a combination of frontend, backend, and machine learning model development. The frontend is built using React, providing an intuitive and user-friendly interface where users can input relevant data and view predictions in an interactive manner. The backend is developed using .NET, ensuring efficient data processing, model integration, and database management. The Random Forest algorithm is implemented without external machine learning libraries, meaning that all calculations, data processing, and model training are manually coded. The system is designed to be deployed as a web-based platform, allowing easy access to farmers, researchers, and government agencies across Nepal. By providing a reliable and accessible prediction system, this project aims to contribute to better agricultural decision-making, improved wheat yield forecasting, and overall enhancement of Nepal's wheat farming sector.

1.2 Problem Statement

Nepal's wheat production is essential for food security and the economy, but it faces many challenges like unpredictable weather, uneven rainfall, varying soil quality, and improper use of fertilizers. These issues cause inconsistent wheat yields across regions. Additionally, the lack of a reliable system to analyze and predict wheat production makes it harder for farmers and policymakers to make informed decisions. This leads to inefficient use of resources and missed opportunities to improve wheat farming. A system that provides insights and accurate predictions is needed to address these problems and support better decision-making to boost wheat production and ensure food security.

1.3 Objective

- To develop a system that predicts wheat production in Nepal based on various environmental and soil factors using random forest algorithm.

1.4 Scope and Limitation

- **Scope:**
    - The system will analyze data related to wheat production in Nepal.
    - It will focus on factors such as climate conditions, soil properties, and fertilizer usage.
    - The system will utilize the Random Forest algorithm to predict wheat production.
- **Limitations:**
    - The project will not use external libraries, requiring manual implementation of algorithms.

- Data availability and quality could affect the accuracy of predictions.

1.5 Development Methodology

The project will follow an Agile methodology, enabling iterative development and continuous feedback. Key stages include requirement gathering, design, implementation, testing, and deployment, ensuring that the system meets user needs and adapts to challenges. Agile methodologies do not suffer from the traditional problems of delayed deliverables from the start of the project stage since every project phase lasts only a few weeks. Each phase is a lot more productive than the last, thanks to all the team members pooling their efforts to accomplish the common goal of working out the required functionality of the software during that phase.

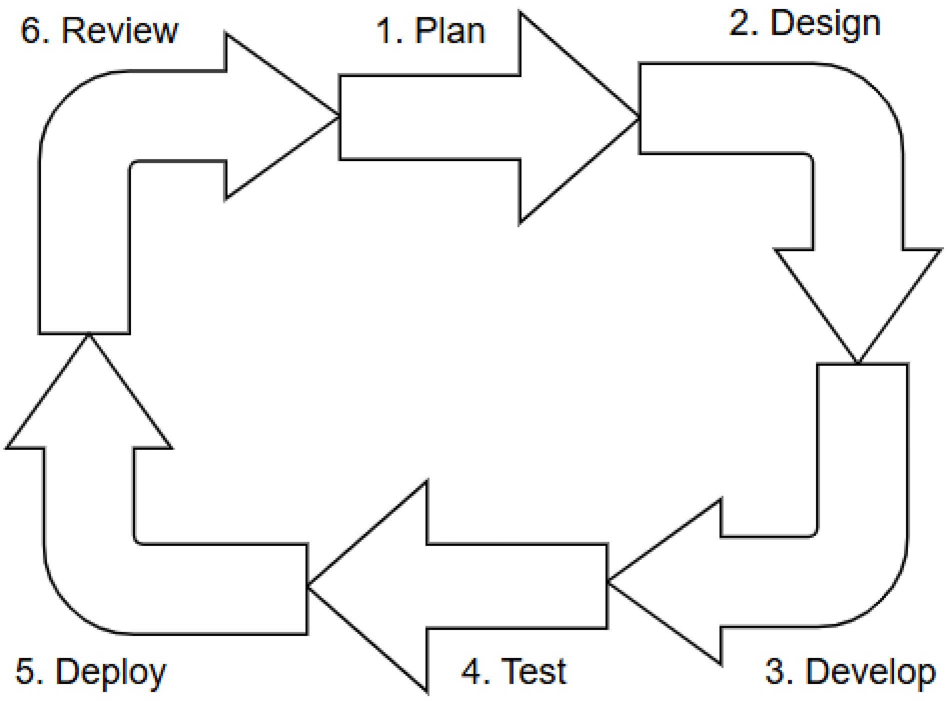


Figure 1.1: An Agile Approach for Development of Wheat Yield Prediction System

1.6 Report Organization

This document is organized as follows:

- **Chapter 1:** Introduction, including problem statement, objectives, scope, and methodology.

- **Chapter 2:** Background study and literature review, providing an overview of relevant theories, concepts, and existing studies.
- **Chapter 3:** System analysis, detailing requirements and feasibility analysis.
- **Chapter 4:** System design, structured approach and algorithm details.
- **Chapter 5:** Implementation and Testing, including the details of tools used, implementation details of module, test cases and result analysis.
- **Chapter 6:** Conclusion and future recommendations

CHAPTER 2:

BACKGROUNG STUDY AND LITERATURE REVIEW

2.1 Background Study:

Fundamental Theories and Concepts

The project "Wheat Yield Prediction System" revolves around understanding and analyzing the factors affecting wheat production using a data-driven approach. This section outlines the key theories, concepts, and terminologies fundamental to this project.

Agricultural Production Systems

Agricultural production systems are influenced by various environmental, biological, and management factors. Wheat, a staple crop in Nepal, is particularly sensitive to climate variations, soil properties, and fertilizer applications. Understanding these relationships is critical to predicting and optimizing yield.

Climate and Wheat Production

Climate factors such as temperature, rainfall, and humidity play a pivotal role in determining wheat yield. For instance, excessive rainfall can lead to waterlogging, affecting crop growth, while insufficient rainfall during critical growth phases may reduce yield. Nepal, with its diverse topography, presents unique challenges in balancing these climatic influences across its districts.

Soil Characteristics and Fertility

Soil type, pH levels, and nutrient content significantly affect wheat growth. The presence of organic matter and essential nutrients like nitrogen, phosphorus, and potassium determines the fertility of the soil, which directly correlates with production levels.

Fertilizer Usage and Crop Yield

Efficient fertilizer usage is essential for increasing productivity. Overuse or underuse of fertilizers can degrade soil health or reduce yield. This project examines how fertilizer application patterns vary across districts and their impact on production.

2.2 Literature Review

The field of wheat production analysis involves the study of numerous environmental and agricultural factors, each impacting crop yields in different ways. Over the past years, several studies have focused on analyzing wheat production using advanced technologies, integrating data related to climate, soil, and fertilizer

“Impact of climate change on district-level wheat production in Nepal,” This paper follows the Ricardian approach to estimate district-level fixed effect panel regressions on per-hectare net wheat revenues. The study integrates climate variables such as precipitation and temperature, alongside traditional agricultural inputs like fertilizer and labor. The results indicate that both temperature and precipitation negatively impact net wheat revenues, but in a diminishing way over time. [1]

“Effect of organic manures and chemical fertilizers on grain yield of maize in rainfed area,” This paper explain the experiment that was done to examine how different types and amounts of fertilizers and manures affect maize yield and nutrient uptake. The experiment used farmyard manure (FYM), poultry manure (PM), sewage sludge (SS), and chemical fertilizers, applied alone or in combinations, before planting the maize variety Agaiti 85. Conducted under a Randomized Complete Block Design (RCBD) with three repetitions, the results showed that poultry manure alone or with half of the recommended chemical fertilizer significantly improved maize yield compared to farmyard manure, sewage sludge, and NP fertilizers. [2]

“Climate Change and Wheat Production in Pakistan: An Autoregressive Distributed Lag Approach" Pakistan is considered particularly vulnerable to climate change due to its geographical location. Human activities have increased greenhouse gases like carbon dioxide, methane, and nitrous oxide, which trap sunlight and raise global temperatures. This higher temperature, especially in tropical areas, could negatively impact wheat growth and productivity. This study examines how climate change affects wheat production in Pakistan using the ARDL model and annual data from 1960 to 2009. [3]

“Wheat yield prediction using machine learning and advanced sensing techniques,” This study aimed to predict wheat yield variations within a 22-hectare field in Bedfordshire, UK, using high-resolution soil data and satellite imagery. On-line proximal soil sensing was used to estimate soil properties, significantly reducing labor and time in soil sampling and analysis. The research compared the performance of three neural network models: Counter-Propagation Artificial Neural Networks (CP-ANNs), XY-Fused Networks (XY-Fs), and Supervised Kohonen Networks (SKNs). Input nodes for these models were formed from normalized soil parameters and satellite-derived NDVI data, while output nodes predicted yield classes (low, medium, and high). [4]

“Genomic Prediction of Wheat Grain Yield Using Machine Learning,” This study evaluated the performance of machine learning (ML), deep learning (DL), and classical Bayesian methods for genomic prediction (GP) of wheat grain yield across three datasets, examining how these methods interact with different feature selection (FS) techniques. The results showed that prediction algorithms had a greater influence on model performance than the choice of FS methods. Tree-based ML methods, such as random forests and gradient boosting, along with classical Bayesian approaches, demonstrated the best performance. However, Bayesian methods encountered fitting problems, especially when using Bayes for feature selection. Other FS methods produced models without fitting issues but with similar performance. Overall, random forests and gradient boosting were found to be the most robust and highly predictive algorithms for wheat grain yield GP. [5]

CHAPTER 3:

SYSTEM ANALYSIS

3.1 System Analysis:

The system analysis focuses on understanding and defining the requirements and feasibility of the Wheat Production Analysis system. It ensures the project meets its objectives effectively and efficiently.

3.1.1 Requirement Analysis:

3.1.1.1 Functional Requirements:

The functional requirements define the core operations and behavior of the system. Below are the functional requirements for the Wheat Production Analysis system:

1. **User Authentication**:
    - Admin and users must log in to access the system
    - Admin can manage specific data for wheat production
2. **Data Input and Management**:
    - The system should allow the user to provide input data on climate, soil type, and fertilizer usage.
    - Data validation should ensure the accuracy and consistency of the entered values.
3. **Processing of Data and Result Generation**:
    - The system should predict wheat production according to provided input data.
4. **Result Display:**
    - Prediction results should be displayed in an easy-to-understand format, such as graphs or tables.

Use Case Diagram

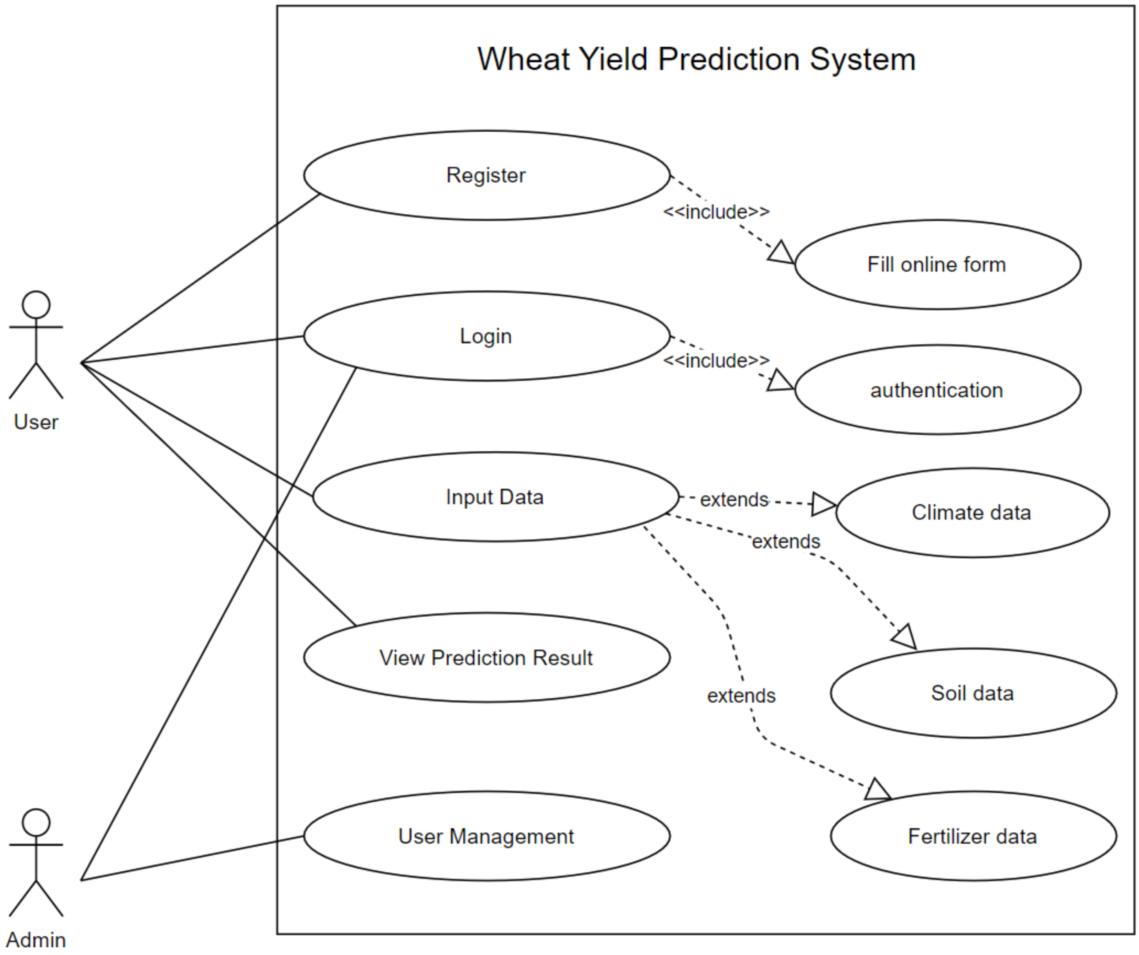


Figure 3.1: Use case diagram of Wheat Yield Prediction System

3.1.1.2 Non-Functional Requirements

Non-functional requirements describe the quality attributes of the system:

1. **Performance**:
   - The system should process predictions for wheat yield in less than 5 seconds.
2. **Scalability**:
   - The system should be scalable to handle future enhancements, such as adding more crops or districts.

3. **Usability**:

    - The system interface should be easy for admin and users to navigate.

4. **Security**:

    - Ensure secure login for the admin.
    - Encrypt sensitive data such as production data and predictions.

5. **Reliability**:

    - The system should function consistently without crashes or errors.

3.1.2 Feasibility Analysis:

3.1.2.1 Technical Feasibility:

This analysis evaluates whether the proposed system can be implemented using the available technologies and resources.

Frontend:

React is a robust and widely-used JavaScript library for building dynamic user interfaces. Its component-based architecture makes it well-suited for modern web applications. React integrates seamlessly with backend APIs, allowing efficient data fetching and UI updates. Additionally, the availability of a large number of libraries and tools enhances its development capabilities.

Backend:

The backend is built using .NET, a highly scalable and versatile framework. .NET supports multiple programming languages (such as C#) and is known for its high performance, security, and compatibility with various database systems. It efficiently handles server-side logic, APIs, and user authentication, making it a reliable choice for the backend.

The random forest model is built and trained using python.

Integration:

The React frontend communicates with the .NET backend via RESTful APIs, and .NET backend communicate with python model via flask api, ensuring a clean separation of concerns and ease of development. The interoperability between the two technologies ensures smooth data exchange and responsiveness.

Tools and Hosting:

Hosting platforms like IIS or AWS can support .NET applications efficiently, and React applications can be deployed on platforms like Vercel.

3.1.2.2 Operational Feasibility:

The system is designed to simplify and automate wheat production analysis. It is user-friendly, ensuring easy adoption by the end-users (e.g., agricultural experts or policymakers). Training materials, guides, and minimal technical assistance will be provided to ensure smooth operation.

3.1.2.3. Economic Feasibility:

Development Cost:

Using .NET for the backend and React for the frontend incurs minimal licensing costs as both have free development environments. Visual Studio (for .NET) and open-source React libraries make development cost-effective.

Hosting Costs:

Platforms like Azure or AWS may involve moderate hosting costs, but the benefits of scalability outweigh the investment.

Maintenance:

Regular maintenance and updates will require a small team of developers familiar with .NET and React.

3.1.2.4 Schedule Feasibility:

The project follows an Agile development methodology with well-defined sprints. Milestones such as requirement gathering, system design, and module-wise development ensure timely delivery. Both .NET and React frameworks support rapid development and iteration, making it feasible to complete the project within the given timeframe.

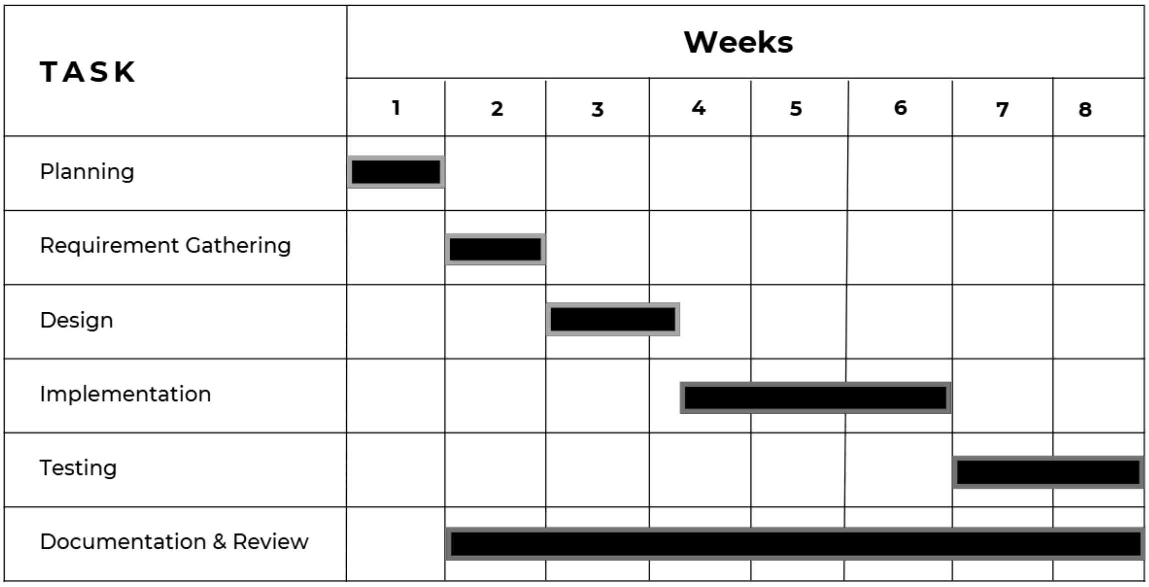
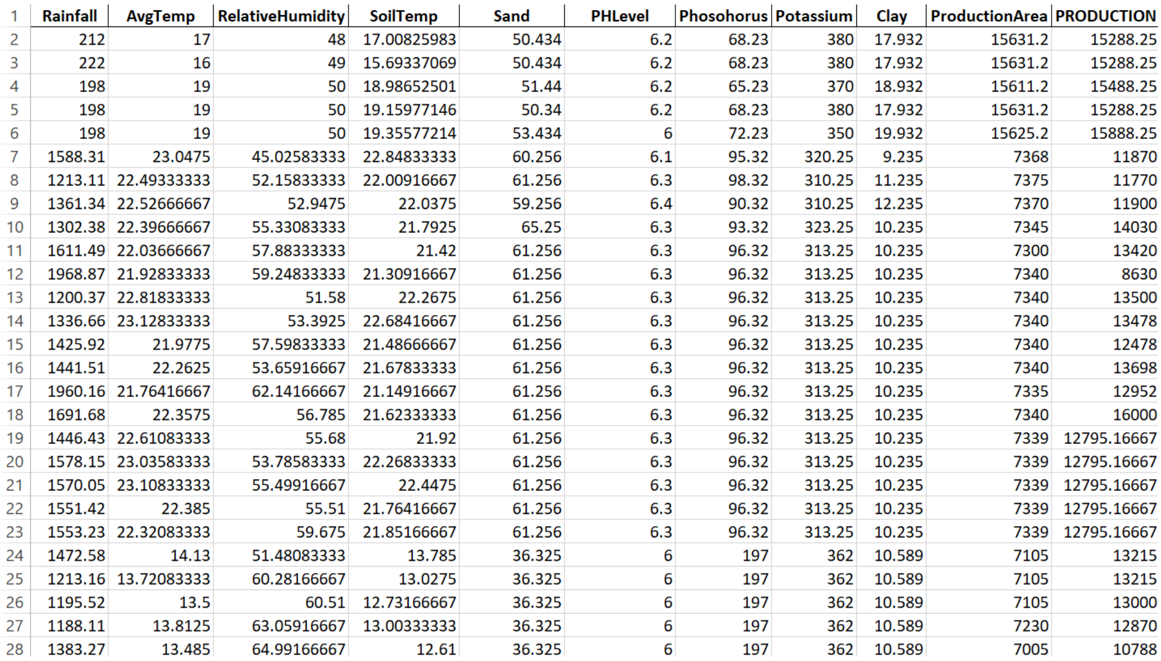


Figure 3.2: Gantt Chart

3.1.3 Analysis

3.1.3.1 Data Collection

Dataset in .csv format. Total 1296 data is then split into train and test (70% and 30% respectively) to build and test our system. All the data are collected from different sources and merged together according to the date. Climate data is pulled from Kaggle repository, soil data is collected from NAARC website and merged together. There are 10 different attributes of data.



1	Rainfall	AvgTemp	RelativeHumidity	SoilTemp	Sand	PHLevel	Phosohorus	Potassium	Clay	ProductionArea	PRODUCTION
2	212	17	48	17.00825983	50.434	6.2	68.23	380	17.932	15631.2	15288.25
3	222	16	49	15.69337069	50.434	6.2	68.23	380	17.932	15631.2	15288.25
4	198	19	50	18.98652501	51.44	6.2	65.23	370	18.932	15611.2	15488.25
5	198	19	50	19.15977146	50.34	6.2	68.23	380	17.932	15631.2	15288.25
6	198	19	50	19.35577214	53.434	6	72.23	350	19.932	15625.2	15888.25
7	1588.31	23.0475	45.02583333	22.84833333	60.256	6.1	95.32	320.25	9.235	7368	11870
8	1213.11	22.49333333	52.15833333	22.00916667	61.256	6.3	98.32	310.25	11.235	7375	11770
9	1361.34	22.52666667	52.9475	22.0375	59.256	6.4	90.32	310.25	12.235	7370	11900
10	1302.38	22.39666667	55.33083333	21.7925	65.25	6.3	93.32	323.25	10.235	7345	14030
11	1611.49	22.03666667	57.88333333	21.42	61.256	6.3	96.32	313.25	10.235	7300	13420
12	1968.87	21.92833333	59.24833333	21.30916667	61.256	6.3	96.32	313.25	10.235	7340	8630
13	1200.37	22.81833333	51.58	22.2675	61.256	6.3	96.32	313.25	10.235	7340	13500
14	1336.66	23.12833333	53.3925	22.68416667	61.256	6.3	96.32	313.25	10.235	7340	13478
15	1425.92	21.9775	57.59833333	21.48666667	61.256	6.3	96.32	313.25	10.235	7340	12478
16	1441.51	22.2625	53.65916667	21.67833333	61.256	6.3	96.32	313.25	10.235	7340	13698
17	1960.16	21.76416667	62.14166667	21.14916667	61.256	6.3	96.32	313.25	10.235	7335	12952
18	1691.68	22.3575	56.785	21.62333333	61.256	6.3	96.32	313.25	10.235	7340	16000
19	1446.43	22.61083333	55.68	21.92	61.256	6.3	96.32	313.25	10.235	7339	12795.16667
20	1578.15	23.03583333	53.78583333	22.26833333	61.256	6.3	96.32	313.25	10.235	7339	12795.16667
21	1570.05	23.10833333	55.49916667	22.4475	61.256	6.3	96.32	313.25	10.235	7339	12795.16667
22	1551.42	22.385	55.51	21.76416667	61.256	6.3	96.32	313.25	10.235	7339	12795.16667
23	1553.23	22.32083333	59.675	21.85166667	61.256	6.3	96.32	313.25	10.235	7339	12795.16667
24	1472.58	14.13	51.48083333	13.785	36.325	6	197	362	10.589	7105	13215
25	1213.16	13.72083333	60.28166667	13.0275	36.325	6	197	362	10.589	7105	13215
26	1195.52	13.5	60.51	12.73166667	36.325	6	197	362	10.589	7105	13000
27	1188.11	13.8125	63.05916667	13.00333333	36.325	6	197	362	10.589	7230	12870
28	1383.27	13.485	64.99166667	12.61	36.325	6	197	362	10.589	7005	10788

Figure 3.3 Snapshot of Data for Wheat Yield Prediction System

3.2.3.2 Data Modeling Using ER Diagram

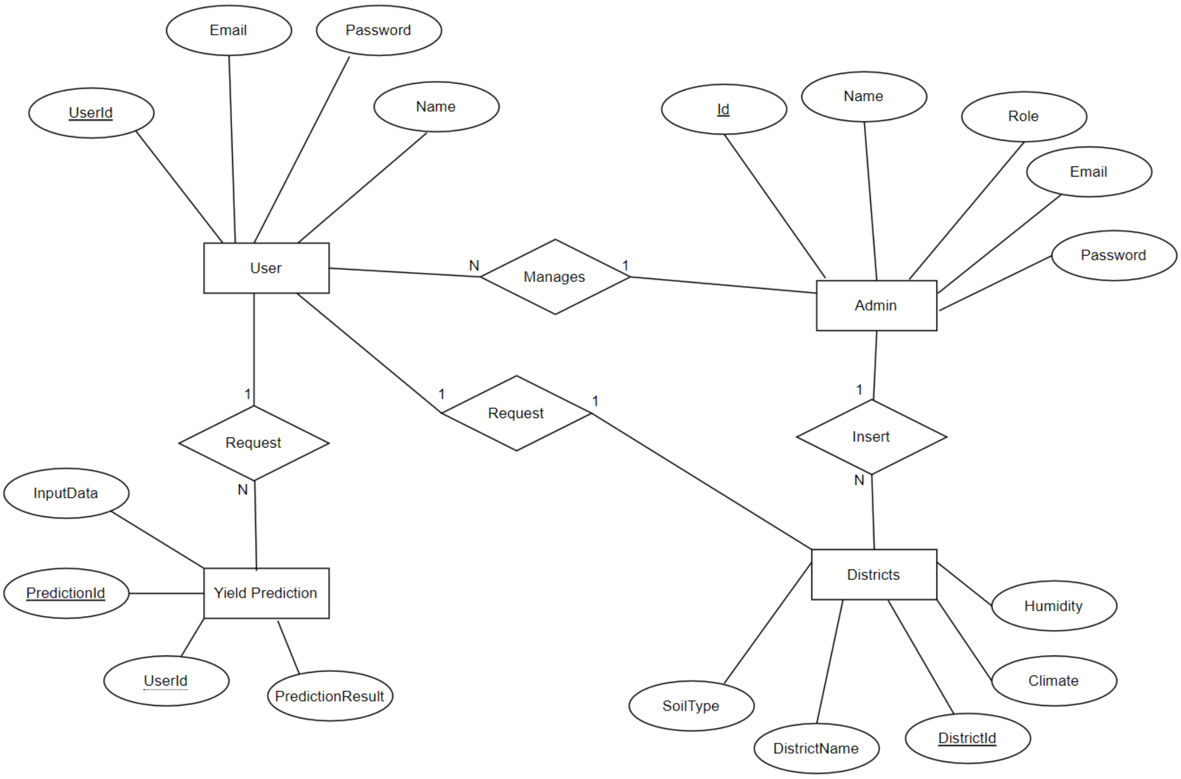


Figure 3.4: ER Diagram of Wheat Yield Prediction System

This ER diagram represents a system where administrators manage users and district data, and users interact with the system for agricultural yield prediction. The Admin entity includes details such as Id, Name, Role, Email, and Password, and has a "Manages" relationship with the User entity, allowing one admin to oversee multiple users. Users, identified by attributes like UserId, Email, Password, and Name, can make multiple yield prediction requests through the "Request" relationship. The Yield Prediction entity stores prediction details, including PredictionId, InputData, PredictionResult, and the associated UserId. Admins can also insert district information, such as DistrictId, DistrictName, SoilType, Humidity, and Climate, into the **Districts** entity via the "Insert" relationship. This system is designed to facilitate crop yield predictions while enabling administrators to manage users and update district-related data.

3.2.3.3 Process Modelling using DFD

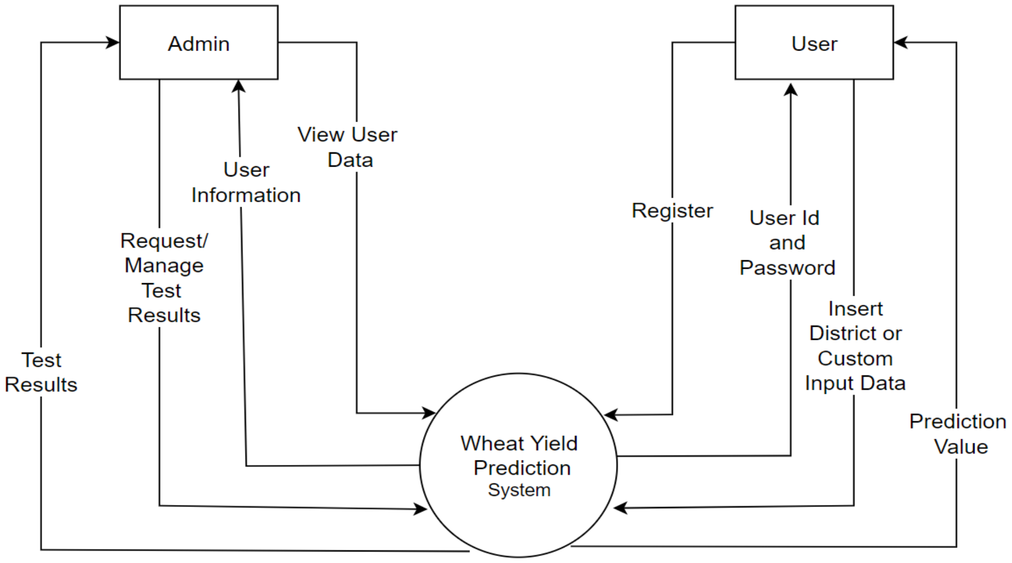


Figure 3.5: DFD Level 0 of Wheat Yield Prediction System

This DFD shows the interaction flow in a wheat yield prediction system involving two primary roles: Admin and User**.** The admin can view user data, manage test results, and request test data to oversee the system's operations. Users register by providing their User ID and Password, then input district-specific or custom data for yield prediction. The system processes the input and provides a prediction value back to the user. Test results generated by the system can be accessed by the admin for review and management. This system facilitates efficient collaboration between administrators and users, ensuring accurate predictions and streamlined management of the process.

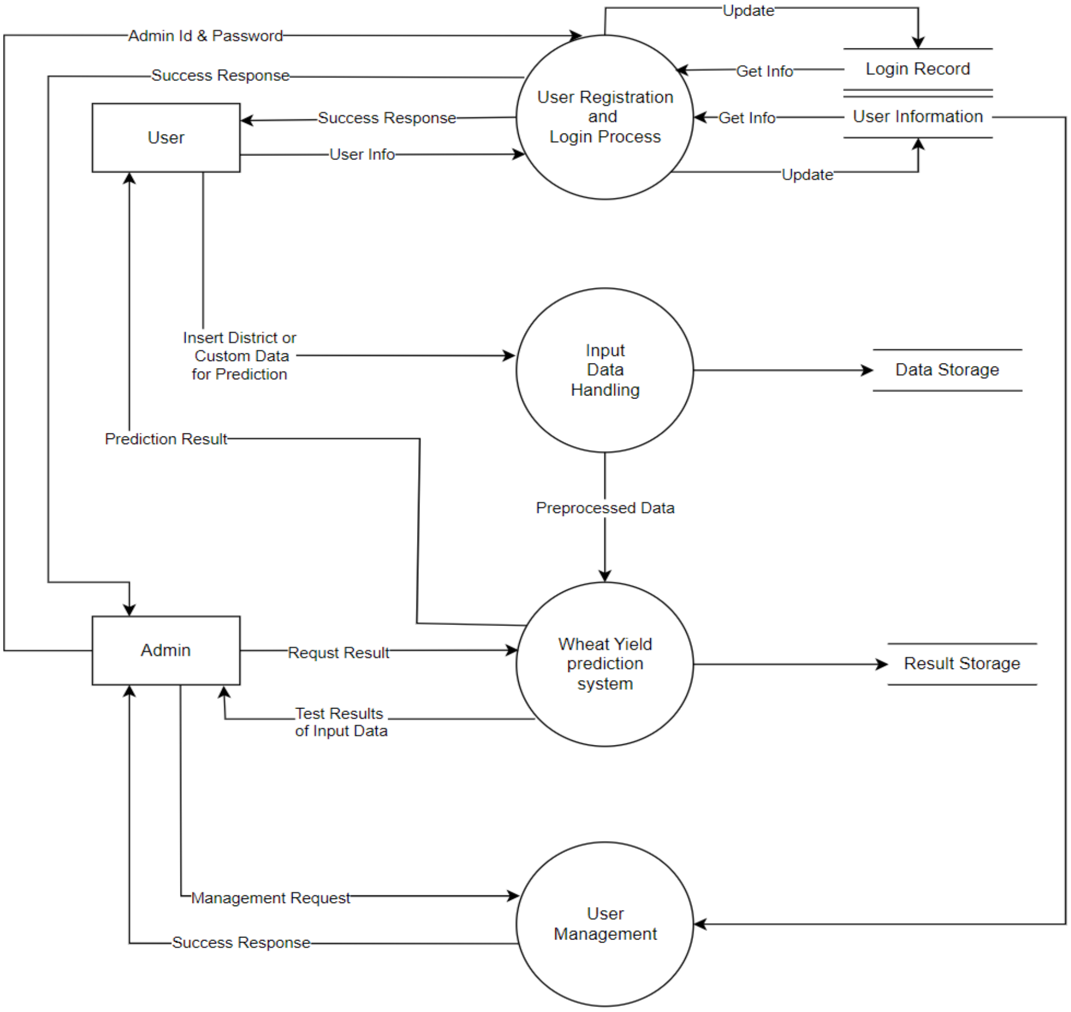


Figure 3.6: DFD Level 1 of Wheat Yield Prediction System

This diagram represents the workflow of a wheat yield prediction system, highlighting key processes such as user registration, input data handling, prediction generation, and user management. Users register or log in through the User Registration and Login Process, which stores and retrieves login records and user information. After logging in, users can submit district-specific or custom data via the Input Data Handling process, where the data is preprocessed and stored. The preprocessed data is sent to the Wheat Yield Prediction System, which generates prediction results and stores them. Admins oversee the system, managing user information, retrieving test results, and handling requests for prediction results through the User Management process. This integrated flow ensures a streamlined system for accurate wheat yield predictions and effective user and data management.

CHAPTER 4:
SYSTEM DESIGN

4.1 Design

Since the project follows Structured Approach**,** this section will document the refinement and creation of various structured oriented diagrams.

4.1.1 Database Design: Transformation of ER to relations and normalizations

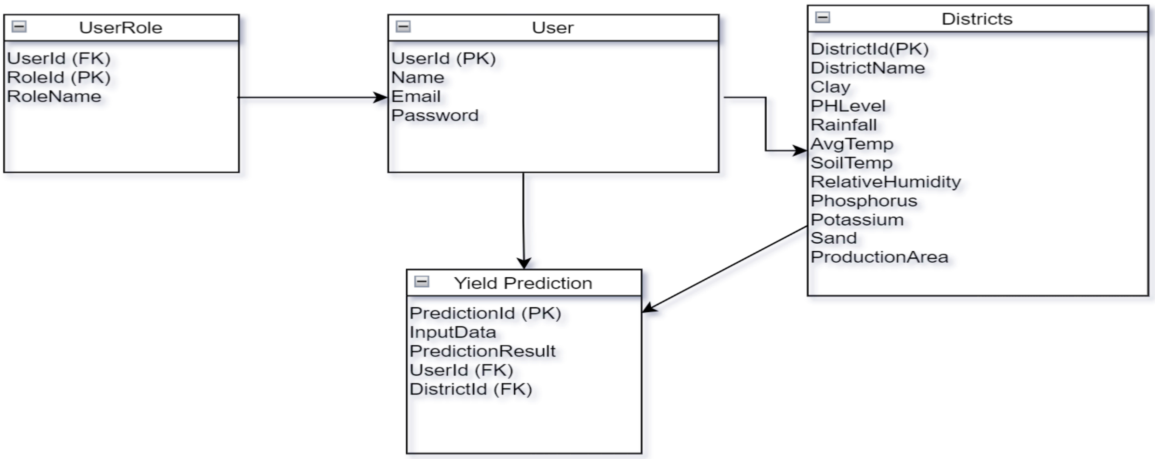


Figure 4.1: Database Design of the System

The above diagram represents the database schema for a yield prediction system. It consists of four main entities: User, UserRole, Districts, and Yield Prediction. The User table stores user details like UserId (Primary Key), Name, Email, and Password. The UserRole table links users to specific roles via a composite primary key of UserId (Foreign Key) and RoleId. The Districts table contains information about different districts, including attributes like soil composition (Clay, PH Level, Phosphorus, etc.), climatic conditions (Rainfall, AvgTemp, etc.), and production area. The Yield Prediction table stores prediction data, where each prediction is linked to a specific user and district through foreign keys (UserId, DistrictId). The relationships indicate that users can have multiple roles, make yield predictions, and reference district-related data for predictions.

4.1.2 Forms and Report Design

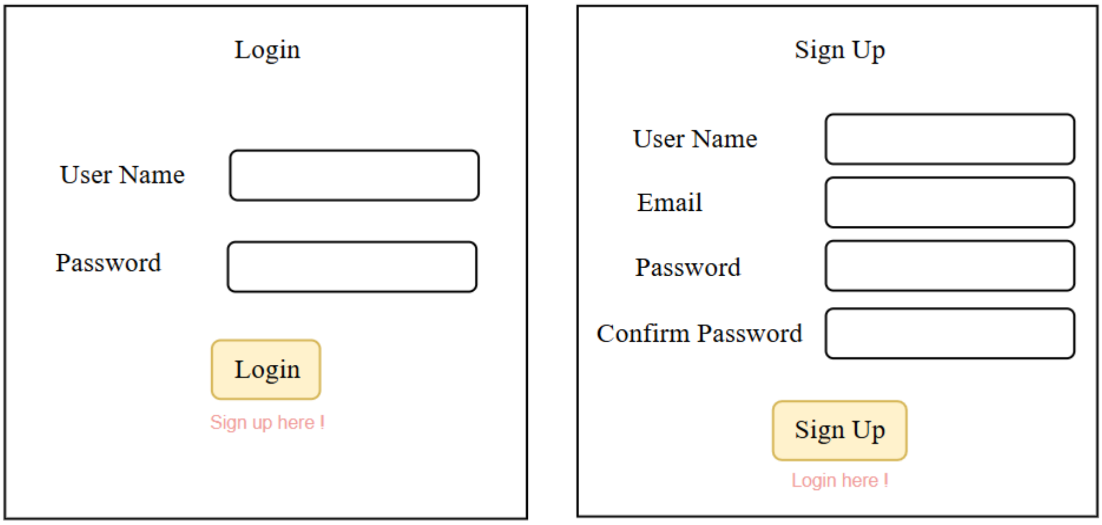
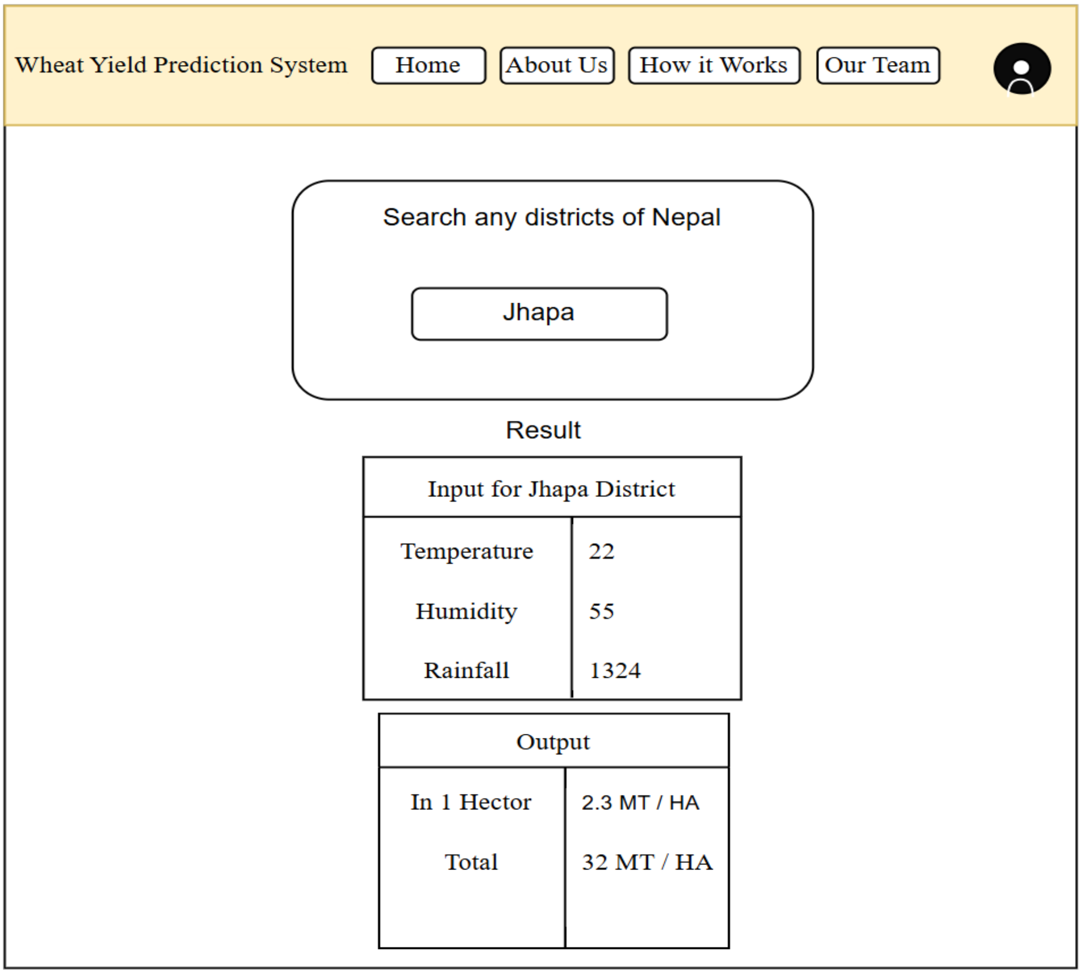


Figure 4.2: Login and Sign up Form Design of the System

These are the designs for the login and signup pages of the system. The login section on the left contains fields for entering a username and password, along with a "Login" button and a "Sign up here!" link for navigation. The signup section on the right includes fields for a username, email, password, and confirm password, with a "Sign Up" button and a "Login here!" link to switch to the login page. Both sections are enclosed within bordered boxes, providing a structured and user-friendly authentication interface.

4.1.3 Interface Design



Search any districts of Nepal

Jhapa

Result

Input for Jhapa District	
Temperature	22
Humidity	55
Rainfall	1324

Output	
In 1 Hector	2.3 MT / HA
Total	32 MT / HA

Figure 4.3: Interface Design of Search page of the System

This is the user interface design for a Wheat Yield Prediction System. The top section contains a navigation bar with buttons for "Home," "About Us," "How it Works," and "Our Team," along with a user profile icon. Below, there is a search section where users can enter and search for any district in Nepal, as shown with "Jhapa" in the input field. The results section displays input parameters such as temperature (22°C), humidity (55%), and rainfall (1324 mm) for the selected district. The output section provides the predicted wheat yield, indicating that in 1 hectare, the yield is 2.3 metric tons per hectare (MT/HA), and the total production is 32 MT/HA. This system likely utilizes machine learning to predict wheat yield based on environmental factors.

4.2 Algorithm Details

The Random Forest algorithm stands as a widely–used machine learning method for regression and classification tasks. It operates as non-parametric and instance-based algorithm, leveraging feature vector similarity within the dataset for decision making. RF employs an ensemble approach, constructing numerous decision trees during training and aggregating prediction to generate final outcomes. Its robustness and simplicity make it a favored choice across various domains for accurate and efficient predictions.

To implement the Random Forest Algorithm from the trees that the system received from the binary decision tree. The system followed the following steps:

Step-1: The data given by the admin is considered to be observed data.

Let X be the matrix of input features for a district, where X_{ij} represents the value of feature j for the district i. The feature vector for a district is denoted as X_i.

Step-2: From a observed data set, a bootstrap data set is taken.

A bootstrap data set is a collection of data set that is randomly picked from the observed dataset. The same data or event from the observed data that may be repeated more than once or may not even be there while taking the data for bootstrap data set. But lesser the repetition better the result.

Step 3: Decision Tree is built from the data in bootstrap dataset

In the Random Forest Regression algorithm, a decision tree is constructed from a bootstrap dataset. The construction process follows a structured approach to ensure that the tree effectively predicts the continuous target variable.

1. Feature Subset Selection

At each node in the decision tree, a random subset of features is selected instead of using all available features. This helps in reducing overfitting and ensures that different trees in the forest make varied decisions.

2. Finding the Best Split Using MSE or MAE

Since the objective of Random Forest Regression is to predict a continuous variable, the best split at each node is determined by minimizing Mean Squared Error (MSE).

The MSE at a node can be expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\, y_i - \hat{y}\,)^2$$

Where:

y_i is the actual value.

ŷ is the predicted value.

n is the number of data points.

The algorithm evaluates different potential split points for each feature and selects the one that minimizes the weighted MSE of the resulting child nodes.

For continuous features such as Rainfall, Temperature, and Soil pH, a threshold is determined to split the data into two groups. For example, if splitting on Rainfall with a threshold Θ = 500*mm*, the data is divided as follows:

Left Child: Rainfall ≤ 500*mm*

Right Child: Rainfall > 500*mm*

This process is repeated recursively to grow the decision tree.

5. Recursive Splitting and Stopping Criteria

Once a split is made, the algorithm continues splitting further until a stopping condition is met. The stopping conditions can be:

A minimum number of samples per leaf node is reached.

Further splits do not significantly reduce the MSE.

The maximum depth of the tree is reached.

Step-4: The average value of all the decision tree is considered to be the final value for the target attribute.

The Random Forest regression involves constructing multiple decision trees, and predictions are based on an ensemble of decision trees. Each individual decision tree provides a prediction, and the final prediction is typically the average of all the tree predictions.

For an individual decision tree, the prediction is determined by the path the data follows through the tree, which can be expressed as:

$$\hat{y}_{tree} = f(x,T)$$

$\hat{y}_{tree}$ is the prediction of specific tree.

x is the feature vector (e.g., values for Rainfall, AvgTemp, etc.),

T represents the decision tree's structure and learned parameters.

For the Random Forest as a whole, the prediction $\hat{y}_{RF}$ is the average of the predictions from all trees in the forest:

$$\hat{y}RF = \frac{1}{N}\sum_{i=1}^{N} f(x, Ti)$$

$\hat{y}_{RF}$ is the final prediction,

N is the total number of trees in the Random Forest,

T_i is the i-th decision tree in the ensemble.

Logic for Random Forest Algorithm:

```
n nodes of trees

decision_tree_result [n]

for _ in range(forest)

        predictions = [predict_with_tree(tree, row) for decision_tree_result in forest]

        result = np.mean(predictions)
```

CHAPTER 5:
IMPLEMENTATION AND TESTING

5.1 Implementation

5.1.1 Tools Used

Python: For building and training the model.

Frameworks and Libraries:
Asp.Net Core: For, creating the backend that is connected with the external API i.e., Flask for data processing and with the frontend to take inputs.
React: For assembling the frontend of the project.

Database:
SQL Server: For archiving the user's data and test results, as well as other related information.

Version Control:
Git/GitHub/GitLab: To coordinate the work of developers and track the changes made to the project in case of using version control systems.

5.1.2 Implementation Details of Module

5.1.2.1 Login/Register Module

The system has user authentication module hosted on a local server. The frontend of the authentication system is built and designed using react framework. Providing intuitive and user friendly webpages for registration and login. These webpages are seamlessly into our system using dotnet api routes, allowing users to access them with ease. For data storage and management, it utilize MSSQL Server. This database stores crucial user information such as emails, usernames and password securely. To enhance user experience and ensure data integrity, the system has form validation. This validation mechanism insure that user inputs the correct and valid information in the login and registration forms, minimizing errors and improving system reliability. Through this comprehensive approach, the system is developed with robust and secure user authentication module that seamlessly integrates into the system, providing user with seamless and secure experience.

5.1.2.2 Preprocessing Module

In this phase of preparing the dataset for machine learning, several crucial steps are undertaken to ensure the data is optimized for training RF algorithms. The raw dataset undergoes a transformation into a structured format conducive to machine learning.

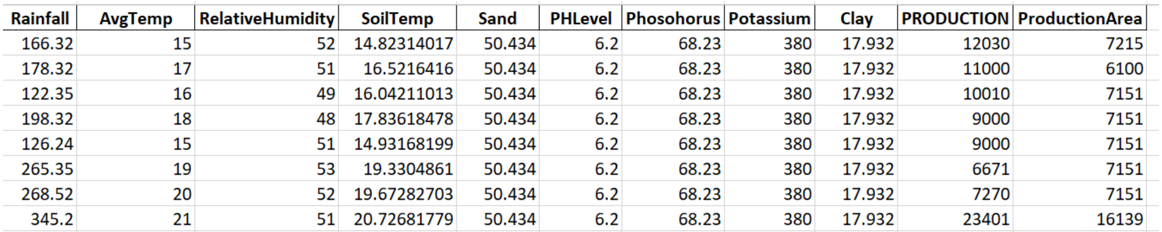
One pivotal aspects involves feature engineering, a process where features are extracted from the raw data utilizing data mining techniques. This step is fundamental in enhancing the performance of RF algorithms by ensuring that the relevant aspects of the data are effectively captured.

To facilitate the RF algorithm we have handled the missing values in dataset by introduction global constant and random values.

We have merged climate data with soil data, then with production area data and finally with production data to make a complete dataset. While doing this we have considered the data in different sets which have same date.

In addition to feature engineering, addressing missing data, dimensionality reduction techniques are employed to streamline the dataset. This involves identifying and removing columns or features that are deemed less important for the statistical analysis or machine learning model. By reducing the dimensionality of the dataset, computational efficiency is improved, and the model performance may be enhanced.

By executing this steps, the dataset is primed for training RF algorithms, ensuring that the model can effectively leverage the available information to make accurate predictions in the context of wheat production analysis.



Rainfall	AvgTemp	RelativeHumidity	SoilTemp	Sand	PHLevel	Phosohorus	Potassium	Clay	PRODUCTION	ProductionArea
166.32	15	52	14.82314017	50.434	6.2	68.23	380	17.932	12030	7215
178.32	17	51	16.5216416	50.434	6.2	68.23	380	17.932	11000	6100
122.35	16	49	16.04211013	50.434	6.2	68.23	380	17.932	10010	7151
198.32	18	48	17.83618478	50.434	6.2	68.23	380	17.932	9000	7151
126.24	15	51	14.93168199	50.434	6.2	68.23	380	17.932	9000	7151
265.35	19	53	19.3304861	50.434	6.2	68.23	380	17.932	6671	7151
268.52	20	52	19.67282703	50.434	6.2	68.23	380	17.932	7270	7151
345.2	21	51	20.72681779	50.434	6.2	68.23	380	17.932	23401	16139

Figure 5.1: Dataset after preprocessing

5.1.2.3 Training module

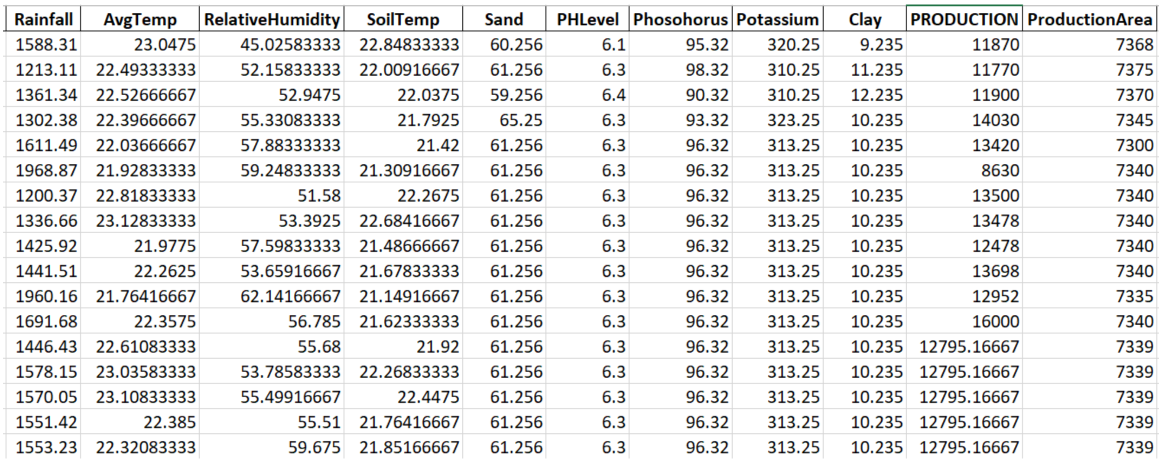
The training data is extracted from the original dataset after performing train test split operation. It comprises data for 892 instances, encompassing 10 distinct attributes. These attributes include essential features such as rainfall, avg temp, relative humidity, soil temperature, sand, ph level, phosphorus, potassium, clay, production area which are crucial for predicting wheat production in different districts.

The training set, typically the larger portion of original dataset, is utilized to train the machine learning model. This involves feeding the model with labeled data (features and corresponding target labels, such as production) and adjusting its internal parameters to learn pattern and relationships within the data.

The model learns from the training data through and iterative process, adjusting its internal parameters based on observed errors. Once the model has been trained sufficiently on the training set, its performance is evaluated using the test set.

The training module in our wheat yield prediction system is responsible for training the machine learning model using the training dataset. It employs algorithms such as Random Forest or Decision Trees, leveraging the features in training data to predict the production of wheat. By analyzing historical production data and learning from labeled examples, the model becomes adept at predicting wheat production accurately.

The sample format of train data:



Rainfall	AvgTemp	RelativeHumidity	SoilTemp	Sand	PHLevel	Phosohorus	Potassium	Clay	PRODUCTION	ProductionArea
1588.31	23.0475	45.02583333	22.84833333	60.256	6.1	95.32	320.25	9.235	11870	7368
1213.11	22.49333333	52.15833333	22.00916667	61.256	6.3	98.32	310.25	11.235	11770	7375
1361.34	22.52666667	52.9475	22.0375	59.256	6.4	90.32	310.25	12.235	11900	7370
1302.38	22.39666667	55.33083333	21.7925	65.25	6.3	93.32	323.25	10.235	14030	7345
1611.49	22.03666667	57.88333333	21.42	61.256	6.3	96.32	313.25	10.235	13420	7300
1968.87	21.92833333	59.24833333	21.30916667	61.256	6.3	96.32	313.25	10.235	8630	7340
1200.37	22.81833333	51.58	22.2675	61.256	6.3	96.32	313.25	10.235	13500	7340
1336.66	23.12833333	53.3925	22.68416667	61.256	6.3	96.32	313.25	10.235	13478	7340
1425.92	21.9775	57.59833333	21.48666667	61.256	6.3	96.32	313.25	10.235	12478	7340
1441.51	22.2625	53.65916667	21.67833333	61.256	6.3	96.32	313.25	10.235	13698	7340
1960.16	21.76416667	62.14166667	21.14916667	61.256	6.3	96.32	313.25	10.235	12952	7335
1691.68	22.3575	56.785	21.62333333	61.256	6.3	96.32	313.25	10.235	16000	7340
1446.43	22.61083333	55.68	21.92	61.256	6.3	96.32	313.25	10.235	12795.16667	7339
1578.15	23.03583333	53.78583333	22.26833333	61.256	6.3	96.32	313.25	10.235	12795.16667	7339
1570.05	23.10833333	55.49916667	22.4475	61.256	6.3	96.32	313.25	10.235	12795.16667	7339
1551.42	22.385	55.51	21.76416667	61.256	6.3	96.32	313.25	10.235	12795.16667	7339
1553.23	22.32083333	59.675	21.85166667	61.256	6.3	96.32	313.25	10.235	12795.16667	7339

Figure 5.2: Train Dataset

5.1.2.4 Testing Module

In Wheat Yield Prediction System the test dataset comprises of 383 instances, each containing each attributes essential for wheat yield prediction system. Following the train-test spilt process, the test dataset serves as a critical component for evaluating the performance of our trained model. It represents unseen data that allows us to assess the models ability to generalize and make accurate predictions on new instances. By feeding the test dataset into trained model, analyze of its predictive performance, identify any potential issues such as overfitting and ensure the reliability and the effectiveness of the Wheat Yield Prediction System can be performed.

5.1.2.5 Description of Classes

The Random Forest class consists of several important parameters that influence its performance behavior. For instance, the ‘n_trees’ parameter determines the number of decision trees in the ensemble, affecting the model’s complexity and robustness. ‘Max_depth’ limits the maximum depth of individual trees, guarding against the complexity and aiding in generalization. ‘n_features’ specifies the maximum number of features considered for splitting each node, introducing randomness and reducing correlation between trees. Additionally ‘min_size’ dictates the minimum number of samples required to split a node, preventing overfitting by restricting node splits with few samples. By adjusting these parameters, practitioners can fine-tune the Random Forest model to match the characteristics of dataset and achieve optimal predictive performance.

Decision Tree class also involve essential parameters such as ‘Max_depth’, ‘n-features’, ‘min_size’. ‘Max_depth’ defines the maximum depth of individual trees, guarding against the complexity and aiding in generalization. ‘n_features’ specifies the maximum number of features considered for splitting each node, introducing randomness and reducing correlation between trees. ‘min_size’ dictates the minimum number of samples required to split a node, preventing overfitting by restricting node splits with few samples. These parameters are critical in shaping the structure and behavior of decision trees.

5.2 Testing

Testing was carried out manually to ensure each component works fine.

5.2.1 Unit Testing

Table 5.1: Unit testing of Wheat Yield Prediction System

Test Case ID	Test Case Description	Test Cases	Result	Remarks
TC-1	Data Input Validation	Verify that the system accepts validate climate, soil and fertilizer input values.	Data is accepted and stored correctly	Pass
TC-2	Invalid Input Handling	Verify that the system rejects invalid or missing input values	System displays an appropriate error message.	Pass
TC-3	Prediction Algorithm	Verify that the Random Forest algorithm processes the input correctly and generates a yield prediction	Predicted yield value is displayed based on the input data	Pass
TC-4	Database Connectivity	Verify that input data is correctly fetched from the database for each district	Data is fetched without errors	Pass
TC-5	Prediction Accuracy	Verify that the system provides 85%+ accurate results based on test datasets	Prediction results are within an acceptable margin of error.	Pass

5.2.2 System Testing

Table 5.2: System testing of Wheat Yield Prediction System

Test Case ID	**Test Case Description**	**Test Case**:	**Steps**	**Expected Result**	**Actual Result**
TC-1	Input Data Flow	Verify that users can input	Enter valid data in all fields	Data is stored and passed to the prediction algorithm	Pass
TC-2	Yield Prediction Flow	Verify the complete workflow from data input to yield prediction	Enter input data -> Submit -> View prediction results.	The system generates and displays yield predictions	Pass
TC-3	Performance Testing	Verify that the system processes large input datasets without delays	Upload a large dataset -> Run prediction	System processes data and generates predictions	Pass

Input Data For Jhapa	
Rainfall (mm)	1344.74
Avg Temp (C)	21.54
Relative Humidity ()	64.09
Soil Temp (C)	21.13
Sand (%)	41.26
PH Level	5.9
Phosphorus (kg/ha)	60.19
Potassium (kg/ha)	240.23
Clay (%)	18.03
Production Area (Hectare)	15500

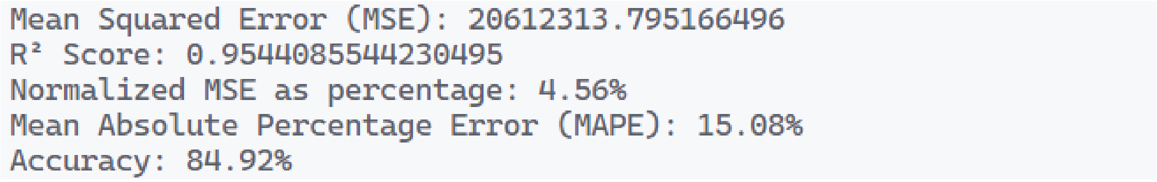
Predicted Result For Jhapa	
Prediction (Metric Ton)	39009.5045

Input Data For Bajhang	
Rainfall (mm)	943.86
Avg Temp (C)	5.04
Relative Humidity ()	73.13
Soil Temp (C)	4.83
Sand (%)	50.01
PH Level	6.6
Phosphorus (kg/ha)	116.22
Potassium (kg/ha)	466.58
Clay (%)	10.24
Production Area (Hectare)	10982

Predicted Result For Bajhang	
Prediction (Metric Ton)	19906.0698

5.3 Result Analysis

Performance Evaluation: This evaluates the alignment between the prediction and the real outcomes. The assessment might include metrics like accuracy, mean square error, RSquare and mean absolute percentage error.



```
Mean Squared Error (MSE): 20612313.795166496
R² Score: 0.9544085544230495
Normalized MSE as percentage: 4.56%
Mean Absolute Percentage Error (MAPE): 15.08%
Accuracy: 84.92%
```

Mean Squared Error (MSE): This measures the average squared difference between the actual and predicted values. A lower MSE indicates better model performance. Here, an MSE of 20,612,313 suggests that the predictions deviate significantly from actual values on average, depending on the scale of the data.

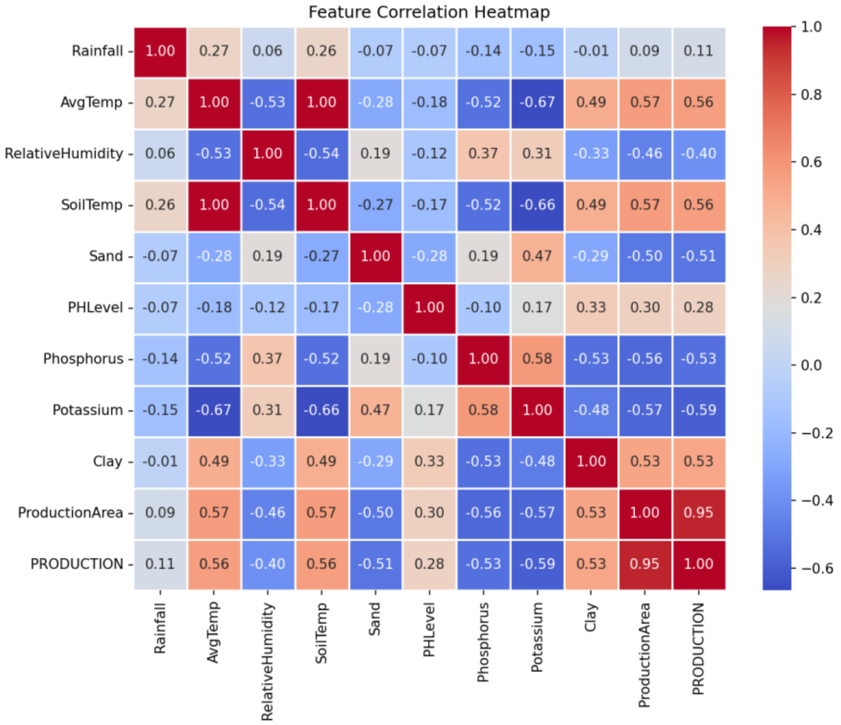
R² Score: Also known as the coefficient of determination, this indicates how well the model explains the variance in the target variable. A score of 0.954 suggests that 95.4% of the variability in the data is captured by the model, meaning it has high predictive power.

Normalized MSE as Percentage: This scales the MSE relative to the range of the data to provide a more interpretable value. A normalized MSE of 4.56% suggests that the error is small compared to the data range.

Mean Absolute Percentage Error (MAPE): This metric expresses error as a percentage of actual values, making it useful for comparing across datasets. A MAPE of 15.08% means that, on average, predictions are about 15.08% off from the actual values.

Accuracy: This is likely calculated as (100% - MAPE), indicating how close the model’s predictions are to the actual values. An accuracy of 84.92% suggests that the model’s predictions are reasonably accurate but leave some room for improvement.

Correlation Heatmap



This correlation heatmap visually represents the relationships between various features in the dataset, with colors indicating the strength and direction of correlations. Red shades signify strong positive correlations, meaning that as one variable increases, the other also tends to increase, while blue shades indicate strong negative correlations, where one variable rises as the other falls. A key observation is the high correlation (0.95) between Production and ProductionArea, suggesting that larger production areas generally yield higher production, which could lead to multicollinearity in regression models. Additionally, AvgTemp (0.56) and SoilTemp (0.56) show a moderate positive correlation with production, indicating that temperature plays a crucial role in crop yield. However, Phosphorus (-0.53) and Sand (-0.51) exhibit negative correlations with production, implying that excessive phosphorus levels and sandy soil might not be favorable for high yields. Furthermore, the inverse relationship between AvgTemp and Relative Humidity (-0.53) aligns with natural weather patterns, where higher temperatures typically reduce humidity. Understanding these correlations can aid in feature selection, soil optimization, and predictive modeling to improve agricultural outcomes.

CHAPTER 6:
CONCLUSIONS AND FUTURE RECOMMENDATIONS

6.1 Conclusion

The Wheat Yield Prediction System successfully addresses the challenge of forecasting wheat production based on key environmental and soil parameters such as rainfall, temperature, humidity, soil composition, and nutrient levels. By implementing a Random Forest Regression model from scratch, the system provides reliable and accurate yield predictions, helping farmers, agricultural researchers, and policymakers make data-driven decisions to optimize wheat production.

The system's ability to handle a variety of input factors makes it scalable and adaptable for different climatic and geographical conditions. Designed specifically for Nepal’s 77 districts, the model can be further customized to support additional regions with similar agricultural landscapes. The inclusion of multiple soil and climate parameters ensures that the predictions reflect real-world agricultural conditions, making the model highly practical for real-time decision-making. Throughout the project, rigorous testing and validation were conducted to ensure the reliability of predictions. The system successfully met the objectives set out in the project proposal, demonstrating the potential of machine learning in modern agriculture. By leveraging historical agricultural data, the system empowers farmers with predictive insights, allowing for better resource allocation, improved crop management strategies, and increased productivity.

The success of this project highlights the broader impact of AI and data science in addressing global food security challenges. With further enhancements, such as integrating real-time weather data, satellite imagery, and remote sensing technologies, the model can be extended to predict yield fluctuations more accurately and suggest optimal farming practices.

In conclusion, this project serves as a significant step forward in the application of machine learning in agriculture. By providing actionable insights into wheat production, it offers a practical, scalable, and innovative solution that has the potential to improve agricultural sustainability and food security in Nepal and beyond.

6.2 Future Recommendations

To enhance the functionality and impact of the Wheat Yield Prediction System, the following improvements are recommended:

1. **Integration of External APIs:**
    - Incorporate APIs for real-time weather and soil data to improve prediction accuracy.
2. **Mobile Application Development:**
    - Extend the system to mobile platforms for increased accessibility, particularly for rural farmers.
3. **Incorporate More Factors:**
    - Include additional factors like pest infestation, irrigation patterns, and crop rotation history.
4. **Enhanced Data Visualization:**
    - Develop more advanced graphical tools for visualizing trends in yield predictions.
5. **AI Model Optimization:**
    - Explore other algorithms like Gradient Boosting or Neural Networks to further improve accuracy and reduce error margins.
6. **Collaboration with Agricultural Institutions:**
    - Partner with agricultural experts and institutions to continuously refine the system with real-world feedback and data.

REFERENCES

[1] R. Thapa-Parajuli and N. Devkota, “Impact of Climate Change on Wheat Production in Nepal,” *Asian Journal of Agricultural Extension, Economics & Sociology, vol. 9, no. 2, pp. 1–14,.* 2016

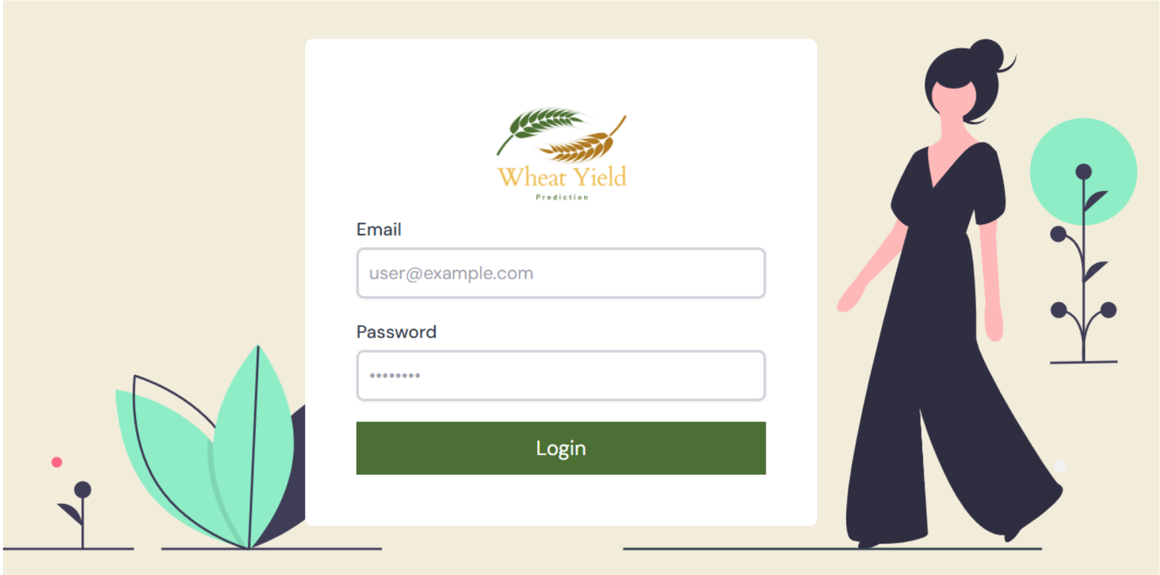
[2] R. A. Sial, E. H. Chaudhary, S. Hussain, and M. Naveed, “ Effect of organic manures and chemical fertilizers on grain yield of maize in rainfed area,” *Soil and Environment (Pakistan) Volume 26 Issue 2 ISSN 1019-729X,* 2007

[3] P. Z. Janjua, G. Samad, and N. Khan, “ Climate Change and Wheat Production in Pakistan: An Autoregressive Distributed Lag Approach,” *NJAS - Wageningen Journal Of Life Sciences, vol. 68, pp. 13–19,* 2014

[4] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, “Wheat yield prediction using machine learning and advanced sensing techniques,” *Computers And Electronics in Agriculture, vol. 121, pp. 57–65,* 2016

[5] Manisha Sanjay Sirsat, Paula Rodrigues Oblessuc, and R. S. Ramiro, “ Genomic Prediction of Wheat Grain Yield Using Machine Learning,” *Agriculture, vol. 12, no. 9, pp.1406,* 2022

APPENDIX



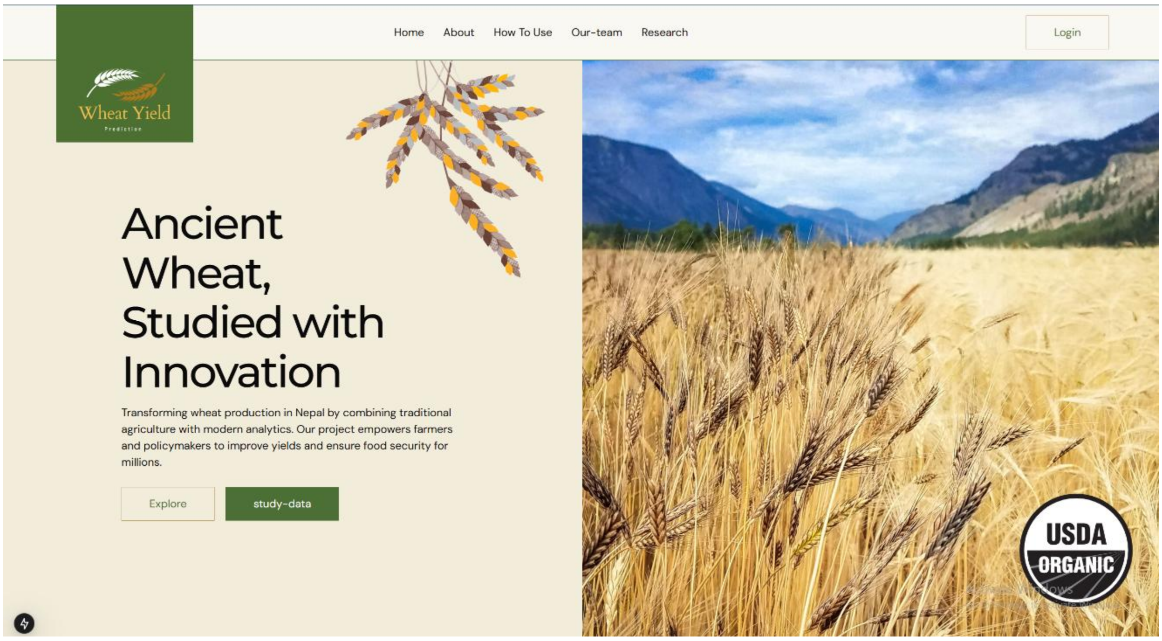
Email

user@example.com

Password

••••••••

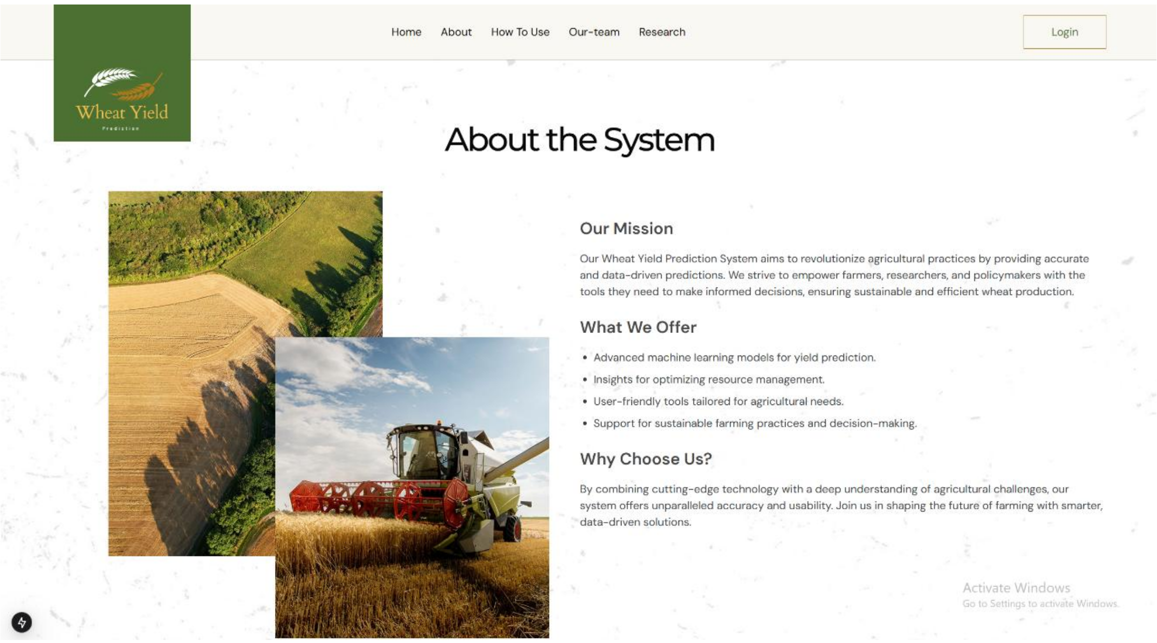
Figure: Login Form of Wheat Yield Prediction System



Ancient Wheat, Studied with Innovation

Transforming wheat production in Nepal by combining traditional agriculture with modern analytics. Our project empowers farmers and policymakers to improve yields and ensure food security for millions.

Figure: Landing page of Wheat Yield Prediction System



About the System

Our Mission

Our Wheat Yield Prediction System aims to revolutionize agricultural practices by providing accurate and data-driven predictions. We strive to empower farmers, researchers, and policymakers with the tools they need to make informed decisions, ensuring sustainable and efficient wheat production.

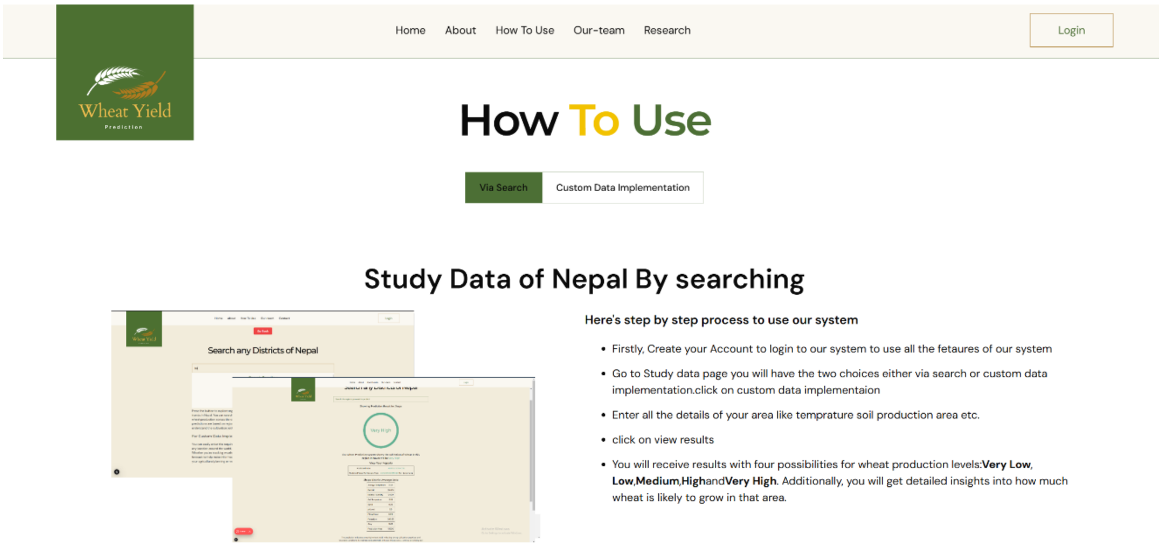
What We Offer

- Advanced machine learning models for yield prediction.
- Insights for optimizing resource management.
- User-friendly tools tailored for agricultural needs.
- Support for sustainable farming practices and decision-making.

Why Choose Us?

By combining cutting-edge technology with a deep understanding of agricultural challenges, our system offers unparalleled accuracy and usability. Join us in shaping the future of farming with smarter, data-driven solutions.

Figure: About section of the Wheat Yield Prediction System

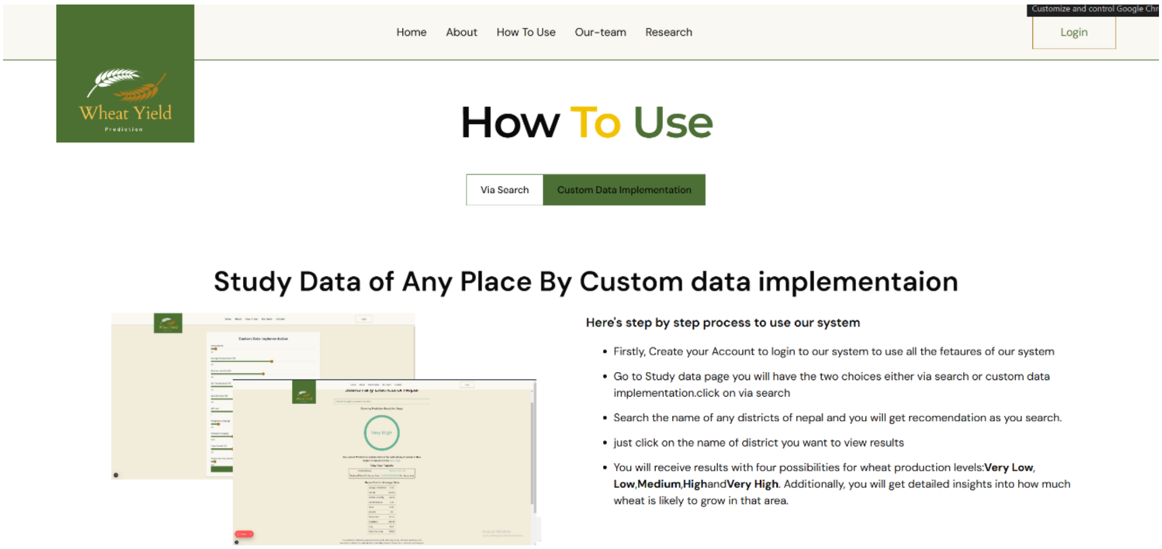


How To Use

Study Data of Nepal By searching

Here's step by step process to use our system

- Firstly, Create your Account to login to our system to use all the fetaures of our system
- Go to Study data page you will have the two choices either via search or custom data implementation.click on custom data implementaion
- Enter all the details of your area like temprature soil production area etc.
- click on view results
- You will receive results with four possibilities for wheat production levels:**Very Low**, **Low**,**Medium**,**High**and**Very High**. Additionally, you will get detailed insights into how much wheat is likely to grow in that area.



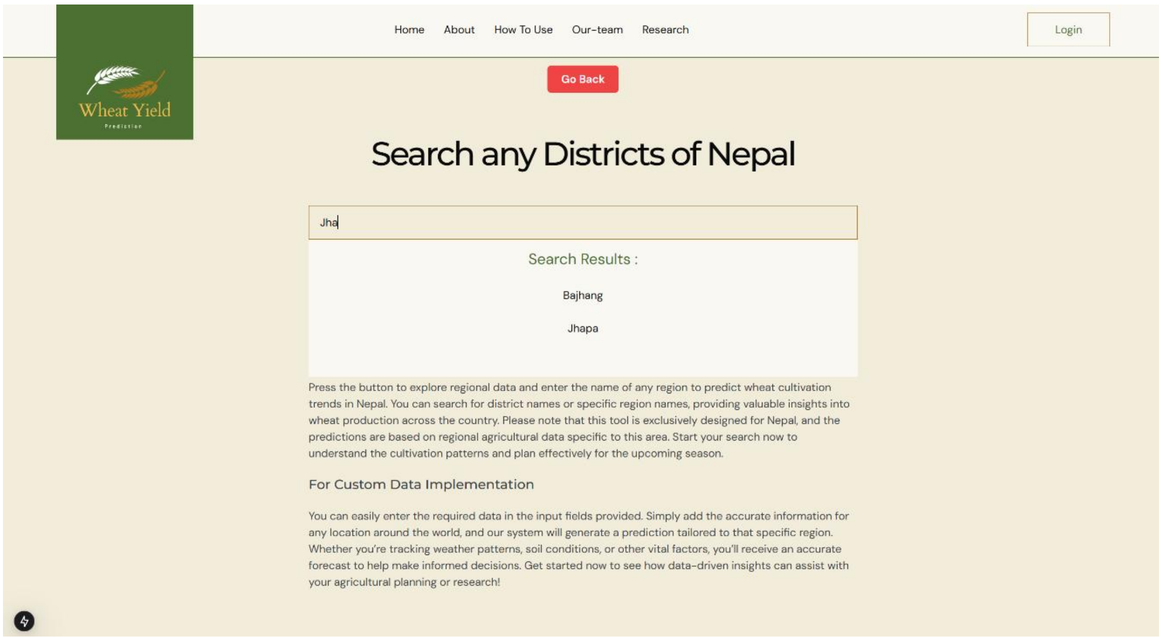
How To Use

Study Data of Any Place By Custom data implementaion

Here's step by step process to use our system

- Firstly, Create your Account to login to our system to use all the fetaures of our system
- Go to Study data page you will have the two choices either via search or custom data implementation.click on via search
- Search the name of any districts of nepal and you will get recomendation as you search.
- just click on the name of district you want to view results
- You will receive results with four possibilities for wheat production levels:**Very Low**, **Low**,**Medium**,**High**and**Very High**. Additionally, you will get detailed insights into how much wheat is likely to grow in that area.

Figure: User guide page of Wheat Yield Prediction System



Search any Districts of Nepal

Jha

Search Results :

Bajhang

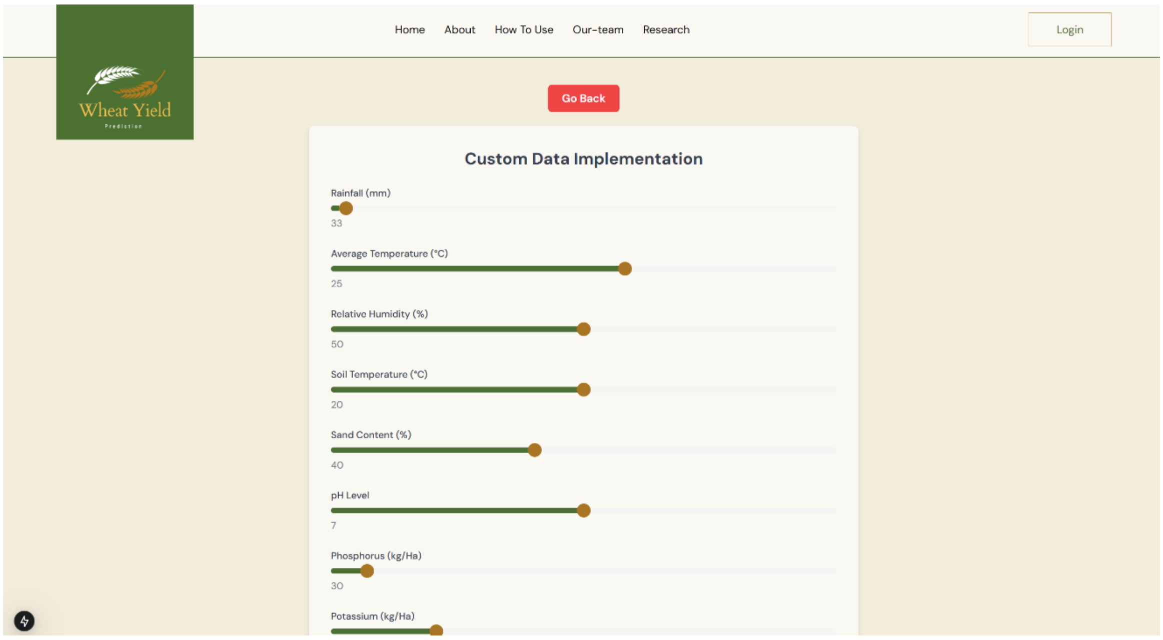
Jhapa

Press the button to explore regional data and enter the name of any region to predict wheat cultivation trends in Nepal. You can search for district names or specific region names, providing valuable insights into wheat production across the country. Please note that this tool is exclusively designed for Nepal, and the predictions are based on regional agricultural data specific to this area. Start your search now to understand the cultivation patterns and plan effectively for the upcoming season.

For Custom Data Implementation

You can easily enter the required data in the input fields provided. Simply add the accurate information for any location around the world, and our system will generate a prediction tailored to that specific region. Whether you're tracking weather patterns, soil conditions, or other vital factors, you'll receive an accurate forecast to help make informed decisions. Get started now to see how data-driven insights can assist with your agricultural planning or research!

Figure: Search by district page of Wheat Yield Prediction System



Custom Data Implementation

Rainfall (mm)
33

Average Temperature (°C)
25

Relative Humidity (%)
50

Soil Temperature (°C)
20

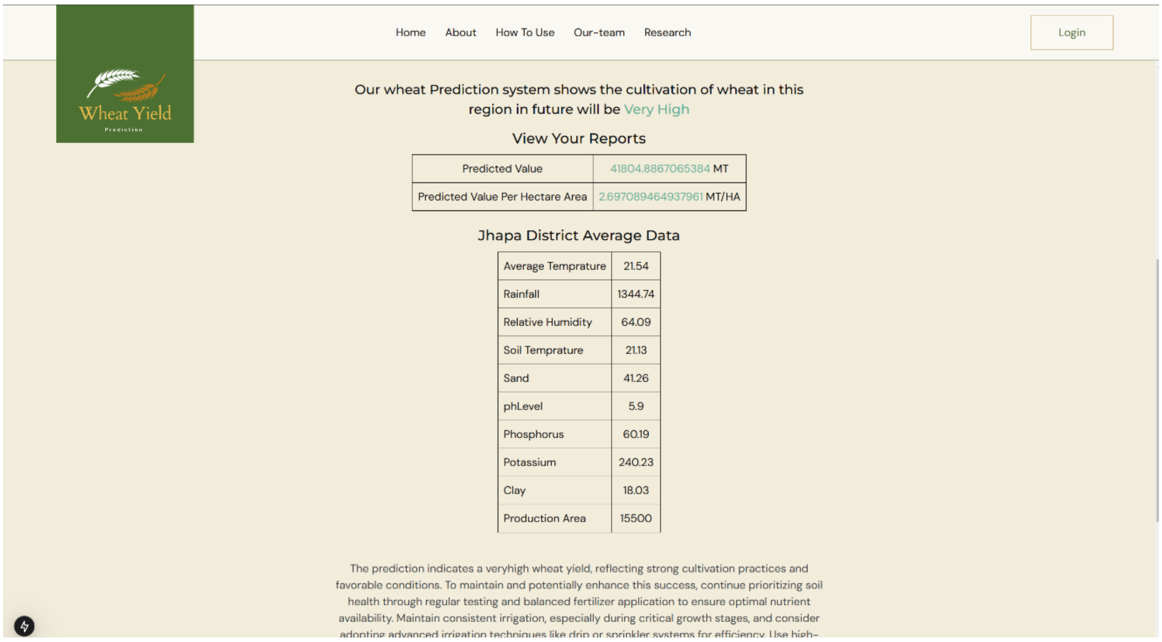
Sand Content (%)
40

pH Level
7

Phosphorus (kg/Ha)
30

Potassium (kg/Ha)

Figure: Custom data implementation page of Wheat Yield Prediction System



Our wheat Prediction system shows the cultivation of wheat in this region in future will be Very High

View Your Reports

Predicted Value	41804.8867065384 MT
Predicted Value Per Hectare Area	2.697089464937961 MT/HA

Jhapa District Average Data

Average Temprature	21.54
Rainfall	1344.74
Relative Humidity	64.09
Soil Temprature	21.13
Sand	41.26
phLevel	5.9
Phosphorus	60.19
Potassium	240.23
Clay	18.03
Production Area	15500

The prediction indicates a veryhigh wheat yield, reflecting strong cultivation practices and favorable conditions. To maintain and potentially enhance this success, continue prioritizing soil health through regular testing and balanced fertilizer application to ensure optimal nutrient availability. Maintain consistent irrigation, especially during critical growth stages, and consider adopting advanced irrigation techniques like drip or sprinkler systems for efficiency. Use high-

Figure: Result of the Prediction of Wheat Yield Prediction System