TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Kathmandu Engineering College

Department of Computer Engineering

Minor Project On

ClimaCrop - Optimal Crops Predictor Using Machine Learning

[Code No: CT 654]

By

Rejina Dangol (KAT077BCT057)

Shreyas Acharya (KAT077BCT080)

Sneha Thakur (KAT077BCT086)

Swapnil Poudyal (KAT077BCT092)

Kathmandu, Nepal

2080

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Kathmandu Engineering College

Department of Computer Engineering

ClimaCrop - Optimal Crops Predictor Using Machine Learning

[Code No: CT 654]

PROJECT REPORT SUBMITTED TO

DEPARMENT OF COMPUTER ENGINEERING

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR

THE BACHELOR IN ENGINEERING

By

Rejina Dangol (KAT077BCT057)

Shreyas Acharya (KAT077BCT080)

Sneha Thakur (KAT077BCT086)

Swapnil Poudyal (KAT077BCT092)

Kathmandu, Nepal

2080

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Kathmandu Engineering College
Department of Computer Engineering

CERTIFICATE

The undersigned certify that they have read and recommended to the Department of Computer Engineering, a minor project work entitled "Climacrop - Optimal Crop Predictor Using Machine Learning" submitted by Rejina Dangol - 77057, Shreyas Acharya - 77080, Sneha Thakur - 77086 and Swapnil Poudyal - 77092 in partial fulfillment of the requirements for the degree of Bachelor of Engineering.

Er Manish Aryal (External Examiner) Department of Computer Engineering Sagarmatha Engineering College	Er Kunjan Amatya (Project Supervisor) Department of Computer Engineering Kathmandu Engineering College
Ms. Krista Byanju (Project Coordinator) Department of Computer Engineering Kathmandu Engineering College	Mr Sudeep Shakya (Head of Department) Department of Computer Engineering Kathmandu Engineering College

ACKNOWLEDGEMENT

We must acknowledge our deepest of gratitude to everyone who have contributed in this project on Crop Prediction using Machine Learning**.** The efforts, guidance and support of various colleagues involved in this project has played an important role. First and foremost, we would like to thank Institute of Engineering for providing us with an opportunity to develop a project under the academic requirement as a minor project.

We are very grateful towards **Er. Sudeep Shakya,** Head of the Department and **Er. Kunjan Amatya,** Deputy Head of the Department, Computer Engineering and as Our Supervisor for their invaluable guidance, effort and time. Their feedback and insights were very crucial in improving the quality of the project and shaping it towards the right direction. Likewise, we must acknowledge our obligation towards **Er. Anju Khanal** ma'am, **Er. Krista Byanju** ma'am and **Er. Dhawa Song Dong** for enlightening us with their expertise and helping us throughout the project.

Moreover, we would like to thank Department of Computer Engineering, Kathmandu Engineering College for providing us with an opportunity to explore the nature of our field and build a project of our interest which has highly helped us to implement the knowledge and skills gained over the years as the minor project.

Lastly, we wish to record our appreciation to all of our friends and family who have directly or indirectly helped us throughout this journey.

ABSTRACT

The agricultural sector is going through considerable changes brought on by a variety of things, including technological development, environmental concerns, and shifting consumer tastes. Precision agriculture, remote sensing, the Internet of Things (IoT), machines, and automation are examples of technical advancements that the agricultural sector is adopting. ClimaCrop will allow the farmers to have access to various food crop and cash crop and their yield on the land. ClimaCrop uses Neural Network for its implementation and upon testing with 500 epochs, the Neural Networking had about 65 % accuracy and MAE of 10000. But the Traditional Machine learning Method had an approximate accuracy of 94% using Random Forest Algorithm. With a dataset of 30000, 80% of the data i.e around 24000 data were used to train the model whereas the remaining 20% ie around 6000 data were used to test the model. The objective of this minor project is to provide a platform based on data and calculations with advanced algorithms, powerful machine learning capabilities, and access to a vast array of data, making it an indispensable tool for the farmers. This minor project will significantly contribute in this field by introducing a system which will help the farmers in crop prediction and yield increment.

Keywords: *ClimaCrop, Machine Learning, RNN, Random Forest, Data processing*

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CSM	Crop Selection Method
GDP	Gross Domestic Product
IoT	Internet of Things
JS	Java Script
KNN	K-Nearest Neighbor
NARC	National Agriculture Research Council
OSI	Open Systems Interconnection
PA	Precision Agriculture
RNN	Recurrent Neural Network
SDK	Software Development Kit
SDLC	Software Development Life Cycle
SQL	Structured Query Language
UI/UX	User Interface / User Experience
ULM	User Lifecycle Management
UML	Unified Modeling Language

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND THEORY

Nepal is an agricultural country, around 66% of the total population is significantly involved in the agriculture sector. It contributes to about one-third of the nation's GDP and has a significant contribution to the nation's economy. Crops are the foundation of our global food system, and it plays a critical role in meeting the always increasing demands for nutrition, and economic stability. Agriculture generates a wide variety of employment opportunities, from farming to small-scale businesses. In addition to frequent natural disasters such floods, landslides, earthquakes, illnesses, and various outbreaks, Nepal is also subject to food insecurity. Farming continues to dominate the nation's agriculture sector, which has led to low agricultural commodity production and productivity. Due to a lack of modern farming skills, most farmers in Nepal follow the traditional methods of farming. They cultivate the most popular food crops such as rice, wheat, maize, and lentils in their area. Similarly, cash crops such as coffee, tea, cotton, cardamom are popular in major regions of Nepal. But it has a negative impact on the fertility of the ground.

Food crops are essential to Nepali agriculture, the nation's economy, and society. Due to the country's primarily agrarian economy and high farming labor force, food crops are essential to livelihoods, cultural traditions, and food security. Many different types of food crops, such as rice, maize, wheat, millet, barley, pulses, and potatoes, can be grown in the nation due to its varied topography and climates.

A cash crop is one that is grown with the goal of selling it for a profit. The majority of crops farmed today are cash crops, meaning they are grown with the intention of being sold on domestic and international markets. The majority of cash crops grown in underdeveloped countries are exported to industrialized countries for a higher price. Food crops are those that are grown primarily for human consumption.

The idea of cash crops first appeared thousands of years ago as agricultural economies and trade networks began to grow. In the past, societies raised cash crops like sugar, cotton, and spices for export to far-off markets, which fueled economic expansion and promoted cross-cultural interaction.

Cash crop cultivation frequently calls for particular soil types, climates, and cultivation practices designed to optimize yield and quality. Consequently, cash crops have been instrumental in influencing global agricultural practices, with different regions focusing on producing particular commodities due to their inherent benefits and financial incentives.

Cash crop cultivation increased with the rise of colonialism and globalization, as European powers established plantations in colonies to profitably exploit labor and natural resources. Globally, this practice profoundly altered economies, societies, and landscapes, resulting in both environmental degradation and economic prosperity. With staples like wheat, corn, soybeans, and rice dominating global production, cash crops remain a key component of many economies in modern agriculture. The need for responsible supervision of agricultural resources has been highlighted by the growing emphasis on fair trade practices and sustainable agriculture drawing attention to the social and environmental implications of cash crop production.

Plants raised mainly for human consumption are known as food crops, and they form the basis of world food systems. A wide variety of species are included in these crops, such as vegetables, fruits, in addition to cereals like rice, wheat, maize, and millet. Food crops have been grown since the beginning of agriculture, which is a major turning point in human history shifting animalistic nature of human societies to permanent agricultural communities. Food crops are vital sources of vitamins, minerals, and macronutrients that sustain and nourish people all over the world. They influence dietary customs, culinary traditions, and agricultural landscapes in many places and civilizations, and they are essential to food security, livelihoods, and cultural practices.

The production of food crops in modern agriculture is influenced by a complex interaction of variables, including soil fertility, climate, water availability, insect control, and agricultural technologies. Crop rotation and organic farming are examples of sustainable farming techniques that attempt to reduce their negative effects on the environment and increase long-term food security.

However, a number of obstacles must be overcome in order to produce food crops, such as disruptions brought on by climate change, land degradation, water scarcity, pests and

diseases, market volatility, and social inequality. To ensure reasonable access to healthy food while protecting environmental resources and promoting resilient food systems, addressing these issues calls for creative solutions and cooperative efforts among governments, researchers, and farmers.

In South Asia, rice-wheat systems produce more than 30% of the rice and 42% of the wheat consumed (CIMMYT 2015). It is also the dominant cropping system among other cereal cropping production systems in Terai region of Nepal. In Nepal, rice is farmed on 1.5 million hectares and accounts for 37% of the country's total rice and 85% of its wheat. It is mostly grown as a sequence of crops produced by rainfall. Nawalaparasi being one of the main districts in contributing about 5.44% of total harvesting area of rice (Adhikari 2020), but the productivity is felt declining over the last three decades and henceforth, its yield over last three decades along with the anomalies of agro-climatic indices like fluctuating maximum and minimum temperatures, solar radiation and rainfall have been studied[1].

Climate refers to the long-term weather patterns, including temperature, precipitation, humidity, and wind. Rainfall is a critical component of climate and has a direct impact on crop growth and yielding. By understanding the regional and seasonal rainfall patterns, such as average rainfall amounts, variations, and trends, we can predict the suitability of different crops for specific areas.

Farmers gain a lot of advantages from scientific farming skills, the ClimaCrop initiative works to avoid issues including incorrect crop selection, decreased productivity, and vulnerability to climatic risks, environmental effects, and difficulties with food security. Our project also provides a solution to optimize crop choices depending on rainfall patterns and encourage sustainable and adaptable agriculture practices by utilizing data-driven insights and prediction models.

1.2 PROBLEM STATEMENT

Farmers struggle to get timely, accurate information on rainfall patterns and how it affects the appropriateness of crops. They are unable to choose crops intelligently due to this information gap, which results in less-than-ideal yields and decreased agricultural production. Ineffective crop selection stems from a lack of knowledge about the connection between rainfall patterns and crop suitability. Farmers frequently use tried-and-true techniques or traditional procedures, which may not be compatible with the current rainfall conditions. Lower yields, financial losses, and decreased food security are all results of this inefficiency.

Furthermore, the variety and unpredictability of rainfall patterns have grown because of climate change. It is difficult for farmers in the target region to adjust their crop selection tactics to these shifting environmental factors. It is challenging for farmers to plan and maximize their agricultural techniques due to the uncertainty surrounding anticipated rainfall patterns.

With global warming increasing rapidly, temperature also plays a vital role on deciding the production. Immense drop or rise in temperature may drastically affect the production rate. It is difficult for farmers to precisely choose the crop.

Alongside these parameters pesticides and its availability has a major impact on the yield and production. Farmers rely on pesticides to manage insect pests that can devastate crops by feeding on leaves, stems, roots, or fruits. Similarly, pesticides such as fungicides and bactericides are essential for preventing and managing disease outbreaks that can spread rapidly and cause extensive damage to food crops. Even with proper amount of pesticides used the production of the crops is not up to par. Farmers are uneducated on the field of pesticides and its uses.

In the context of Nepal, the vulnerability in climate change is not something very rare. The nation is vulnerable to alterations in rainfall patterns, an increase in the frequency of extreme weather events, and protracted drought or flood conditions. Farmers require assistance and support in making crop choices that are resistant to these climate risks, but their capacity to lessen the negative effects is affected by the absence of specialized information and resources.

A comprehensive strategy is needed to address these issues, one that emphasizes timely and accurate information access, technical capacity building among farmers, infrastructure investments for data collection and distribution, the promotion of climate-resilient agricultural practices, and targeted assistance for crop adaptation based on rainfall patterns.

1.3 OBJECTIVES

- To find patterns, fluctuations in rainfall and recommend accurate crops to sow for maximum yield using KNN, Linear Regression, Lasso, Ridge, DTR, Random Forest and neural networking.
- To encourage sustainable farming and to address farming problems brought on by ineffective methods.

1.4 SCOPE AND APPLICATIONS

Right now there aren't many mechanisms in Nepal that are aimed at making farming easier. Additionally, the agriculture industry is not entirely exposed to the current technology. The majority of the systems are run by the government and are not readily accessible to the residents or the farmers in accordance with their requirements.

The ClimaCrop project's scope includes the creation and application of a data-driven system for crop selection based on the region's rainfall patterns. As today’s world is based on the internet, our app can provide that benefit to both the farmers and local people to gain knowledge before planting any of the crops for getting a better yield.

Some of the applications of our optimal crop predictor are as follows:

- To give farmers precise and customized guidance, it involves collecting and analyzing rainfall data, climate data, and crop suitability models
- It aids in the informed selection of crop varieties, planting dates, and maximizing output and resource efficiency.
- It helps to assist farmers as well as local people with a digitized platform.

- It helps farmers manage risks associated with water availability and climate variability.
- It can support policymakers in formulating effective agricultural policies.

Overall, the ClimaCrop project has wide-ranging applications that support farmers in making informed decisions about crop selection based on rainfall patterns. It enables climate-resilient agriculture, improves agricultural productivity, and contributes to sustainable resource management in the agricultural sector.

CHAPTER 2: LITERATURE REVIEW

Nepal being an agricultural country, relies most of its income from crops and farming. However, due to the unnatural climate change and irregular rainfall patterns along the Terai region, an effective solution is very important. The primary focus considered while making this project is on the massive impact to the economy, increased crop production, and reduced crop failure.

2.1 CURRENT RESEARCH

2.1.1 Crop Recommendation System to Maximize Crop Yield using Machine Learning Technique

Kumar et al. [2] highlights the significance of crop selection, and factors influencing crop selection are explored, such as production rate, market price, and governmental policy. This study recommends the Crop Selection Method (CSM) as a means of resolving the crop selection issue and raising the crop's net yield rate. While taking weather, soil type, water conditions and crop type into consideration, it gives a variety of crops that can be selected during the course of a season. The estimated value of important parameters affects how accurate CSM is. A prediction method with improved performance and accuracy is therefore needed.

Satish Babu states the requirements and planning necessary for creating a software model for precision farming. [3] It analyzes the basics of precision farming in great detail. The authors start with the basics of precision farming and work toward developing a model that would support it. The Crop Predictor focuses mainly on providing information on what to plant, how to plant, and how to give optimum conditions for maximum production. Furthermore, it also provides information about what sort of plants are best suited for what soil type and climate condition. This study presents a framework incorporating Precision Agriculture (PA) principles at the level of the individual farmer and crop with small, open farms. The overall objective of the plan is to deliver direct encouraging services to even the smallest farmer at the level of his or her smallest plot of crops using the most accessible technology, such as SMS and email. The Kerala State situation, whose normal holding sizes are substantially less than those of the rest of India, inspired the creation of this model. As a result, this model can only be applied in other regions of India after certain modifications.

According to a thesis done at the University of Florida, a crop based AI is classified by the techniques used in machine learning and the algorithm it follows. It may follow convolutional neural networking patterns based on images taken to predict the yield or using recurrent neural networking patterns by providing data in text format [4].

Refining the algorithms and searching techniques can certainly improve the accuracy of the AI significantly. But in the real world the prediction of rainfall, soil condition, natural disasters, crop failures, and animal/insect infestation can become a major hurdle. The user has to be responsible for such random deviations. They can reduce it by utilizing proper fencing and protection of land and crops [5].

2.2 EXISTING SIMILAR SYSTEMS AVAILABLE

2.2.1 Saathi - An AI Innovation For Nepali Farmers

Saathi is an innovation that automatically collects, analyses and communicates data to the farmers. It is meant to provide precise instructions at each stage of planting, from crop selection through maintenance. It gives farmers more power by removing the need to wait for and seek out consultation. Instead, people can obtain help in real time with anything from picking the seeds to sow to keeping track of their health.

2.2.2 Agrisutra

Agrisutra is an integrated agricultural organization that offers services including contract farming, enterprise appraisal, accounting, crop advising, contract farming, crop marketing, farm inspection, crop insurance, and crop marketing. Including seed production, seed trails, and seed breeding. Agrisutra provides B2C services in food and agriculture technology and environmental technology.

2.2.3 Krishi Mitra

Krishi Mitra is a mobile app developed by the National Informatics Center District Unit which can be used by farmers, agricultural officials, agricultural scientists, agricultural supervisors and students of agricultural faculty to increase general information related to agriculture. It offers guidance to farmers on fertilizer use, planting, and crop choices. The system includes a discussion board where farmers may post questions and request advice from experts and other farmers. Krishimitra gathers information on environmental factors including soil moisture, temperature, and rainfall using IoT devices. A machine learning model that can forecast agricultural yields is trained using this data. The model accounts for the resources and desires of the farmer as well as the environmental factors. Farmers then receive the recommendations via the Krishimitra website or mobile app. This helps the interested farmers know more about the plant they are using and also keeps user engagement in the software.

2.3 LIMITATIONS OF EXISTING SYSTEMS

● Lack of technical knowledge: Most of the farmers that rely on manual or primitive ways do not have proper knowledge regarding the crop they are using. Those who are luckily getting high yields are oblivious of the causes for it.

● Accuracy: The accuracy of the app's recommendations may depend on the quality of the data used to train the machine learning model. As a result, if the data is wrong, the model won't be able to generate reliable predictions. The data is inappropriate for Nepal because the majority of existing systems are foreign based.

● Lack of a user-friendly interface: The system's user is not given any specific instructions on how to utilize it or navigate it. Users may find it less appealing to interact with technology when the agriculture sector develops as a result.

With Nepal being an agricultural country, the farmers are not utilizing the modern software to their advantage. Modern farmers still rely on other expert farmers or by trial and error method.

2.4 Solutions provided by our system

By utilizing machine learning algorithms such as random forest algorithms, our proposed System overcomes these flaws by the following ways:

- Our System is a low cost, extremely reliable and modern solution.
- It predicts the yield rate by utilizing different data like rainfall, temperature, humidity, and season.
- It predicts using random forest models using previous census data on rainfall, humidity, and temperature.
- It also provides us information related to the crop of choice of the user.
- It also can also provide information regarding the diseases the plant can suffer and its preventive measures.

CHAPTER 3: METHODOLOGY

3.1 PROCESS MODEL

3.1.1 Incremental Model

Based on the specification of the project and its intended use, we have chosen an Incremental Model for our system development.

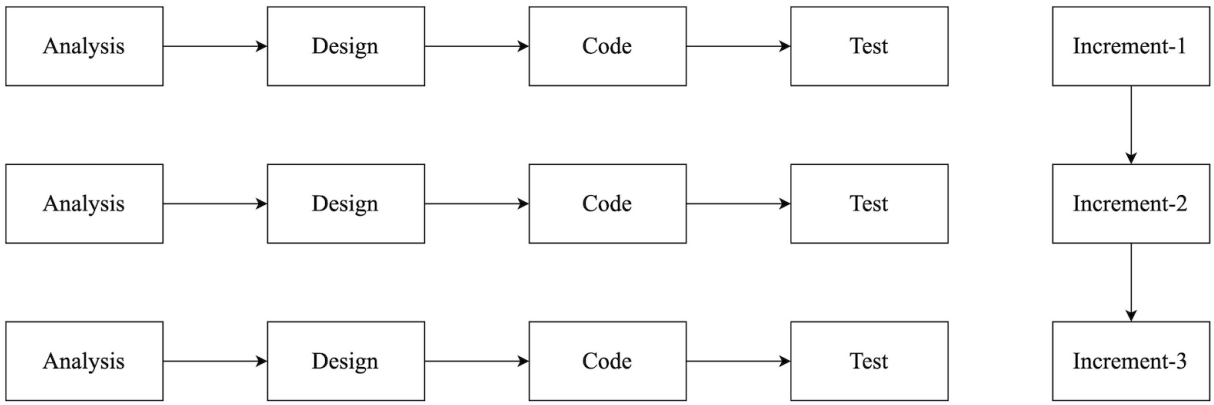


Figure 3.1.1 Block Diagram of Incremental Process Model

The incremental process model is ideal for the project where the software requirements are broken down into many standalone modules in the software development lifecycle. The major goal of our project is to create a system that can predict the best suited crops based on rainfall patterns, soil type, temperature, humidity, and season. The initial stage of the project determines whether the intended requirements are feasible. Once the requirements are determined then incremental development will be carried out in steps covering all the analysis, designing, coding, testing, and implementation.

Various Phases of incremental model:

1. Requirement Analysis

This is the first phase of the incremental model. In this phase, the experts on product analysis identify the product requirements, which include both functional requirements and non-functional requirements; and they also make sure the requirements are compatible. This is a crucial phase when developing software using Incremental models.

2. Design and Development

This is the second phase of the incremental model. In this phase, the development method and system functionality designs must have been successful. The design is then proposed on how to archive and implement this requirement. When the software develops new practicality, the incremental model then uses the development phase and style.

3. Coding

This is the third phase. In this phase, coding is done according to the purpose of the requirements. The coding standards must be followed without any unnecessary hard codes and defaults. This phase also enables the implementation of the designs which are done practically. By completing this phase, the quality of the working product can be upgraded and enhanced.

4. Testing

This is the fourth phase of the incremental phase. In this phase, the performance of each of the existing functions, as well as other additional functionality, are checked. Also, various methods are used to test the various behaviors of each task.

5. Implementation

This is the final phase of the incremental phase. Implementation phase enables the coding phase of the development system. It involves the final coding that design in the designing and development phase and tests the functionality in the testing phase. After completion of this phase, the number of the product working is enhanced and upgraded up to the final system product.

3.2 BLOCK DIAGRAM OF SYSTEM

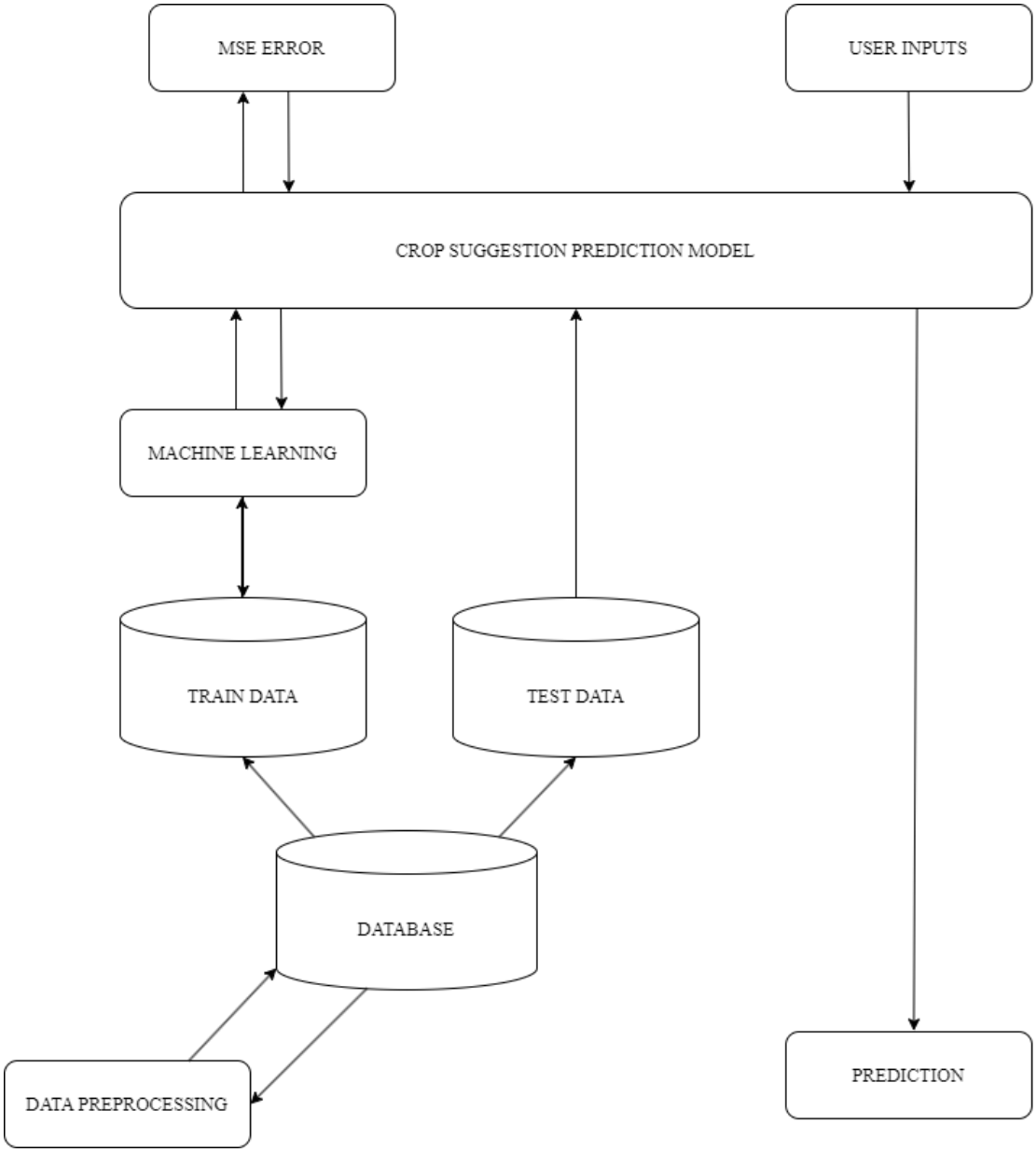


Figure 3.2 Block Diagram of System

The block diagram of the crop prediction system shown in Fig 3.2 offers a general overview of the key components involved in a proposed system.

Access User Interface: The user enters the required information they are looking up. The input data onto an IoT database and is taken to the crop suggestion interface. It is then passed onto a prediction model.

In the machine learning part, the dataset is first pre-processed before model training in order to handle missing values and normalize it. Then, the dataset is split into train data and test data.

The model takes data based on trained as well as tested data to increase its accuracy by reducing the mean squared error or standard deviation. Finally, the optimal crop is then suggested by the model based on the accuracy.

3.3 ALGORITHM

Step 1: Data Preprocessing

Initially, essential libraries including numpy, pandas, seaborn, matplotlib.pyplot, and warnings are imported. The data is read from a CSV file into a pandas DataFrame, followed by the removal of duplicate rows. Non-numeric values in the 'average_rain_fall_mm_per_year' column are eliminated, and a statistical description of the dataset is generated. Visualization techniques such as scatterplots and line plots are employed to explore relationships between variables over time.

Step 2: Data Analysis and Visualization

Various visualizations, including scatterplots and line plots, are created to examine the impacts of different factors like pesticides, rainfall, and temperature on crop yield. Patterns and trends within the data are analyzed, particularly focusing on the effects of these variables on crop productivity.

Step 3: Feature Engineering and Train-Test Split

Feature engineering involves preparing the data for model training. This includes dropping unnecessary columns and splitting the dataset into features (X) and the target variable (y). A train-test split is performed to facilitate model evaluation.

Step 4: Data Transformation

Categorical variables are encoded using one-hot encoding, while numerical features are scaled using standardization and min-max scaling techniques.

Step 5: Model Training and Evaluation

Multiple regression models including Linear Regression, Lasso Regression, Ridge Regression, K-Nearest Neighbors (KNN), Decision Tree Regressor (DTR), and Random Forest Regressor are trained on the dataset. These models are evaluated using Mean Absolute Error (MAE) and R-squared (R2) score to assess their performance.

Step 6: Model Selection

The best performing model is selected based on the evaluation metrics obtained during model training and evaluation.

Step 7: Prediction System

A prediction function is created to make predictions using the selected model. This function is tested with sample input values to demonstrate its functionality.

Step 8: Conclusion

The process concludes by identifying the crop with the highest predicted yield based on the provided input values.

3.3.1 Random Forest Algorithm

An ensemble learning system called random forest mixes various decision trees to produce a forecast. Using a random selection of feature subsets and training each tree on a separate subset of data, this approach generates numerous decision trees. The final prediction is then created by merging all of the trees' predictions. The robust and adaptable random forest technique excels at both classification and regression issues.

To create the random forest, N decision trees are joined initially. Then, predictions are generated for each of the trees that were created in the first step.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

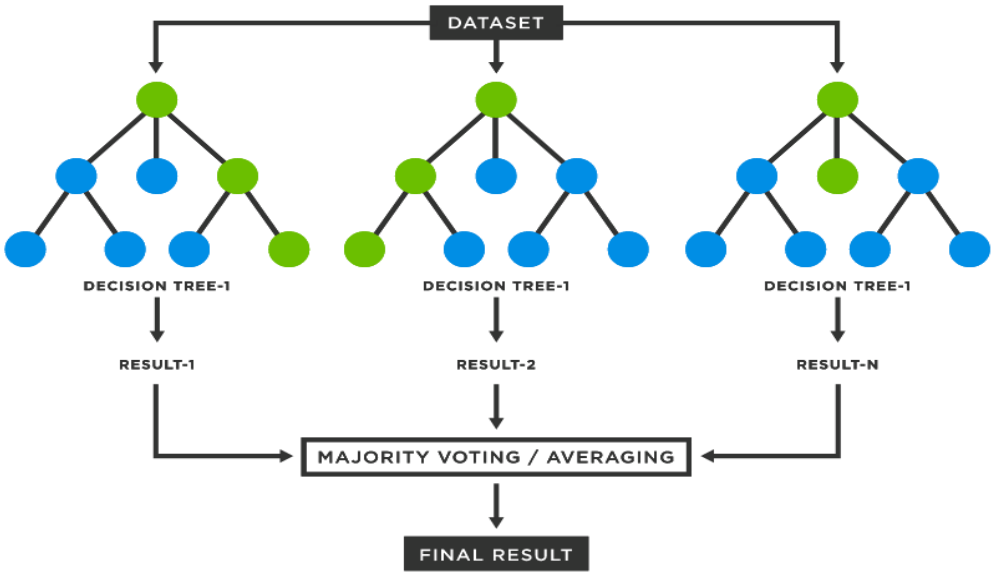


Figure 3.3.1 Simple Random Forest Network

3.3.2 Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

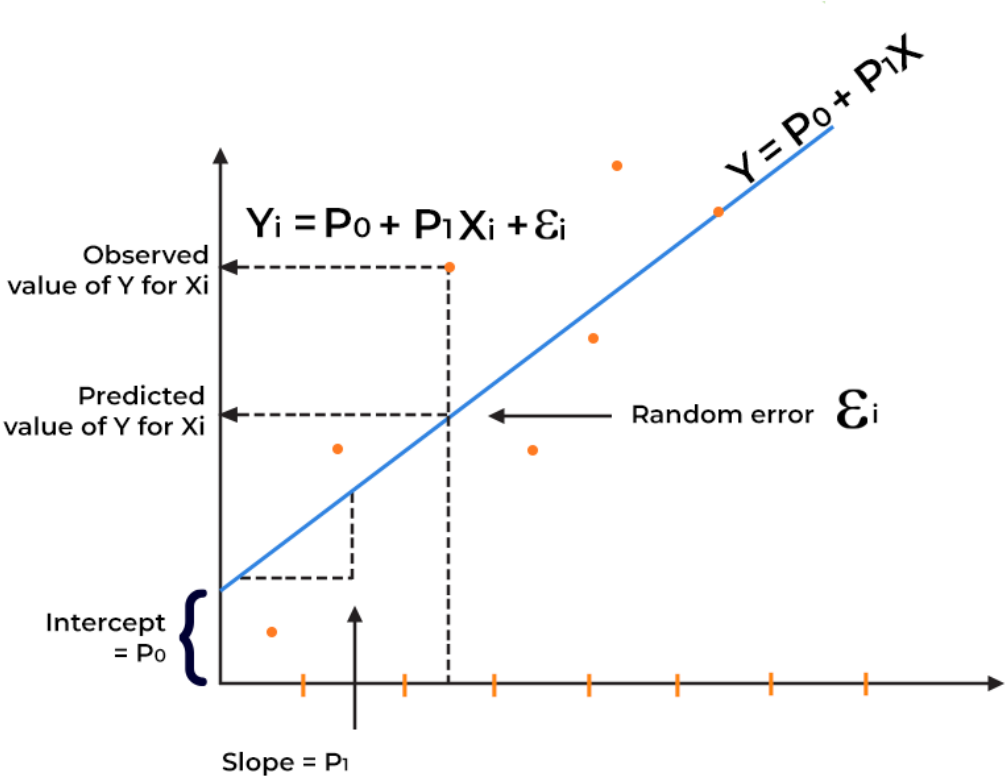


Figure 3.3.2 Linear Regression

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple other linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data.

3.3.3 K-Neighbor Regressor

K-Neighbor Regressor is a non-parametric supervised learning method which stores all the available data and classifies a new data point based on the similarity.

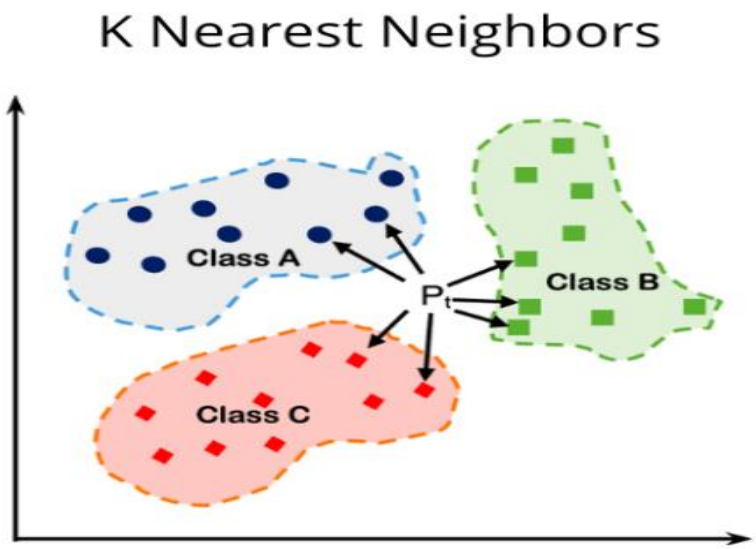


Figure 3.3.3 K-Neighbour Regressor

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

3.3.4 Lasso Regression

"LASSO" stands for Least Absolute Shrinkage and selection Operator.Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

Lasso solutions are quadratic programming problems, which are best solved with software. The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\ \beta_j)2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Which is the same as minimizing the sum of squares with constraint Σ |Bj≤ s (Σ = summation notation). Some of the βs are shrunk to exactly zero, resulting in a regression model that's easier to interpret

3.3.5 Neural Network

A neural network is a machine learning program, or model, that makes decisions in a manner similar to the human brain, by using processes that mimic the way biological neurons work together to identify phenomena, weigh options and arrive at conclusions. Every neural network consists of layers of nodes, or artificial neurons—an input layer, one or more hidden layers, and an output layer. Each node connects to others, and has its own associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Neural networks rely on training data to learn and improve their accuracy over time. Once they are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts.

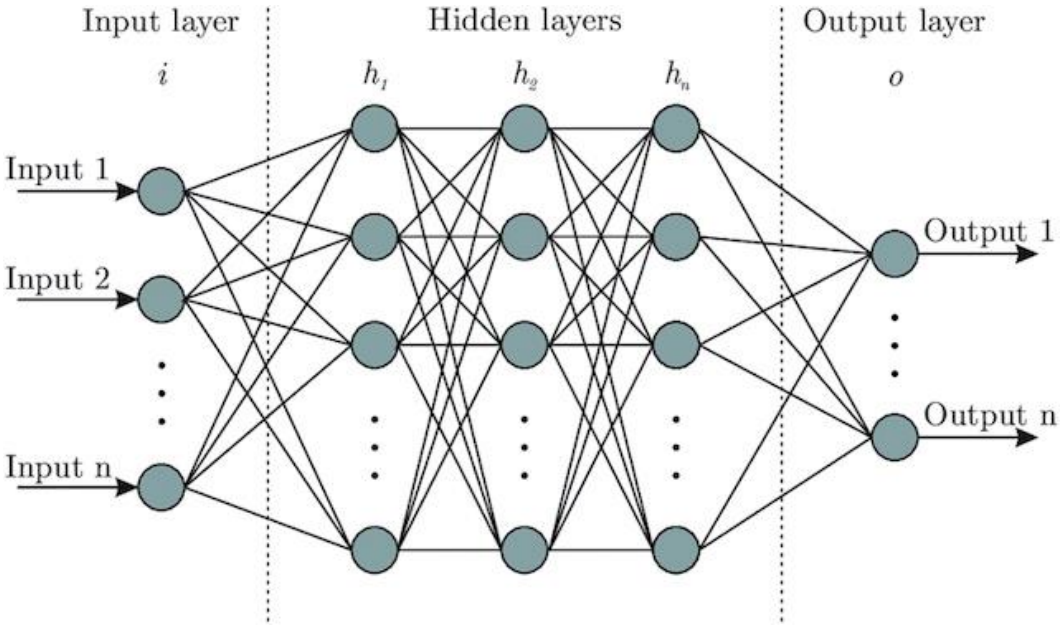


Figure 3.3.5 Neural Networking

3.3.6 Ridge Regressor

Ridge regression—also known as L2 regularization—is one of several types of regularization for linear regression models. Regularization is a statistical method to reduce errors caused by over fitting on training data. Ridge regression specifically corrects for multicollinearity in regression analysis. This is useful when developing machine learning models that have a large number of parameters, particularly if those parameters also have high weights. While this article focuses on regularization of linear regression models, note that ridge regression may also be applied in logistic regression.

L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will not result in sparse models.

A tuning parameter (λ) controls the strength of the penalty term. When λ = 0, ridge regression equals least squares regression. If λ = ∞, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

OLS regression uses the following formula to estimate coefficients:

$$\underline{\hat{B}}=(X'X)^{-1}X'\underline{Y}$$

If X is a centered and scaled matrix, the cross product matrix (X`X) is nearly singular when the X-columns are highly correlated. Ridge regression adds a ridge parameter (k), of the identity matrix to the cross product matrix, forming a new matrix (X`X + kI). It's called ridge regression because the diagonal of ones in the correlation matrix can be described as a ridge. The new formula is used to find the coefficients:

$$\underline{\tilde{B}}=(X'X+kI)^{-1}X'\underline{Y}$$

3.3.7 Decision Tree Regression

Decision Trees (DTs) are a non-parametric supervised learning method used to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf node.

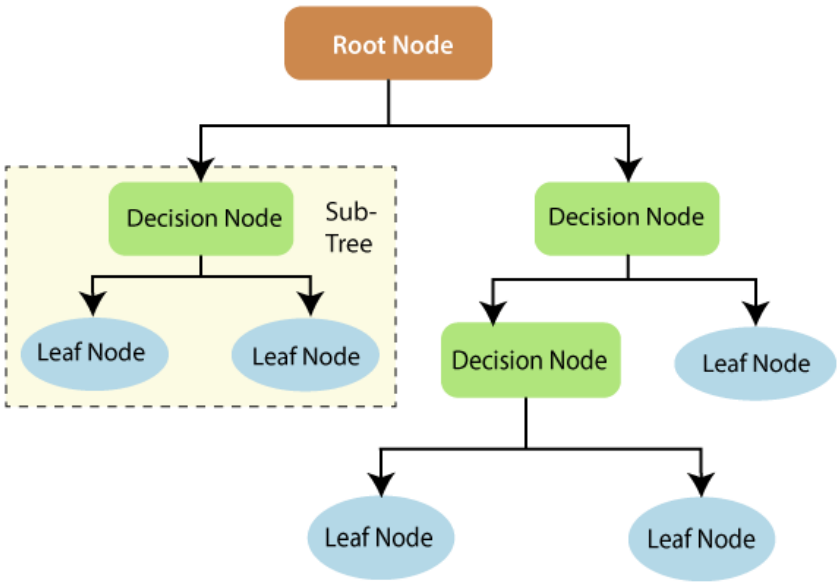


Figure 3.3.7 Decision Tree Regression

3.4 Flowchart

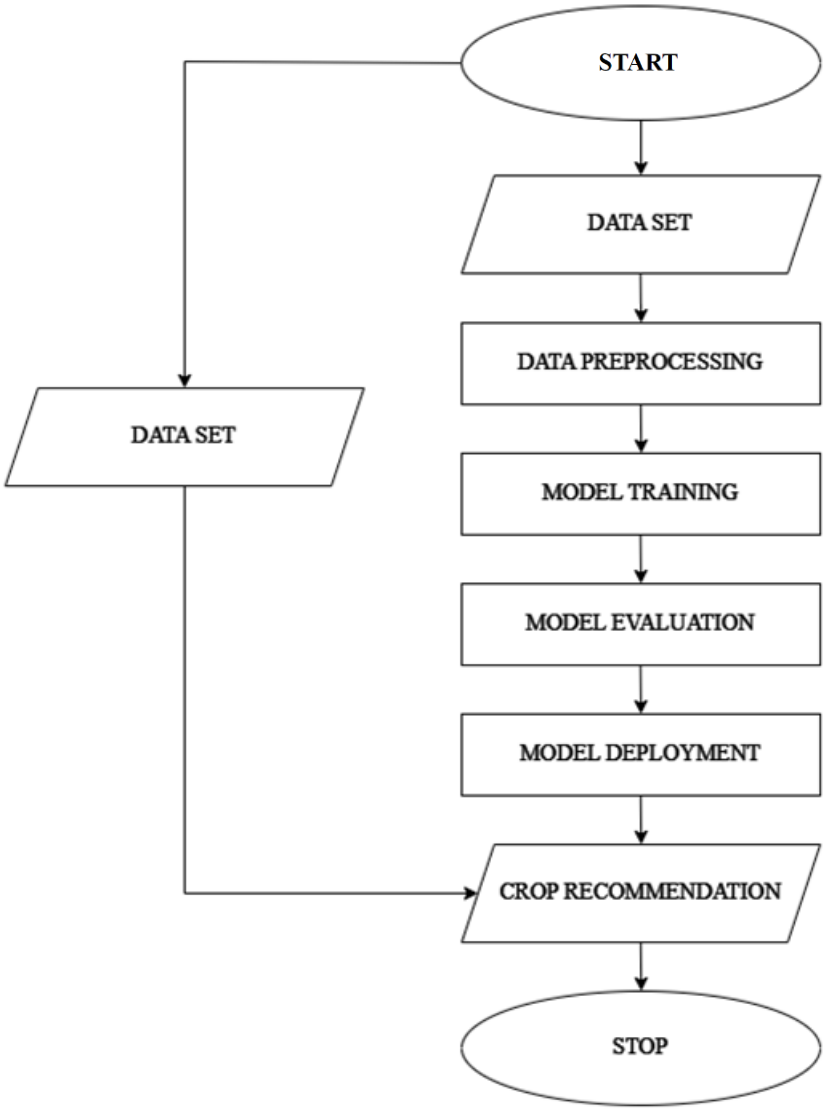


Fig 3.4 Flowchart

3.5 Use Case Diagram

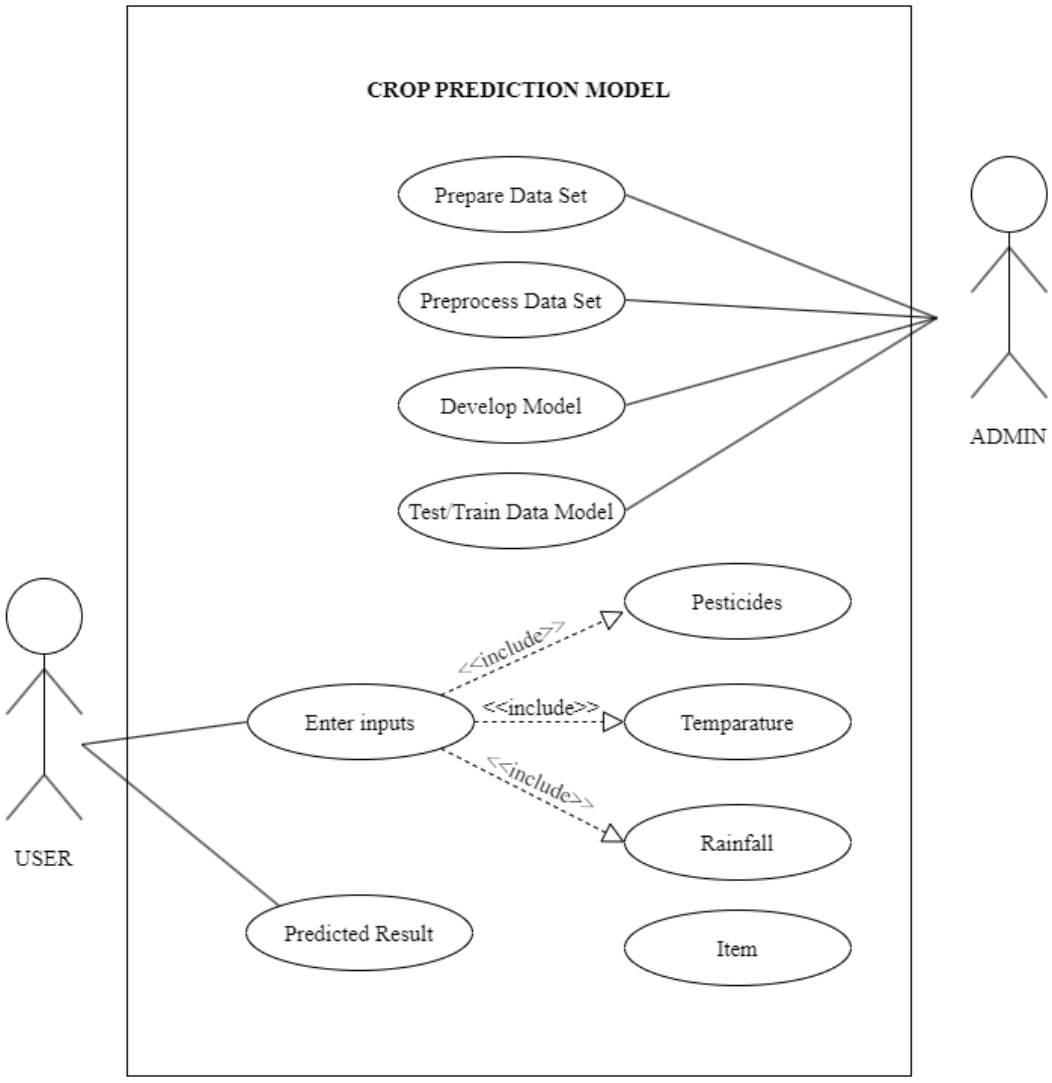


Fig 3.5 Use Case Diagram

3.6 Sequence Diagram

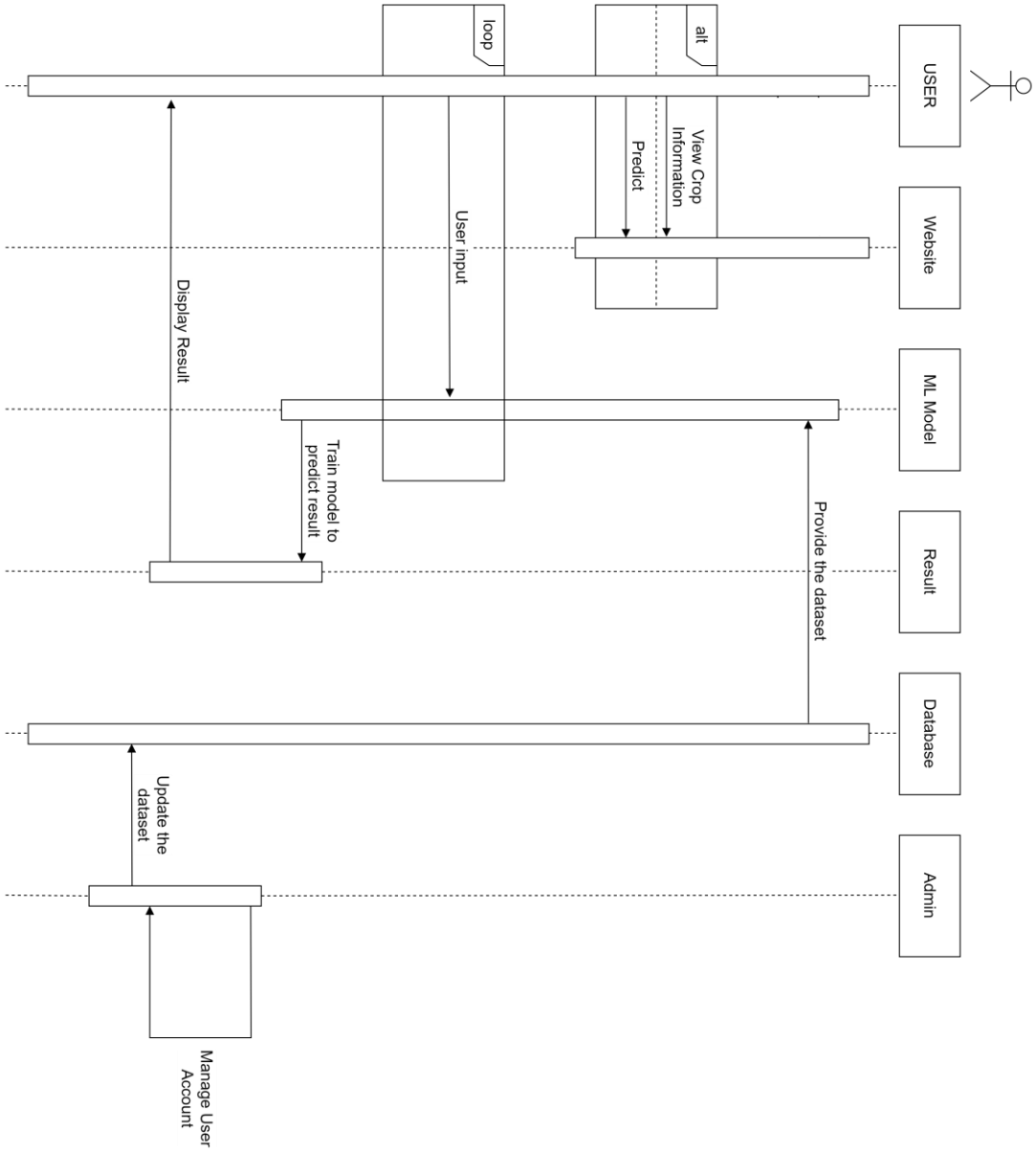


Fig 3.6 Sequence Diagram

3.7 Data Flow Diagram

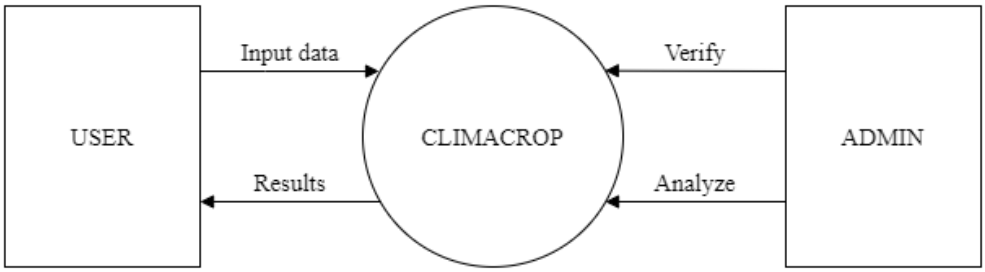


Fig 3.7.1 DFD Level 0

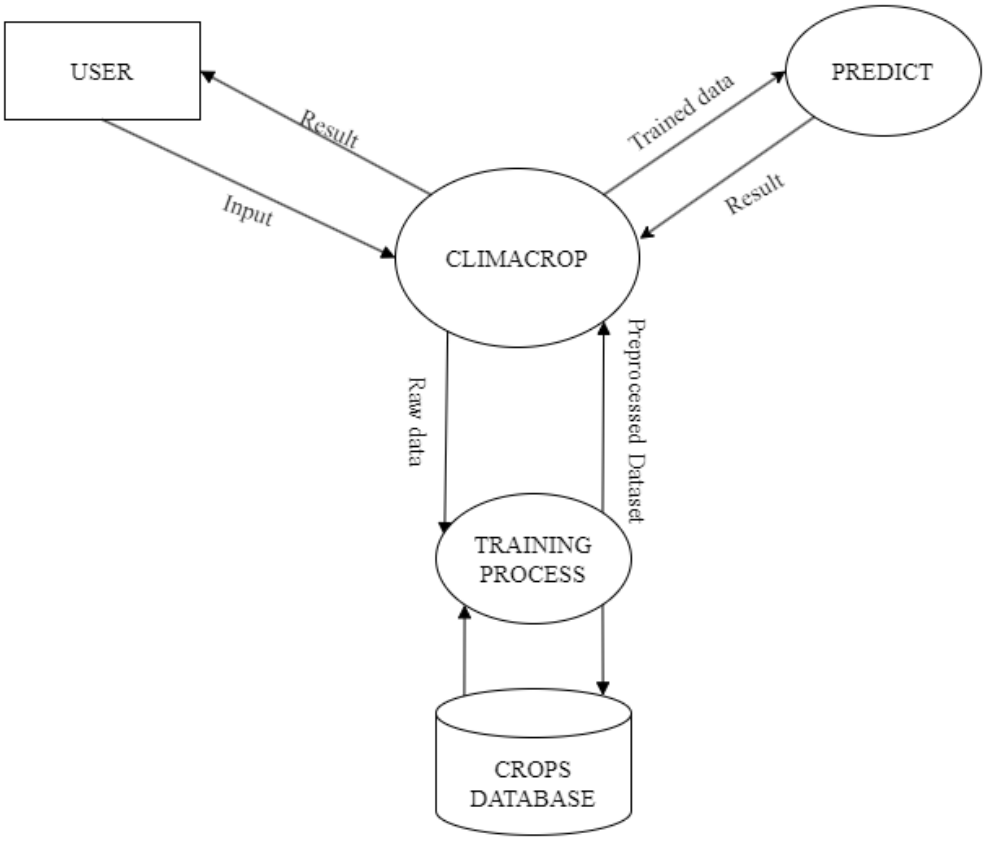


Fig 3.7.2 DFD Level 1

3.8 TOOLS USED

3.8.1 Python

Python is a high-level, interpreted, object oriented, general-purpose programming language. It was developed by Guido Van Rossum and released in 1991. It enables structured (particularly computational), object-oriented, and functional programming on multiple levels. Python is created under an open-source license that has been accepted by OSI, making it freely distributable and useable, even for commercial purposes.

3.8.2 JavaScript (JS)

JavaScript, often abbreviated as JS, is a lightweight, interpreted, or just-in-time compiled programming language with first class functions. JS was created at Netscape Communications by Brendan Eich in 1995. With JavaScript users can build modern web applications to interact directly without reloading the page every time.

3.8.3 Git

Git is a distributed version control system that tracks changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development.

3.8.4 Python libraries

A Python library is a collection of related modules it has collections of code that can be utilized in many program. For the programmer, it simplifies and makes Python programming more practical. Since we don't have to develop the same code for many program repeatedly. In the domains of machine learning, data science, data visualization, etc., python libraries are quite important.

3.8.5 Jupyter

Jupyter Notebook is an open-source interactive web application for creating and sharing live code, equations, visualizations, and narrative text. It supports various programming languages, including Python, R, and Julia, allowing users to develop and present data-driven analyses in a seamless and flexible environment. Jupyter Notebooks are widely

used in data science, research, and education for their collaborative and interactive features.

3.8.6 HTML

HTML (Hypertext Markup Language) is the standard markup language for creating and structuring web pages. It utilizes tags to define elements such as headings, paragraphs, links, and images, organizing content for browsers. HTML provides the foundation for web development, enabling the creation of interactive and visually appealing websites.

3.8.7 Flask

Flask is a lightweight and versatile Python web framework, ideal for building web applications and APIs. It provides essential components, allowing developers to choose and integrate specific tools as needed. Flask emphasizes simplicity and ease of use, making it popular for beginners and seasoned developers alike. Its modular structure and extensive documentation facilitate rapid development, making Flask a preferred choice for projects of varying complexities in the Python ecosystem.

3.8.8 CSS

CSS (Cascading Style Sheets) is a fundamental web technology used to style and format HTML documents. It controls the visual presentation of a webpage by defining aspects like layout, colors, fonts, and spacing. Using selectors and properties, CSS enables developers to create visually appealing and responsive designs, ensuring a consistent and polished user experience across different devices. As a backbone of web development, CSS plays a crucial role in separating content from presentation, enhancing the flexibility and maintainability of websites.

3.8.9 Kaggle

Kaggle is a powerful platform for data science competitions and collaborative machine learning projects. Acquired by Google, it hosts a vast community of data scientists, researchers, and enthusiasts. Kaggle provides datasets, notebooks, and a competition platform, fostering knowledge sharing and innovation. Participants can tackle real-world challenges, showcase their skills, and learn from diverse solutions. It has become

a go-to hub for the global data science community, promoting collaboration and advancing the field through shared insights and solutions.

3.8.10 Visual Crossing Weather API

Visual Crossing Weather API is a service providing comprehensive weather data for global locations. It offers current conditions, forecasts, historical weather, and climate data through RESTful endpoints. A RESTful endpoint represents a specific resource and its associated data, and can be manipulated using the standard HTTP methods. Users can access weather information with ease, enabling applications, businesses, and researchers to integrate accurate and reliable weather data into their projects.

3.9 VERIFICATION AND VALIDATION

We applied various machine learning algorithms namely linear, lasso, ridge, KNN Decision Tree Regression, and Random Forest Algorithm on our data set and used the predicted values against actual values to plot the following graph.

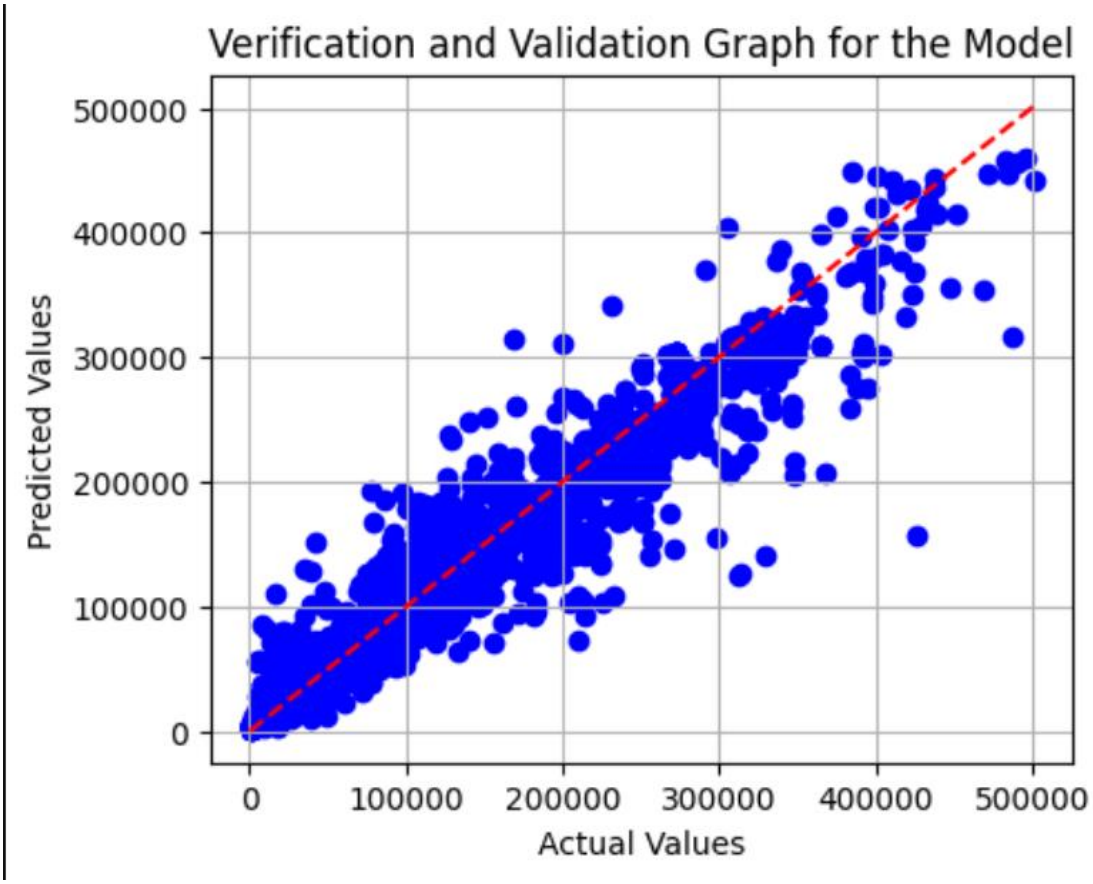


Fig 3.9 Verification and validation Graph

Finally, after training the model using random forest regressor as our chosen model we finally evaluated its performance using the following metrics.

1. Mean Absolute Error (MAE): The Mean Absolute Error (MAE) is a metric that measures the average absolute difference between the actual and predicted values of a target variable. It is calculated by taking the average of the absolute differences between each actual and predicted value. MAE provides a simple and intuitive measure of how close the predictions are to the actual values, with a lower value indicating better performance. MAE is less sensitive to outliers compared to Mean Squared Error (MSE).

2. Root Mean Squared Error (RMSE): The Root Mean Squared Error (RMSE) is a metric that measures the average magnitude of the difference between the actual and predicted values of a target variable. It is calculated by taking the square root of the average of the squared difference between each actual and predicted value. RMSE and MSE is similar to MAE, but it penalizes larger errors more heavily than smaller errors, making it more sensitive to outliers. A lower RMSE indicates better performance.

CHAPTER 4: EPILOGUE

4.1 Result and Conclusion

We have successfully implemented the crop yield prediction system that utilizes ML algorithms trained on historical data to forecast crop yields.
For the part of accuracy comparison of the model, the accuracy with different algorithm are as follows:

Algorithm	Accuracy
Lasso	63.286% with MSE of 32251
Ridge	63.289% with MSE of 32216
KNN	93.213% with MSE of 10437
Decision Tree Regression	85.732% with MSE of 12062
Random Forest	94.76% with MSE of 9539
Neural Network with 500 epochs	65% with MAE of 10000

Thus, out of all algorithms taken into implementation, Random Forest algorithm was found to be best suited algorithm for our project to predict the yield of crops on hg/ha (hectogram/hectare).

4.2 Future Enhancement

We aim to implement the following enhancements in future:

- Implementation of IoT to get far more accurate and real time variables (humidity, temperature, moisture etc).
- User feedback feature can be added.

REFERENCES

[1] Lal Prasad Amgain, Shailesh Adhikari, Samiksha Pandit
Multi-year Prediction of Rice Yield under the Changing Climatic Scenarios in Nepal Central Terai Using DSSAT Crop Model

[2] R.Kumar, M.P. Singh, P. Kumar and J.P. Singh "Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM).

[3] Satish Babu (2013), "A Software Model for Precision Agriculture for Small and Marginal Farmers", at the International Centre for Free and Open Source Software (ICFOSS) Trivandrum, India.

[4] Clyde Fraisse, Yiannis Ampatzidis, Sandra Guzman,Wonsuk Lee Artificial Intelligence for Crop Yield Forecasting

[5] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, Stephano Ermon
Combining Remote sensing data and machine learning to predict crop yield

[6] Dorugade and D. N. Kashid "Alternative Method for Choosing Ridge Parameter for Regression."

[7] Casper Hansen "Neural networks from scratch."

[8] Sharban Kumar Apat, Jyotirmaya Mishra, Dr. Neelamadhab Padhy, Srujan Kotagiri Raju "An Artificial Intelligence-based Crop Recommendation System using Machine Learning"

[9] Male Sowjayna, Motupalli Vinay, Vidya Malepati, Makkena Vishnu Vardhan "Developing the crop yields using Artificial Intelligence and Computer Vision"

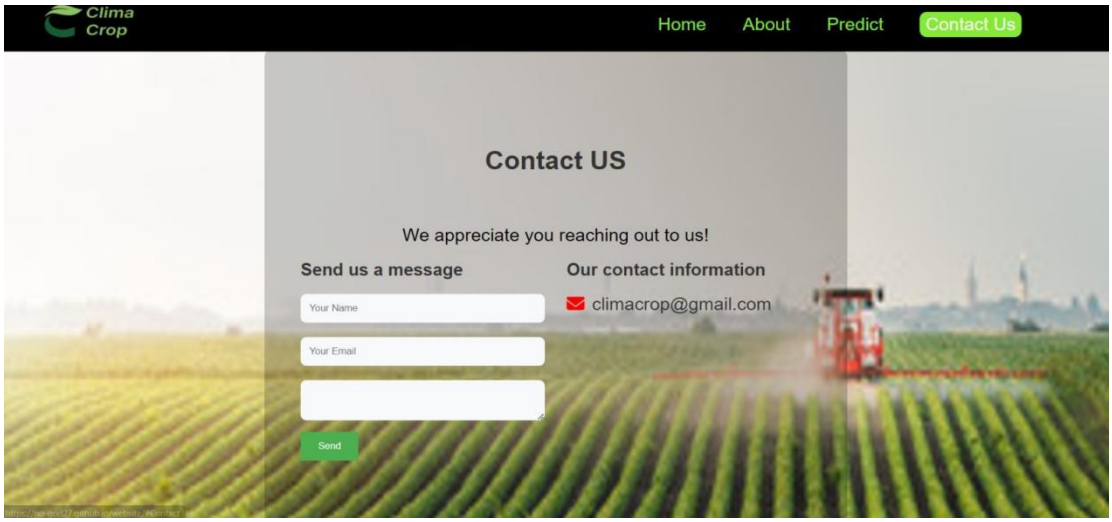
BIBLIOGRAPHY

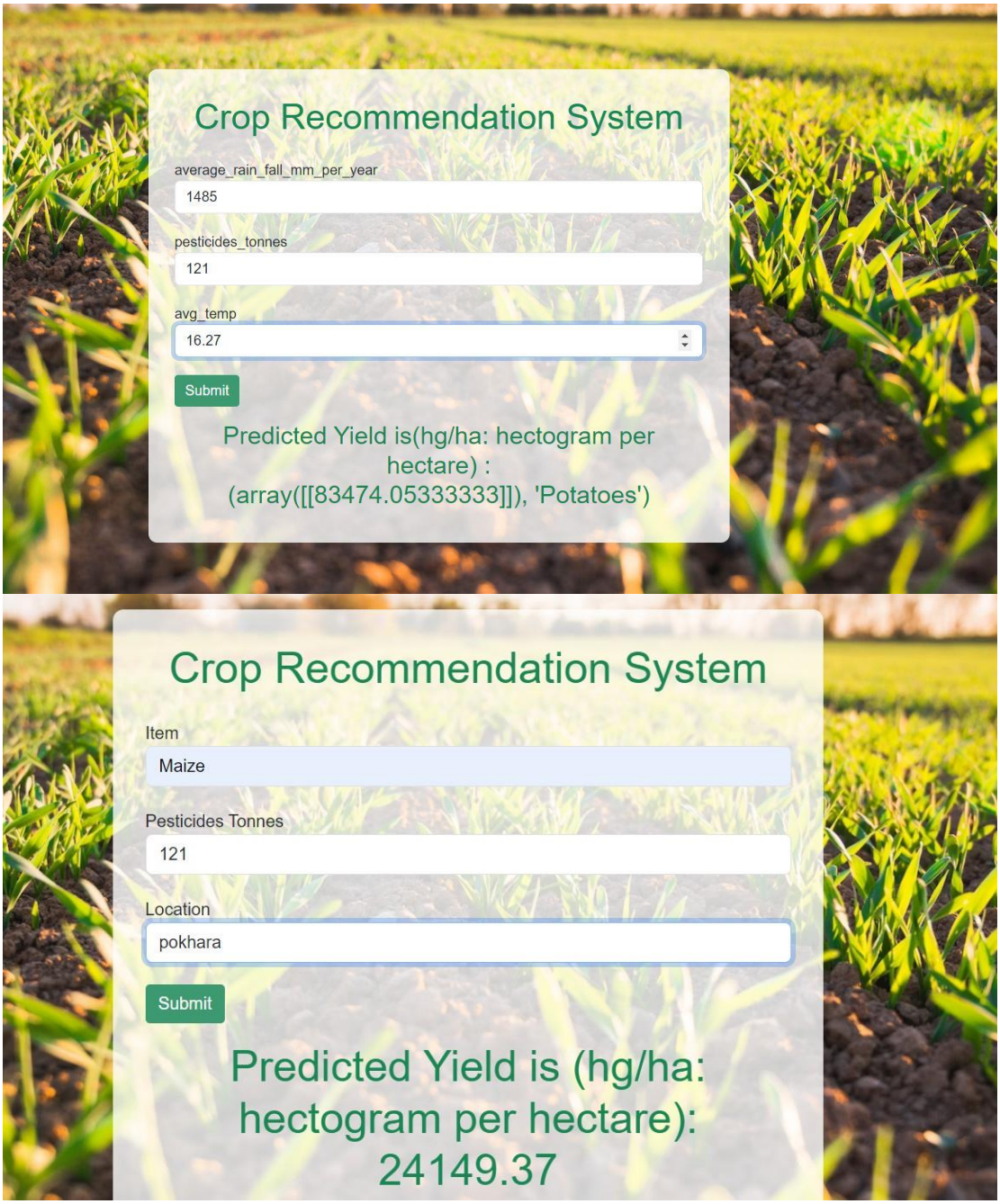
[1] Danilo P. Mandiac, Jonathon A. Chambers., “Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability”. Wiley 2001

[2] Olivier Rebiére, Cristina Rebiére., Mastering the Gantt chart: Understand and use the “Gantt Project” open source software efficiently, Guide Education, 2017

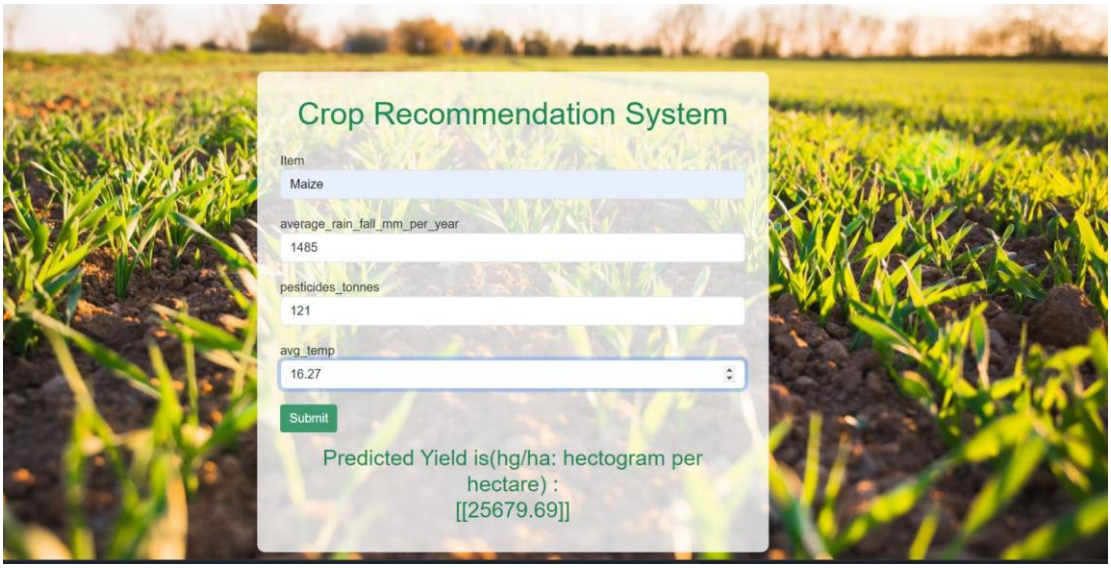
[3] Scott Hartshorn., Machine Learning with Random Forests and Decision Tress: A Visual Guide For Beginners

SCREENSHOTS




Crop Recommendation System
average_rain_fall_mm_per_year
1485
pesticides_tonnes
121
avg_temp
16.27
Submit
Predicted Yield is(hg/ha: hectogram per hectare) :
(array([[83474.05333333]]), 'Potatoes')

Crop Recommendation System
Item
Maize
Pesticides Tonnes
121
Location
pokhara
Submit
Predicted Yield is (hg/ha: hectogram per hectare):
24149.37



Selecting the most accurate model

```
RF = RandomForestRegressor()
RF.fit(X_train_dummy,y_train)
RF.predict(X_test_dummy)
```

```
array([ 67582.10666667,  15094.03      ,  61628.5       , ...,
        12228.1       ,   9102.68      , 162955.88      ])
```

Creating a prediciton system

```
def prediction(Item,average_rain_fall_mm_per_year,pesticides_tonnes,avg_temp):
    features = np.array([[Item,average_rain_fall_mm_per_year,pesticides_tonnes,avg_temp]])
    transformed_features = preprocessor.transform(features)
    predicted_value = RF.predict(transformed_features).reshape(1,-1)
    return predicted_value[0]
```

```
models = {
    'Linear regression': LinearRegression(),
    'Lasso': Lasso(),
    'Ridge': Ridge(),
    'KNN' : KNeighborsRegressor(),
    'DTR' : DecisionTreeRegressor(),
    'Random Forest' : RandomForestRegressor()

}
```

```
for name,model in models.items():
    model.fit(X_train_dummy,y_train)
    y_pred = model.predict(X_test_dummy)
    print(f"{name} have MSE : {mean_absolute_error(y_test,y_pred)} and Score : {r2_score(y_test,y_pred)}")
```

```
Linear regression have MSE : 9289353205079444.0 and Score : -3.2797854697708637e+22
C:\Users\linpa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\linear_model\_coordinate_descent.py:628: ConvergenceWarning: Objective d
id not converge. You might want to increase the number of iterations, check the scale of the features or consider increasing regularisation. Duality gap:
2.972e+11, tolerance: 1.482e+10
  model = cd_fast.enet_coordinate_descent(
Lasso have MSE : 32251.95742643754 and Score : 0.632866538574377
Ridge have MSE : 32216.517838301952 and Score : 0.6328968805005877
KNN have MSE : 10437.674185463658 and Score : 0.932135547373045
DTR have MSE : 11685.721740248056 and Score : 0.8808601038342525
Random Forest have MSE : 9651.295153184396 and Score : 0.9467695181523696
```

21046	Nepal	Soybeans	1991	5964	1500	60.11	14.94
21047	Nepal	Wheat	1991	14103	1500	60.11	14.94
21048	Nepal	Maize	1992	16647	1500	60.11	14.95
21049	Nepal	Potatoes	1992	85916	1500	60.11	14.95
21050	Nepal	Rice, padd	1992	20481	1500	60.11	14.95
21051	Nepal	Soybeans	1992	5805	1500	60.11	14.95
21052	Nepal	Wheat	1992	13338	1500	60.11	14.95
21053	Nepal	Maize	1993	15984	1500	60.11	15.16
21054	Nepal	Potatoes	1993	84268	1500	60.11	15.16
21055	Nepal	Rice, padd	1993	24100	1500	60.11	15.16
21056	Nepal	Soybeans	1993	5811	1500	60.11	15.16
21057	Nepal	Wheat	1993	12460	1500	60.11	15.16
21058	Nepal	Maize	1994	16502	1500	60.11	15.3
21059	Nepal	Potatoes	1994	87539	1500	60.11	15.3
21060	Nepal	Rice, padd	1994	21237	1500	60.11	15.3
21061	Nepal	Soybeans	1994	5855	1500	60.11	15.3
21062	Nepal	Wheat	1994	14704	1500	60.11	15.3
21063	Nepal	Maize	1995	16447	1500	60.11	15.3
21064	Nepal	Potatoes	1995	85926	1500	60.11	15.3
21065	Nepal	Rice, padd	1995	23910	1500	60.11	15.3
21066	Nepal	Soybeans	1995	6540	1500	60.11	15.3
21067	Nepal	Wheat	1995	14416	1500	60.11	15.3
21068	Nepal	Maize	1996	16770	1500	70.43	15.37
21069	Nepal	Potatoes	1996	84750	1500	70.43	15.37
21070	Nepal	Rice, padd	1996	24554	1500	70.43	15.37
21071	Nepal	Soybeans	1996	6601	1500	70.43	15.37