ENSEMBLE MACHINE LEARNING MODEL FOR BETTER CROP PRODUCTION

PROJECT REPORT

In partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING
(Maulana Abul Kalam Azad University of Technology,
Formerly known as West Bengal University of Technology)

Submitted by

NAME	ROLL NO.
Arya Bose	33200120041
Aditya Ghosh	33200120042
Akash Samanta	33200120065
Aditya Ray	33200120056

Under the guidance of

Ms. Tanushree Chakraborty
Assistant Professor, Department of Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
TECHNO INTERNATIONAL BATANAGAR
Maheshtala, Kolkata – 700141, West Bengal, India

IN LOVING MEMORY OF

Archisman Samanta

"For as long as there is a memory of them,
they live on in our hearts."

TECHNO INTERNATIONAL BATANAGAR
MAHESHTALA, KOLKATA, PIN: 700141, WEST BENGAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
2023-2024

CERTIFICATE

Certified that the project work entitled ***ENSEMBLE MACHINE LEARNING MODEL FOR BETTER CROP PRODUCTION*** *is a bona fide work carried out by*

NAME	ROLL NO
Arya Bose	33200120041
Aditya Ghosh	33200120042
Akash Samanta	33200120065
Aditya Ray	33200120056

In partial fulfilment for the award for degree of ***BACHELOR OF TECHNOLOGY*** *in* ***COMPUTER SCIENCE AND ENGINEERING*** *of the* ***MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY****, formerly known as* ***WEST BENGAL UNIVERSITY OF TECHNOLOGY, KOLKATA*** *during the year 2023-2024. It is certified that all corrections / suggestions indicated for Internal Assessment has been incorporated in the report deposited in the Department. The project report has been approved as it satisfies the academic requirements in respect of Project Work prescribed for Bachelor of Technology Degree.*

________________________	________________________	________________________
Ms.Tanushree Chakraborty	**Mr.Subhankar Guha**	**Prof.(Dr.) Ratikanto Sahoo**
Project Guide	HOD, CSE	Director

Name of the Student:
University Roll No:
Name of Examiner: Signature with Date

1.

2.

ACKNOWLEDGEMENT

I am greatly indebted to our project Guide **Ms. Tanushree Chakraborty, Assistant Professor of Computer Science and Engineering Department** for providing us all possible help and support while doing this project. Without her guidance the project would not get such a progress.

I also express my sincere thanks to **Mr. Subhankar Guha, Head of the Department, Computer Science and Engineering,** whose ceaseless cooperation made it possible in completion of this project for extending all possible help.

Finally, I would like to thank **PROF.(Dr.) Ratikanto Sahoo, Director, Techno International Batanagar** for his colossal support and encouragement.

Lastly, I would like to extend our thanks to all respected Faculty Members and friends directly or indirectly associated with our project, who contributed their personal level best which enabled us for a successful completion of the project.

ABSTRACT

Agriculture is the most precious factor for food supply in the world. It is also responsible for supplying the raw materials for other industries. The growth in agriculture cannot cope with the world's population which is increasing day by day. Besides increasing food production for a developing country that has limited land and resources, can cause a shortage of food in the near future. Selecting the right crop for a specific region is crucial for enhancing its production. Historical data is used to accurately anticipate the area's agricultural yield based on measures of soil components (Nitrogen, Phosphorus, Potassium) and climatic measures (Temperature, Humidity, Rainfall, pH). Then we propose ensemble machine learning approaches called 'Naive Forest', 'AdaBayes', 'AdaForest' to effectively predict the crops. After investigating four classification machine learning algorithms (Decision Tree, Gaussian Naïve Bayes, Support Vector Machine (SVM), Logistic Regression), regression machine learning algorithm (Support Vector Regressor (SVR), Gradient Boosting Regression, Extreme Gradient Boosting Regression, Random Forest Regression) and ensemble learning (Random Forest, AdaBoost) algorithms and their respective accuracy, these 'Naive Forest', 'AdaBayes', 'AdaForest' approach are designed. We believe that the proposed Models will help farmers and personnel in the agriculture sector in proper crop production.

TABLE OF CONTENTS

TOPIC **PAGE NO.**

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form	Abbreviation	Full Form
AI	Artificial Intelligence	CR	Cognitive Radio
ML	Machine Learning	SR	Super Resolution
KRR	K-Nearest Neighbour Random Forest Ridge Regression	LGBM	Light Gradient Boosting Machine
SVR	Support Vector Regression	ANN	Artificial Neural Network
NB	Gaussian Naïve Bayes	NN	Neural Network
RR	Ridge Regression	IoT	Internet of Things
RF	Random Forest	RFR	Random Forest Regression
CB	CatBoost	FN	Fuzzy Network
KNN	K-Nearest Neighbour	MCM	Marckov Chain Model
SVM	Support Vector Machine	KMC	K-Means Clustering
DT	Decision Tree	CSM	Classifier System with Memory
XGB	Extreme Gradient Boosting Regression	ADB	AdaBoost
LR	Logistic Regression	MSE	Mean Square Error
GBR	Gradient Boosting Regression	RMSE	Root Mean Square Error
KR	Kernel Ridge	MAE	Mean Absolute Error
BT	Baggage Tree		

1. Overview

1.1 Introduction:

Agriculture is an essential aspect of human life, providing food, fibre, and other resources necessary for survival. Crop production is dependent on various factors such as climatic conditions, soil nutrients, and other environmental factors. Over the years, climate change has become a major concern for agricultural productivity, as it significantly affects crop production by altering the availability of key environmental factors such as temperature, precipitation, humidity, and light. Soil nutrient availability is another key factor that influences crop growth and yield. Therefore, understanding the impact of climatic conditions and soil nutrients on crop production is crucial for sustainable agriculture.

1.2 Research Problem:

The research problem is to predict crop production based on climatic conditions and soil nutrients. This problem arises due to the challenges faced by agricultural practitioners in optimizing crop yield, especially in the context of climate change and soil nutrient deficiencies. Predicting crop production based on these factors can help farmers optimize their cultivation practices, enhance food security, and improve their economic conditions.

1.3 Research Questions:

The research questions that this study aims to answer are:

- What is the impact of climatic conditions on crop production?
- How do soil nutrients influence crop production?
- Can we develop an ensemble model for crop production based on climatic conditions and soil nutrient data?
- How accurate and useful is the ensemble model in practical applications?

1.4 Research Objectives:

The objective of our paper is to achieve the following:

1. To assess the performance of various individual machine learning models, such as Random Forest, Logistic Regression, Decision Tree, Naive Bayes, and Support Vector Machine, AdaBoost, Support vector regressor (SVR), Gradient boosting regression, Random Forest Regression and Extreme gradient boost regression in predicting best crop yields based on soil components (Nitrogen, Phosphorus, Potassium) and climatic measures (Temperature, Humidity, Rainfall, pH).
2. To propose an ensemble model to demonstrate the effectiveness of ensemble machine learning techniques in improving the accuracy and reliability of crop yield predictions.
3. To investigate how ensemble models can effectively address uncertainties arising from variations in climate and soil components data and crop yield responses.
4. Compare the precision and reliability of ensemble models with the traditional agricultural prediction techniques.

1.5 Motivation of the Study:

The motivation behind this research paper is to prevent food shortages and develop agricultural import-export strategies, focusing on effective prediction for government use in short and long-term policies. The study will focus on predicting the crop production of crops based on climatic conditions and soil nutrient data.

1.6 Significance of the Study:

This study's significance lies in its potential to contribute to sustainable agriculture by providing a tool for predicting crop production based on key environmental factors. The predictive model developed in this study can help farmers optimize their crop management practices, reduce crop failures, and improve their economic conditions. The study's findings can also help policymakers and agricultural practitioners develop strategies to mitigate the impact of climate change on crop production.

2. Literature survey

In 2023, M. Hasan et al. [8] proposed an ensemble machine learning approach named K-nearest Neighbour Random Forest Ridge Regression (KRR), to predict crops, especially rice, potato, and wheat. The DM test validates the robustness of the proposed model, showing 1% and 5% significance compared to other benchmark ML models, indicating KRR's superior performance.

In 2022, Sanjay M. D et al. [9] used regression techniques such as Lasso, Kernel Ridge, and ENet algorithms to forecast agricultural production.

In 2021, Priyadharshini et.al. [10] stated that in this paper, precision agriculture is implemented, utilizing modern agricultural technology and advancements, particularly in underdeveloped nations, to optimize crop management tailored to specific locations. The crops suitable for the underlying soil series were recommended using five distinct algorithms: Naive Bayes, Adaboost, Bagged Tree, Support Vector Machine, and Artificial Neural Network were utilized.

In 2017, E. Manjula et al. [11] suggested predicting crop yield based on previous data by applying association rule mining to agricultural data, focusing on creating a prediction model for future crop yield estimation targeting the district of Tamil Nadu in India.

In 2016, Govindarajan Muthukumarasamy et al. [2] proposed two ensemble models, namely AdaSVM and AdaNaive, which are then compared to the SVM and Naive Bayes methods which outperform SVM and Naive Bayes, making them more suitable for the analysed dataset.

Again, Gour Hari Santra et al. [1] evaluated the innovative techniques namely Decision Tree, Artificial neural networks, Regression Analysis, Information Fuzzy Network, and Bayesian belief network. The study employs various statistical techniques such as time series analysis, Markov chain model, k-means clustering, k nearest neighbour, and support vector machine to identify significant relationships in the database.

In 2014, Utkarsha P. Narkhede et al. [6] proposed an evaluation of a modified K-Means clustering algorithm for crop prediction against the traditional K-Means and K-Means++ clustering algorithms. The modified k-Means algorithm has been shown to achieve the highest number of high-quality clusters, accurate crop predictions, and maximum accuracy count.

In 2014, S. Veenadhari et al. [12] proposed a user-friendly web page software tool called ’Crop Advisor’ using the C4.5 algorithm, which is employed to identify the most influential climatic parameter on the crop yields of selected crops in specific districts of Madhya Pradesh.

Table 1: Comparison of the related works on crop production prediction

Technique	Year	Dataset	Remarks
KRR, SVR, NB, RR, RF, and CB [8]	2023	Collected data samples from different agricultural organizations in Bangladesh	The DM test is conducted showing 1% and 5% significance of KRR compared to the benchmark ML models.
KNN, RF, SVM, DT, XGB, LR [18]	2023	Data collected from five districts of Issyk-Kul region.	XGB : 98.86%, RF : 99.09%, LR : 56.59%, SVM : 34.09%, DT : 67.04%, KNN : 65.68%
GBR, XGB, RF, SVR [19]	2022	Data collected from Sugar Industry and Cane Development Department of Uttar Pradesh.	The results of GBR and XGB were very close, and GBR closely outperforms XGB. RF performed fairly better than SVR. SVR was the weakest.
KR, Lasso and Enet [9]	2022	The data comprises details such as State, District, Crop, Season, Year, Area, and Product.	Predicts yield for all sorts of crops grown in India.
LGBM, GBR, XGB, Ridge, CR, SR [21]	2022	Dataset comprises data from an open-source computer simulation model and meteorological information collected from Maine, USA, and the Canadian Maritimes.	Highest R^2 was achieved by SR indicating that the points of output are linear, almost being a perfect performance.
NB, RF [7]	2021	Collected data from the field area using Sensors	The model has an accuracy of 6.89%.

		Arduino Microcontroller and iOT system.	
NB, AdaBoost, BT, SVM, ANN [10]	2021	Collected data from government websites and Kaggle.	“NN = 89.88% accuracy and LR (2) = 88.26% accuracy.”
SVM, KNN, BT [4]	2020	Collected data from some research papers.	Predicting crops based on soil types.
RF [3]	2020	Collected data from different sources and self-generated datasets	The accuracy of predictions is above 75% in all the crops.
RF [13]	2018	Collected data from Indian Govt. historical data	Predicting yield of the crop from available historical and available data.
Association Rule Mining [14]	2017	Collected data from Tamil Nadu Govt.	For predicting all crops, the accuracy of 85%.
IoT enabled Arduino sensors, NN [15]	2017	Collected data from Arduino Soil moisture sensor.	Detect and specify the level of soil moisture.
RFR, GBR, XGB [20]	2017	Kaggle AMS 2013-14 Solar Energy Prediction Contest	No clear overall winner emerges.
Various AI Models [5]	2016	Collected data using both RGB and BGNIR sensors at an altitude of 50 meters.	Comparison between In-situ Yield Data and Estimated Yield Data.
AdaNaive, AdaSVM [2]	2016	The crop production data was obtained from faostat3.fao.org.	Used for Time Series Forecasting.
NN, FN, DT, Regression Analysis, Time Series Analysis, MCM,	2016	Collecting data from different sources.	A comparison study about the performance of ANN’s with exponential smoothing and ARIMA models.

KMC, KNN, and SVM [1]			
CSM [16]	2015	Data is gathered from farmers of Patna district, Bihar (India).	CSM method retrieves all possible crops that are to be sown at a given time stamp.
KMC, DBSCAN, OPTICS, COBWEB, Agglomerative and Divisive [12]	2015	FAO data set (soya bean).	K-means is the most effective method for the given dataset since it takes the least time to complete.
C4.5 [11]	2014	Data are gathered from selected districts of Madhya Pradesh in which the selected crop area is maximum.	The prediction accuracy varied from 76 to 90 %. The overall prediction accuracy of the developed model is 82.00 %.
Modified K-Means [6]	2014	The dataset is provided by the Department of Agriculture Maharashtra.	The modified K-Means clustering algorithm achieved the correct prediction of crop and maximum accuracy count.
SVM [17]	2009	The weather data of Cambridge University for a period of five years (2003- 2007)	In the case of MLP, the Mean Square Error ranges from 8.07 to 10.2, but in the case of SVM, it is in the range of 7.07 to 7.56.
DT, NB, SVM, LR, RF, AdaBoost; SVR, GBR, XGBoost, RFR; Naive Forest, AdaBayes, AdaForest (Ours proposed models)	2024	Data collected from Kaggle	The accuracy of the ensembled ‘Naive Forest’ model is 99.18%, ‘AdaBayes’ model is 99.39%, & ‘AdaForest’ model is 99.09%.

3. Workflow Diagram

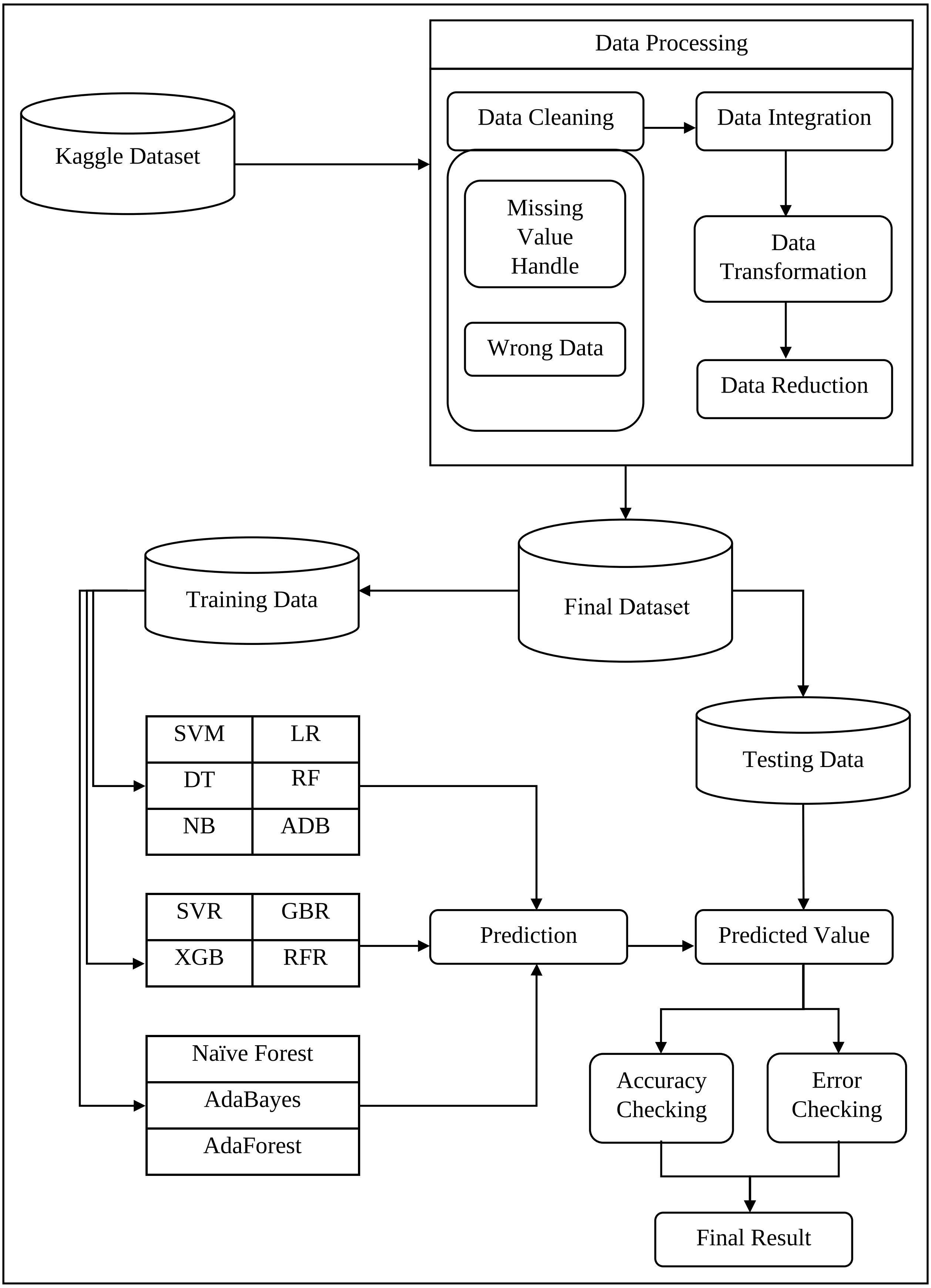


Figure 1: Workflow Diagram

4. Methodology

The methodology encompasses several key steps that collectively contribute to the advancement of agricultural forecasting and sustainable practices. By amalgamating the strengths of multiple individual models, ensemble methods offer a powerful tool to navigate the complexities of climate-crop relationships. It can address the complexities of crop production and enhance predictive accuracy. This section outlines the systematic process to develop and assess an ensemble machine learning model specifically designed to improve crop production predictions.

The proposed ensemble model is illustrated in Figure 1. Here we start with the data collection from Kaggle in the first step. Then after pre-processing, prepare the suitable data for training and testing our ensemble model. After the train-test split, evaluation, and analysis are performed giving us an accurate comparison of our proposed ensemble model with other ML models.

4.1. Data Collection Methods:

In our dataset, the data samples include 2200 records of crops having 22 crops specified namely rice, maize, chickpea, kidney beans, pigeon- peas, moth beans, black gram, mungbean, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee, having 100 samples each. We prepare our dataset with 7 different attributes namely N(Nitrogen), P(Phosphorus), K(Potassium), Temperature, Humidity, pH, and Rainfall.

4.2. Data Analysis Methods:

The data analysis methods would involve analysing the collected data using statistical techniques such as correlation analysis, regression analysis, and machine learning algorithms. These techniques would help the researcher determine the relationships between climatic conditions, soil nutrients, and crop production. For Analysing the data, we are using the following algorithms:

4.2.1. Decision Tree:

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

4.2.2. Gaussian Naïve Bayes:

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

4.2.3. Support Vector Machine (SVM):

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

4.2.4. Logistic Regression:

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

4.2.5. Random Forest:

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4.2.6. AdaBoost:

- Adaptive Boosting, or AdaBoost, is a statistical classification meta-algorithm that is used to enhance performance when combined with numerous other learning algorithm types. The final output of the boosted classifier is represented by a weighted sum that is created by combining the output of the other learning algorithms, or "weak learners."
- AdaBoost is adaptive in that it adjusts weaker learners in the future to take into account cases when prior classifiers misclassified the data. Compared to other learning algorithms, it may be less prone to the overfitting issue in specific situations.

- It has been demonstrated that AdaBoost can successfully integrate strong base learners, like deep decision trees, as well as weak base learners, like decision stumps, to create an even more accurate model.

4.2.7. Support Vector Regressor (SVR):

- One kind of support vector machine (SVM) utilized for regression problems is support vector regression (SVR). It looks for a function that, given an input value, most accurately predicts the continuous output value.
- Support Vector Machines (SVMs) are employed in classification problems, Support Vector Regression (SVR) looks for a hyperplane in a continuous space that best matches the data points. The process involves projecting the input variables onto a high-dimensional feature space, then identifying the hyperplane that minimizes the prediction error and maximizes the margin, or distance, between the hyperplane and the nearest data points.
- Utilizing a kernel function to translate the data into a higher-dimensional space, SVR is able to manage non-linear interactions between the input variables and the goal variable.

4.2.8. Gradient Boosting Regression:

- A machine learning ensemble technique called gradient boosting successively combines the predictions of several weak learners, usually decision trees. By refining the model's weights based on the mistakes of earlier iterations, it seeks to enhance overall predictive performance by progressively lowering prediction errors and raising the model's accuracy.
- Gradient Boosting Regressor is used when the goal column is continuous, while Gradient Boosting Classifier is used when the problem is one of classification.

4.2.9. Extreme Gradient Boosting Regression:

- Extreme gradient boosting is a machine learning technique that is a part of the gradient boosting framework, which is a subset of ensemble learning. It makes use of regularization techniques to improve model generalization using decision trees as foundation learners.

- Extreme gradient boost is a popular choice for computationally demanding tasks including regression, classification, and ranking due to its proficiency in feature importance analysis, management of missing information, and computational economy.

4.2.10. Random Forest Regression:

- Random Forest Regression is an ensemble strategy that uses numerous decision trees along with a technique called Bootstrap and Aggregation, or bagging, to solve both regression and classification tasks.
- The fundamental idea here is to use a combination of decision trees instead of depending only on one to determine the final result.
- We can use variables like mean squared error, root mean square error, R^2, adjusted R^2, and more to examine a regression model's performance measures. I'll be concentrating on mean squared error and root mean square error in this piece.

4.3. Ensemble Model:

Ensemble modelling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread. This improves the accuracy of predictive analytics and data mining applications. Ensemble modelling techniques are commonly used in machine learning applications to improve the overall predictive performance.

Analytical and machine learning models process some inputs and identify patterns in those inputs. Then, based on those patterns, they produce some output (outcome), which is usually some kind of prediction.

In many applications, one model is not enough to produce accurate predictions (i.e., predictions with low generalization error). To improve the prediction process and deliver better predictive output, multiple models are combined and trained. This approach is known as ensemble modelling.

4.3.1. Naïve Forest:

A voting classifier is a machine learning model that gains experience by training on a collection of several models and forecasts an output (class) based on the class with the highest likelihood of becoming the output. To forecast the output class based on the largest majority of votes, it averages the results of each classifier provided into the voting classifier. The concept is to build a single model that learns from various models and predicts output based on their aggregate majority of votes for each output class, rather than building separate specialized models and determining the accuracy for each of them**.**

By using this voting classifier technique, we ensemble Naïve Bayes and Random Forest models. These models are chosen on the basics of accuracy. Thus, Naïve Forest model is mead.

4.3.2. AdaBayes:

Boosting is an ensemble modelling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued, and models are added until either the complete training data set is predicted correctly, or the maximum number of models are added.

By using the Boosting technique, AdaBoost and Naïve Bayes models are added. This approach helps to boost the prediction accuracy of the Naïve Bayes model. Thus, AdaBayes model is made.

4.3.3. AdaForest:

AdaBoost and Random Forest models are added to the ensemble approach with the help of the previously mentioned Boosting technique. This method aids in improving the Random Forest model's prediction accuracy. The result is the AdaForest model.

4.4. Learning and Evaluation:

It is the method of splitting a given dataset into training and testing datasets using the train test split() method of the scikit learn module. The following 2200 data is divided as 70-30% split thereby having 1540 data in training and 660 for testing the Models.

The dataset is divided into training and testing sets, and models like Random Forest, Logistic Regression, Decision Tree, Naïve Bayes, and Support Vector Machine are trained and evaluated for their respective performance.

4.5. Limitations of the Study:

We would acknowledge the limitations of the study, such as the potential for confounding variables, sampling bias, and measurement error. We would also suggest areas for future research to address the limitations of the current study.

5. Technical Platform

Our project is implemented using Python Programming language. In this project, we have imported certain packages namely,

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Sklearn
- Pickle

This entire computation is being performed over Google Colab and Kaggle platform. Besides, Streamlit is used to create an interactive web application interface along with Visual Studio Code to predict crop using the ensemble models separately.

6. Result Analysis

The study aimed to predict crop production based on climatic conditions and soil nutrients. The research design followed a quantitative approach. The study used statistical analysis and machine learning algorithms to build an ensemble model for crop production. The results indicated that the climatic conditions, including temperature, precipitation, humidity, and other factors, along with soil nutrients such as nitrogen, phosphorus, and potassium, have a significant impact on crop production.

6.1. Model Accuracy Comparison:

Table 2: Comparison of the accuracy scores of the used classification and ensemble models

Model Name	**Accuracy (%)**
Decision Tree	85.91
Naïve Bayes	99.24
Support Vactor Machine(SVM)	98.33
Logistic Regression	96.21
Random Forest	98.94
AdaBoost	95.30
Naïve Forest	99.09
AdaBayes	99.39
AdaForest	99.09

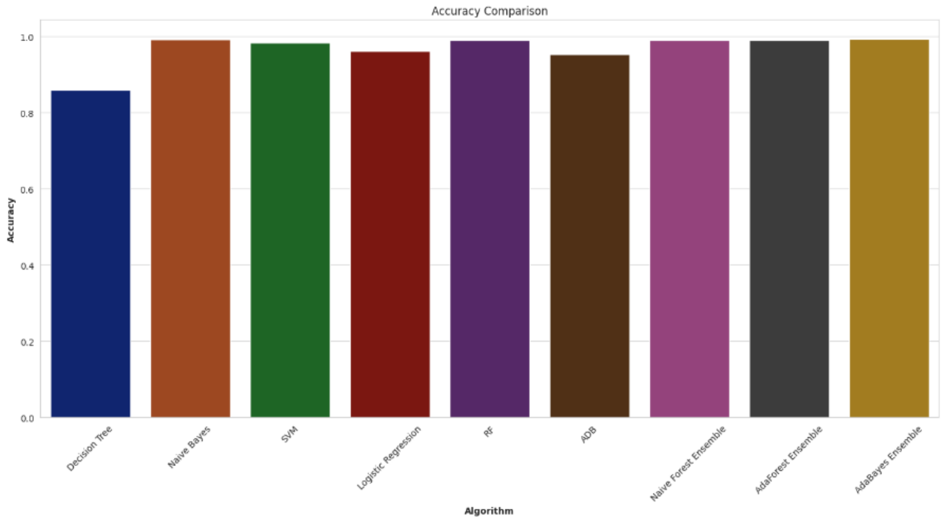


Figure 2: Accuracy Comparison

The accuracy of our ensemble models with the other ML models is presented in Figure 2. The proposed ensembled ’Naive Forest’ model performs with a predicted accuracy of 99.09% (Logistic Regression having 96.21%, Decision Tree with 85.91%, SVM with 98.33%, etc.). Random forest is itself an ensemble model used for both classification and regression combining multiple decision trees to enhance the model’s prediction. Whereas Naïve Bayes is used for multi-class prediction problems taking less training data. Also, AdaBoost is an ensemble learning technique used to improve the predictive accuracy of any given model by combining multiple ‘weak’ learners. It provides the accuracy of 95.30% and after adding AdaBoost with Naïve Biyes and Random Forest it also boosts their accuracy 99.24% to 99.39% and 98.94% to 99.09%.

6.2. Model Error Comparison:

Table 3: Comparison of the error scores of the used regression models

Model	MSE (Mean Square Error)	RMSE (Root Mean Square Error)	MAE (Mean Absolute Error)	R^2 Score
Support Vector Regression	10.95	3.31	2.20	0.71
Gradient Boosting Regression	25.81	5.08	3.89	0.31
Extreme Gradient Boosting Regression	5.83	2.41	1.69	0.84
Random Forest Regression	1.21	1.10	0.34	0.97

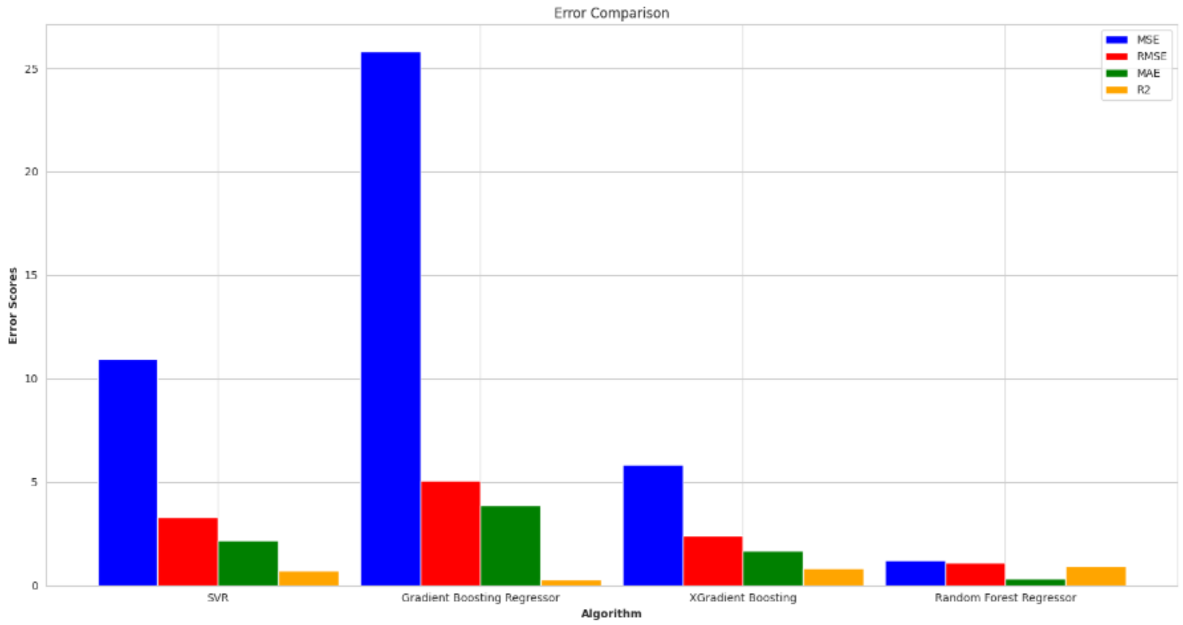


Figure 3: Error Comparison

Here we compare the performance of four regression models—Support Vector Regression (SVR), Gradient Boosting Regression (GBR), Extreme Gradient Boosting Regression (XGBR), and Random Forest Regression (RFR)—using several metrics: Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. Among these models, Random Forest Regression demonstrates the best performance, with the lowest MSE (1.21), RMSE (1.10), and MAE (0.34), and the highest R^2 Score (0.97), indicating superior accuracy and fit to the data. Extreme Gradient Boosting Regression also performs well, with relatively low error metrics (MSE of 5.83, RMSE of 2.41, MAE of 1.69) and a high R^2 Score (0.84). Support Vector Regression shows moderate performance, with an R^2 Score of 0.71 and moderate error metrics (MSE of 10.95, RMSE of 3.31, MAE of 2.20). In contrast, Gradient Boosting Regression has the highest error values (MSE of 25.81, RMSE of 5.08, MAE of 3.89) and the lowest R^2 Score (0.31), suggesting it is the least effective model for this data set. Overall, Random Forest Regression is the most effective model, followed by Extreme Gradient Boosting Regression, Support Vector Regression, and Gradient Boosting Regression.

6.3. Data Visualization:

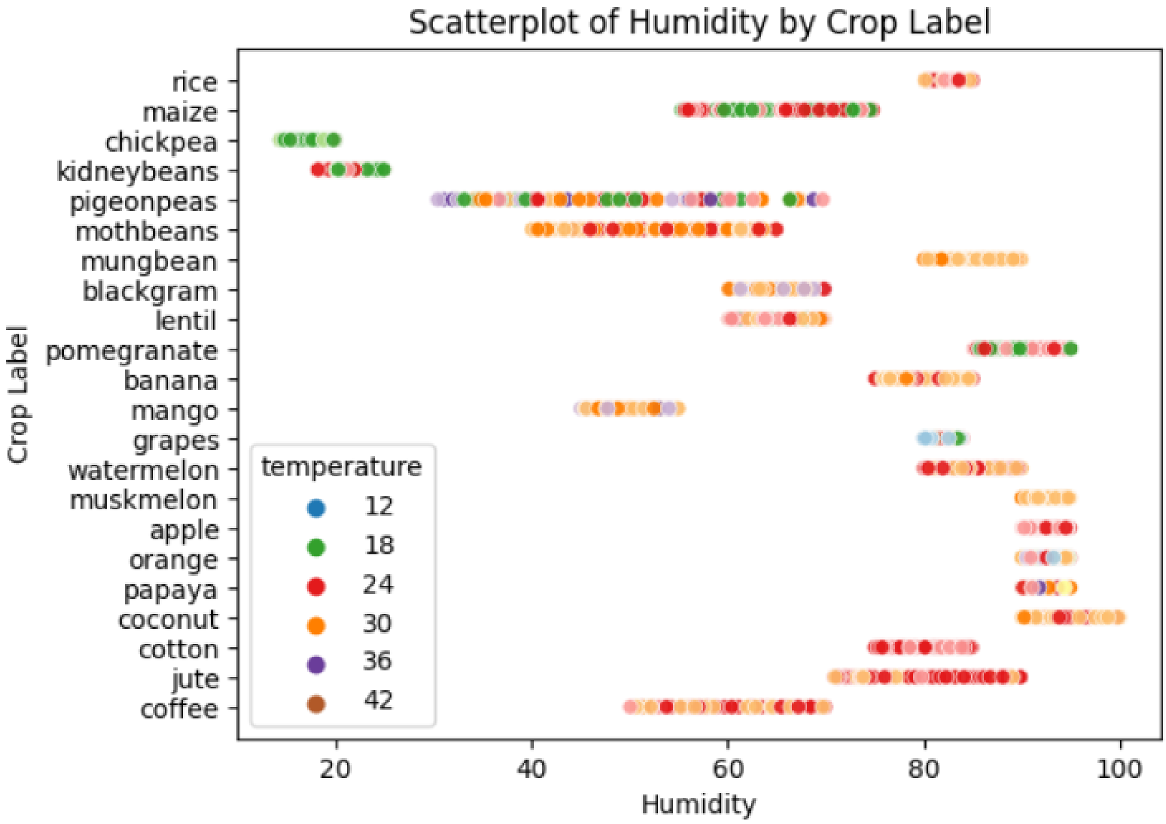


Figure 4: Scatterplot (Crops vs Humidity, Hue = Temperature)

This scatterplot illustrates the humidity preferences and temperature ranges for various crops. Rice prefers very high humidity (80-100%), while crops like Maize, Chickpea, and Kidneybeans favour moderate to high humidity (40-80%). Pomegranate, Banana, Mango, and Grapes prefer higher humidity (60-80%). Temperature-wise, Rice, Pomegranate, Banana, Mango, and Grapes grow in warmer climates (indicated by red and orange dots), while Maize and Chickpea can tolerate a broader temperature range. Coffee shows a distinct preference for high humidity around 90%. This information is valuable for agricultural planning in different climatic regions.

‘

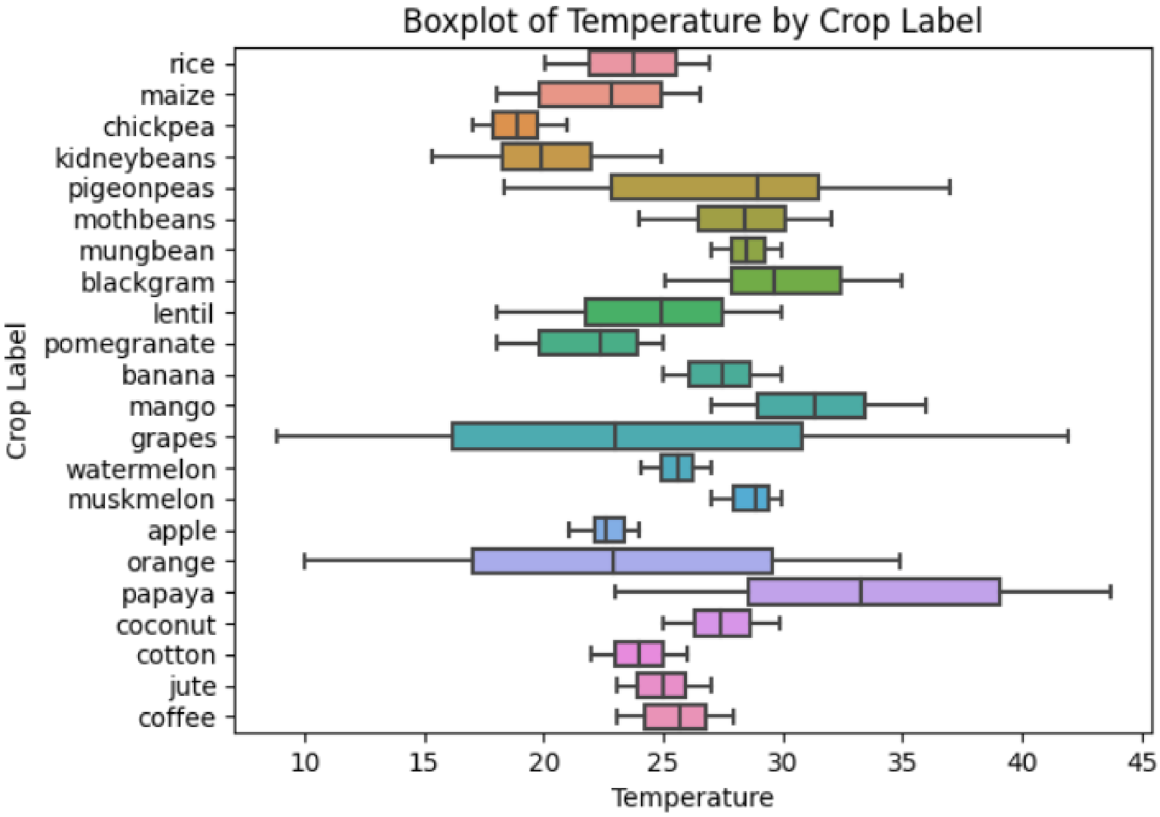


Figure 5: Boxplot (Crops vs Temperature)

This boxplot shows the temperature ranges preferred by various crops. Rice and Maize thrive in warmer temperatures (around 25-35°C). Chickpea, Kidneybeans, Pigeonpeas, Mothbeans, and Mungbean prefer slightly cooler ranges (20-30°C). Pomegranate, Banana, Mango, and Grapes favour a broad range from 20-35°C. Watermelon and Muskmelon have narrow preferences around 25°C. Apple, Orange, Papaya, Coconut, Cotton, Jute, and Coffee show a wider range, with Coffee and Cotton favouring higher temperatures (25-40°C). This highlights each crop's optimal growing temperature, aiding in agricultural planning.

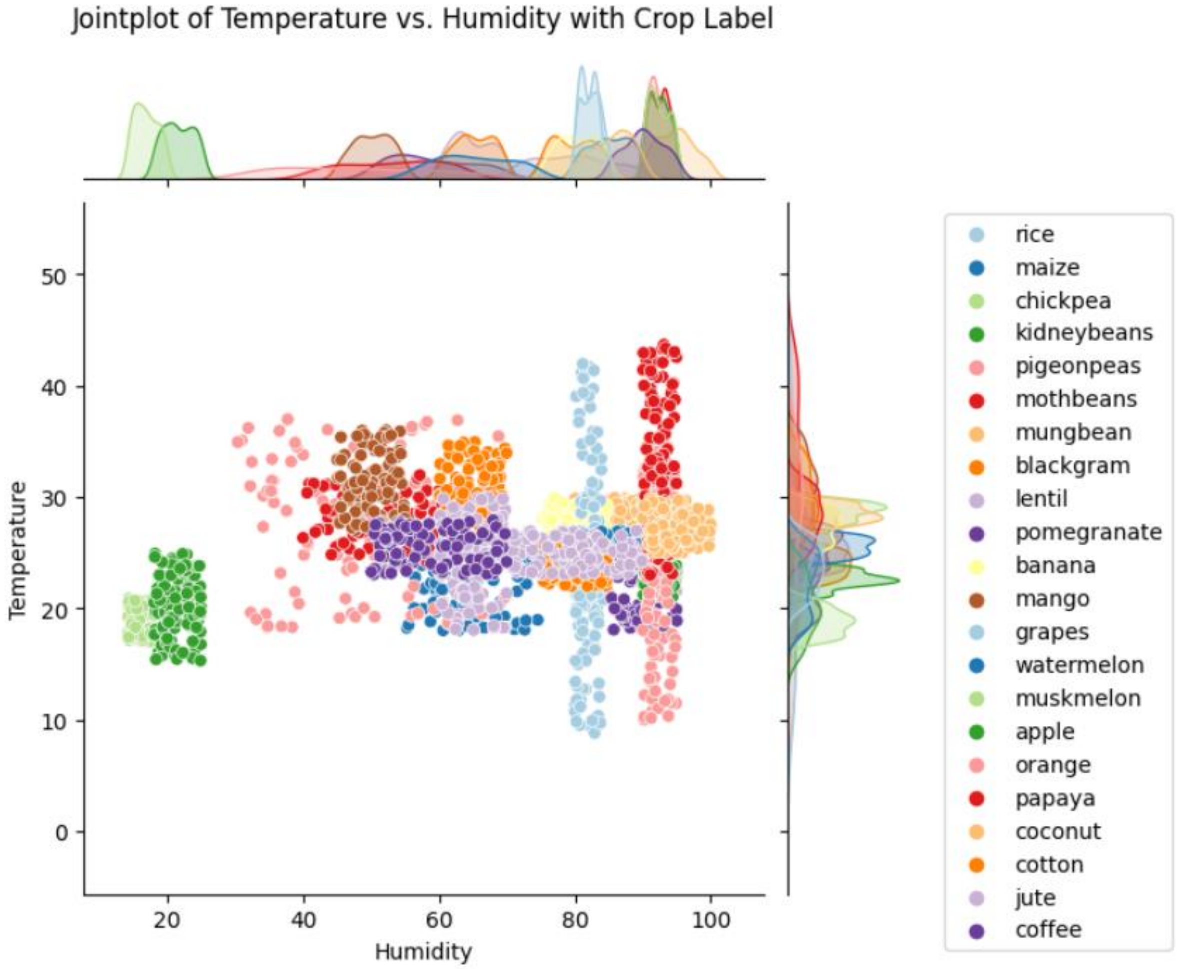


Figure 6: Jointplot (Temperature vs Humidity)

This joint plot displays the relationship between temperature and humidity for various crops. Rice thrives at high humidity (80-100%) and moderate temperatures (20-30°C), while Maize and Chickpea prefer moderate humidity (40-60%) and slightly higher temperatures (20-35°C). Pomegranate, Banana, Mango, and Grapes grow in a wider humidity range (40-80%) with temperatures around 20-35°C. Cotton and Coffee favour high humidity (60-90%) and higher temperatures (25-40°C). The density plots on the axes show the distribution of temperature and humidity preferences, providing a comprehensive view of optimal growing conditions for each crop.

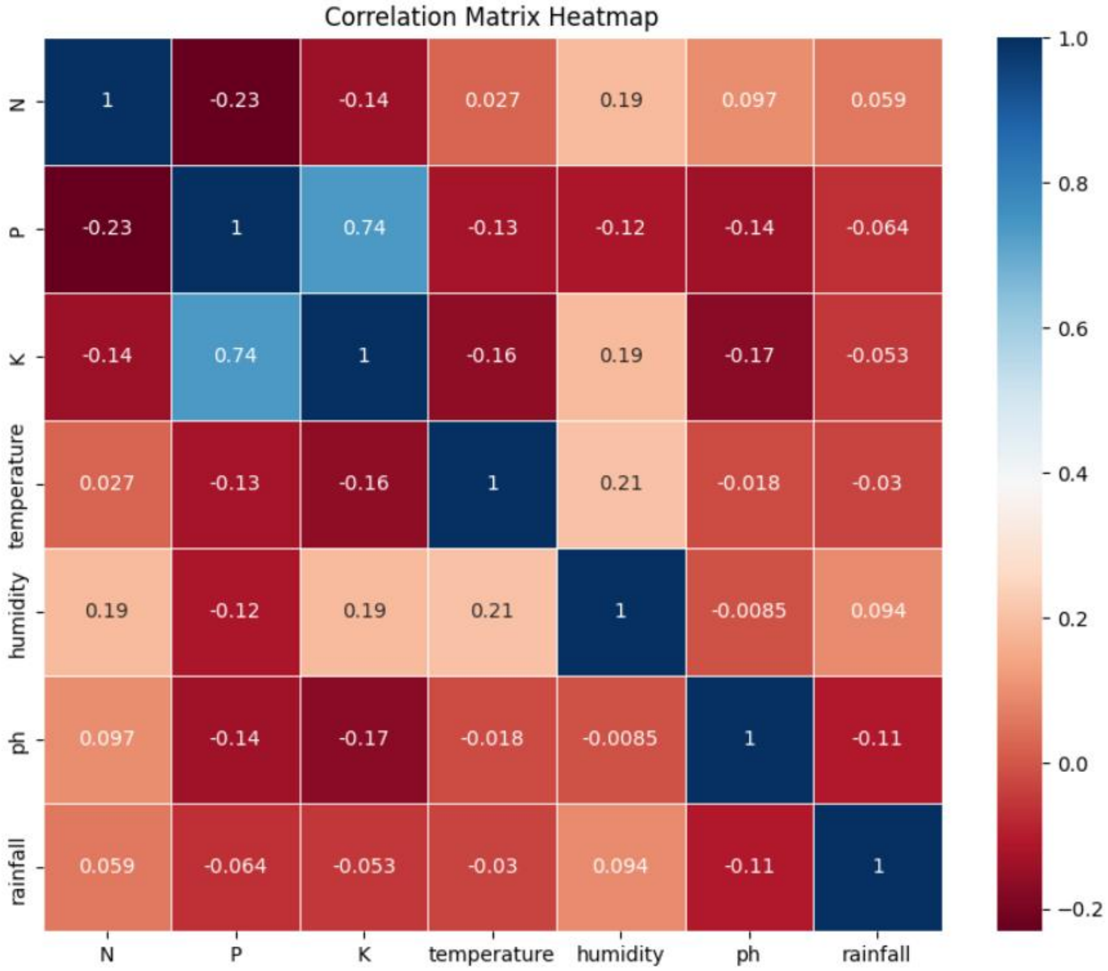


Figure 7: Correlation Matrix Heatmap

This correlation matrix heatmap shows the relationships between various factors affecting soil health: N (Nitrogen), P (Phosphorus), K (Potassium), temperature, humidity, pH, and rainfall. A strong positive correlation (0.74) exists between P and K, indicating that as phosphorus levels increase, potassium levels tend to increase as well. Conversely, there is a weak or no significant correlation among most other pairs, suggesting that these factors generally do not influence each other strongly. The colours help quickly identify these relationships, with dark blue representing strong positive correlations and dark red representing negative correlations.

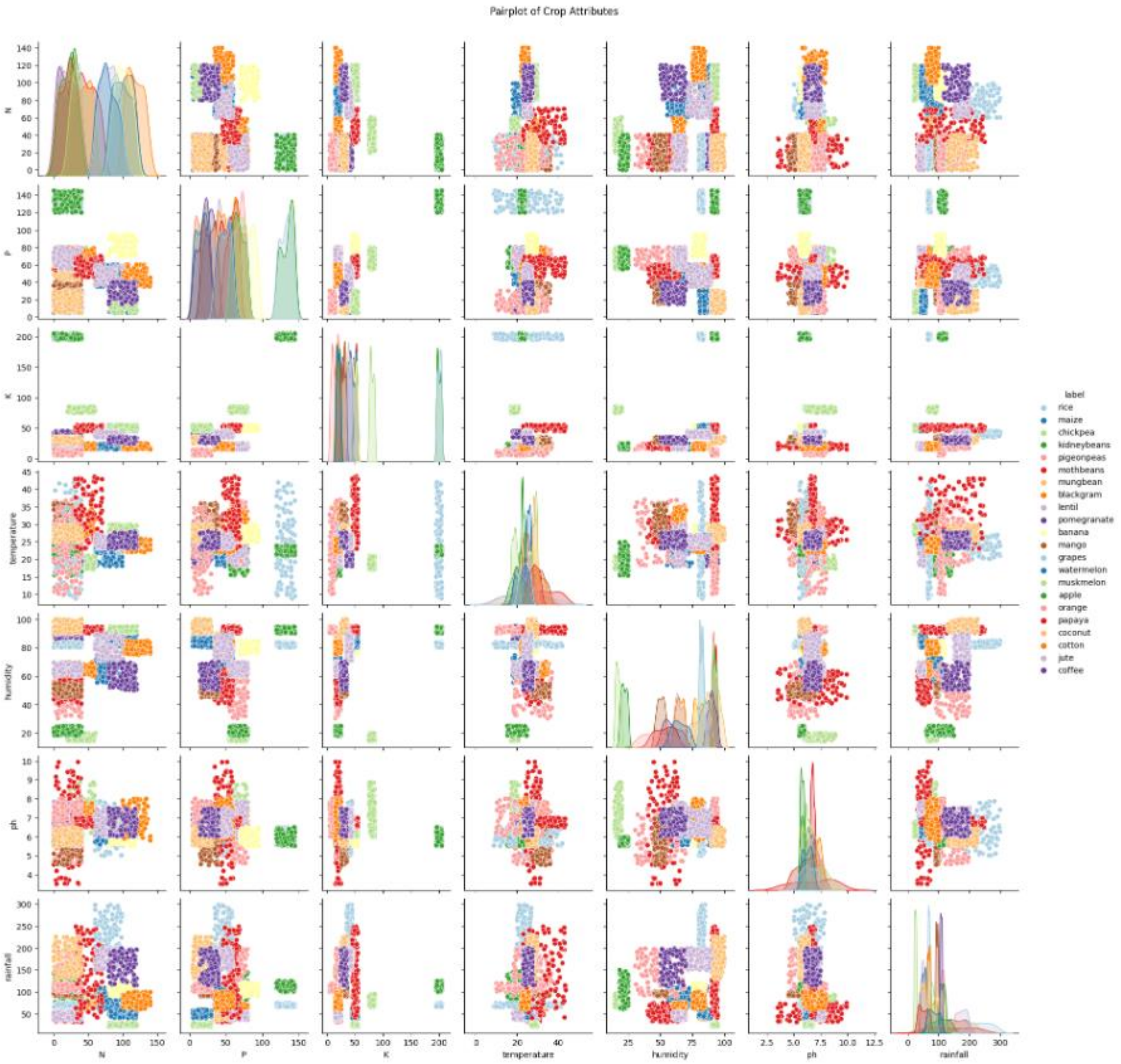
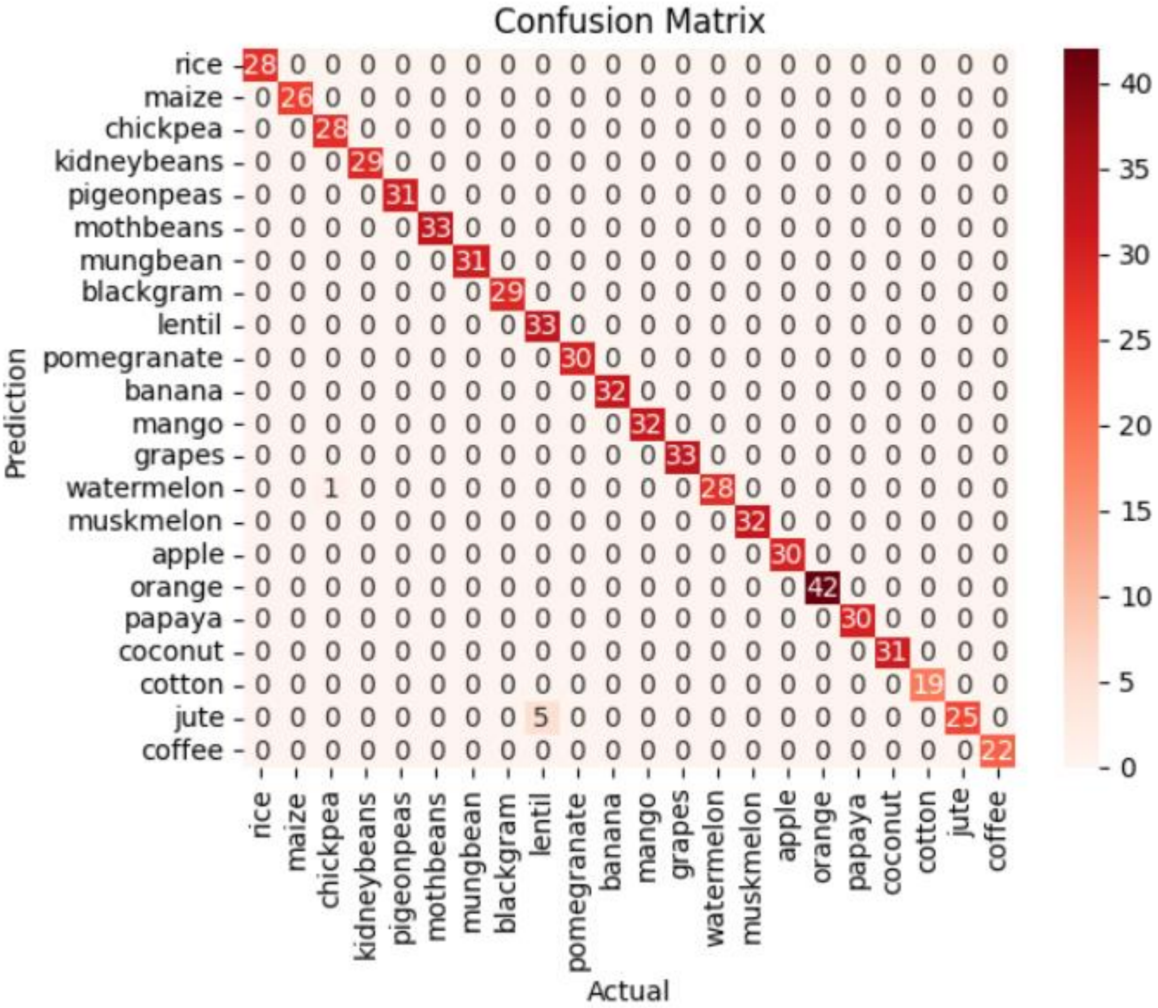


Figure 8: Pairplot of Crop attributes

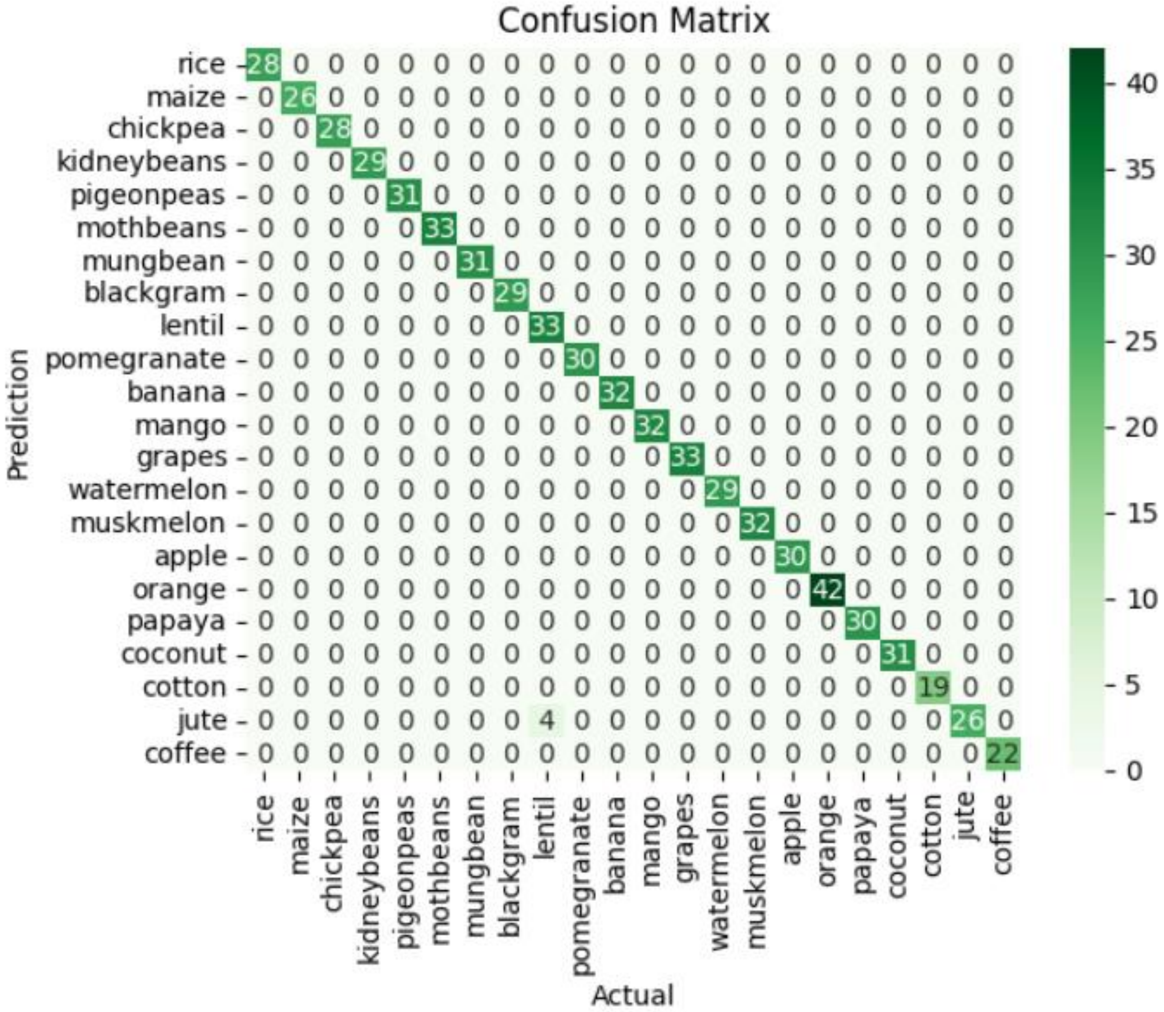
This Pairplot of crop attributes shows the relationships between different soil and environmental factors (N, P, K, temperature, humidity, pH, and rainfall) across various crops. Each crop type is color-coded, allowing us to observe how different crops cluster based on these attributes. For example, certain crops like rice and papaya show distinct groupings, indicating specific ranges of environmental conditions they thrive in. The diagonal histograms display the distribution of each attribute, while the scatter plots reveal potential correlations and clusters among the crops, highlighting how different crops prefer distinct combinations of soil and climate conditions.



	rice	maize	chickpea	kidneybeans	pigeonpeas	mothbeans	mungbean	blackgram	lentil	pomegranate	banana	mango	grapes	watermelon	muskmelon	apple	orange	papaya	coconut	cotton	jute	coffee
rice	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
maize	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chickpea	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidneybeans	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pigeonpeas	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mothbeans	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mungbean	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
blackgram	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lentil	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0
pomegranate	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0
banana	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0
mango	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0
grapes	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0
watermelon	0	0	1	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0
muskmelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0
apple	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0
papaya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0
coconut	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0
cotton	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0
jute	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	25	0
coffee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22

Figure 9: Confusion Matrix (Naïve Forest)

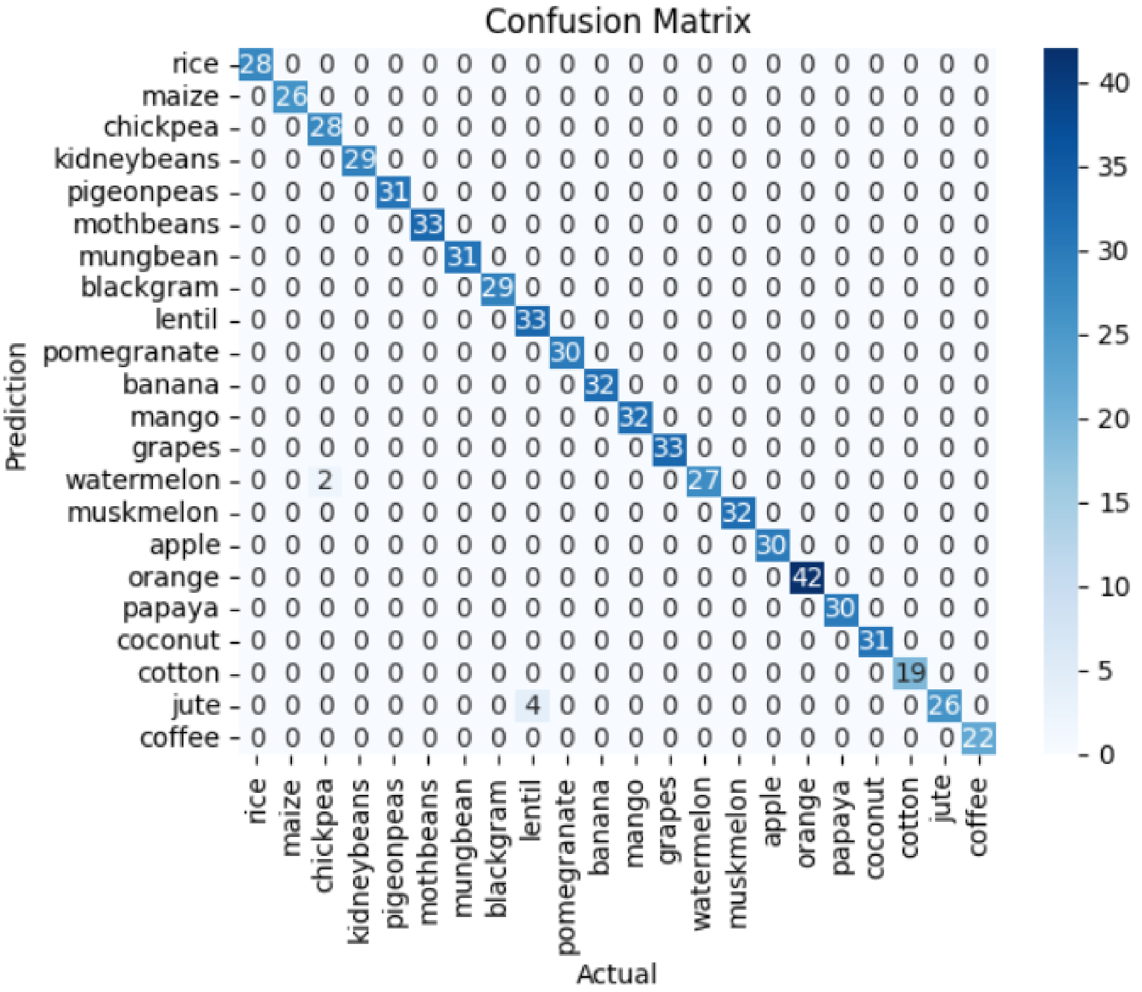
This confusion matrix for the 'Naive Forest' ensemble model, combining Naive Bayes and Random Forest, demonstrates high classification accuracy across various crops. The diagonal entries show the correctly predicted instances for each crop, with most values being non-zero, indicating successful predictions. Off-diagonal entries are zero or very low, suggesting minimal misclassification. The model excels particularly in predicting crops like orange (42 correct predictions), banana (32), and mango (32), while having minor errors such as a few misclassified jute samples. Overall, the model shows robust performance with precise crop identification.



	rice	maize	chickpea	kidneybeans	pigeonpeas	mothbeans	mungbean	blackgram	lentil	pomegranate	banana	mango	grapes	watermelon	muskmelon	apple	orange	papaya	coconut	cotton	jute	coffee
rice	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
maize	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chickpea	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidneybeans	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pigeonpeas	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mothbeans	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mungbean	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
blackgram	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lentil	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0
pomegranate	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0
banana	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0
mango	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0
grapes	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0
watermelon	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0
muskmelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0
apple	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0
papaya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0
coconut	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0
cotton	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0
jute	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	26	0
coffee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22

Figure 10: Confusion Matrix (AdaBayes)

The confusion matrix for the AdaBayes model, which combines AdaBoost and Naive Bayes, demonstrates strong predictive performance with most predictions falling on the diagonal, indicating high accuracy. Each class, represented along the rows and columns, shows clear separations with minimal off-diagonal entries, suggesting that the model effectively distinguishes between different crop types. The highest counts on the diagonal (e.g., orange with 42 correct predictions, and most others around 28-33) reflect the model's robustness and reliability in correctly classifying these categories. Only a few misclassifications are observed (e.g., jute predicted as other crops), indicating areas where the model could potentially improve. Overall, the model shows excellent performance in accurately predicting crop types with a high level of precision.



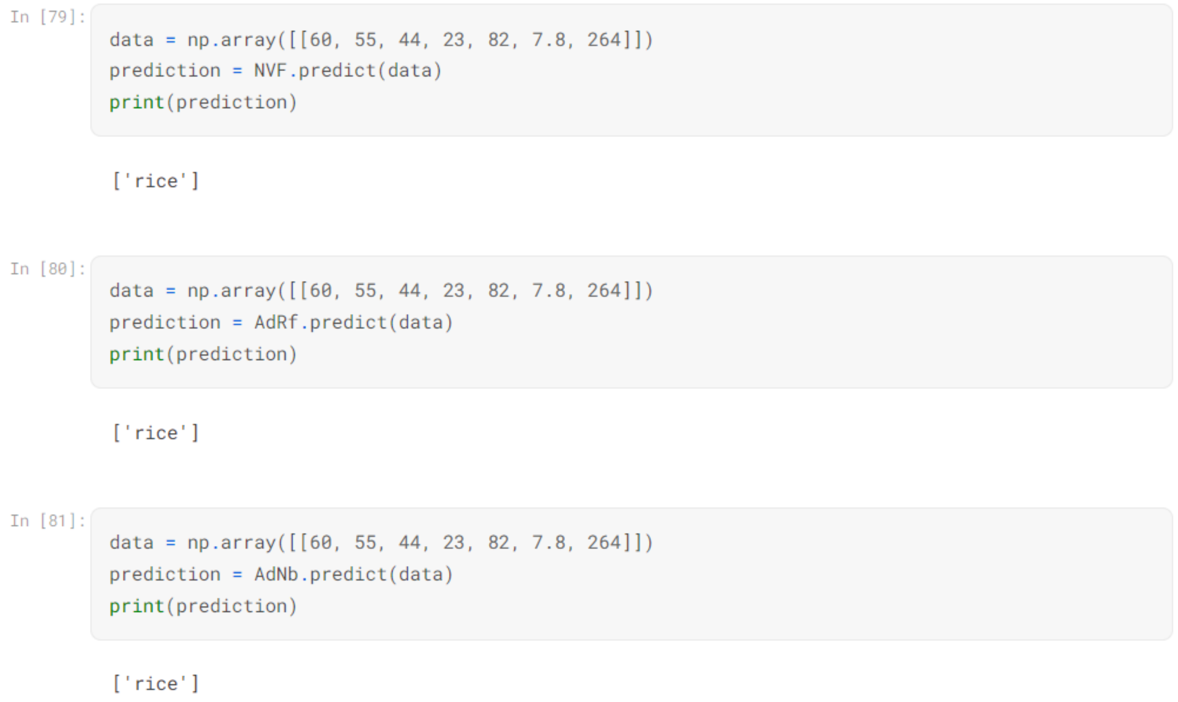
	rice	maize	chickpea	kidneybeans	pigeonpeas	mothbeans	mungbean	blackgram	lentil	pomegranate	banana	mango	grapes	watermelon	muskmelon	apple	orange	papaya	coconut	cotton	jute	coffee
rice	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
maize	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chickpea	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidneybeans	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pigeonpeas	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mothbeans	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mungbean	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
blackgram	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lentil	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0
pomegranate	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0
banana	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0
mango	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0
grapes	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0
watermelon	0	0	2	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0
muskmelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0
apple	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0
papaya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0
coconut	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0
cotton	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0
jute	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	26	0
coffee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22

Figure 11: Confusion Matrix (AdaForest)

The confusion matrix for the AdaForest model, which combines AdaBoost and Random Forest, shows that this ensemble approach achieves high accuracy in predicting crop types, similar to the AdaBayes model. Most predictions align perfectly with the diagonal, indicating correct classifications across various crop categories. High values on the diagonal (e.g., orange with 42 correct predictions, other crops generally in the range of 26-33) signify the model's strong performance and ability to correctly identify crop types. Misclassifications are minimal, with only a few instances of predictions falling off the diagonal, such as a slight confusion with watermelon and jute. Overall, the AdaForest model demonstrates excellent precision and reliability, effectively differentiating between the crop types with a high level of accuracy.

6.4. Crop Prediction:

Crop Prediction results using our ensembled models ‘Naïve Forest’, ‘AdaBayes’ and ‘AdaForest’ are shown below in Fig. 12. By specifying the values of the following attributes namely N, P, K, Temperature, Humidity, pH and Rainfall, our models predict the best possible crop to produce for the given conditions.



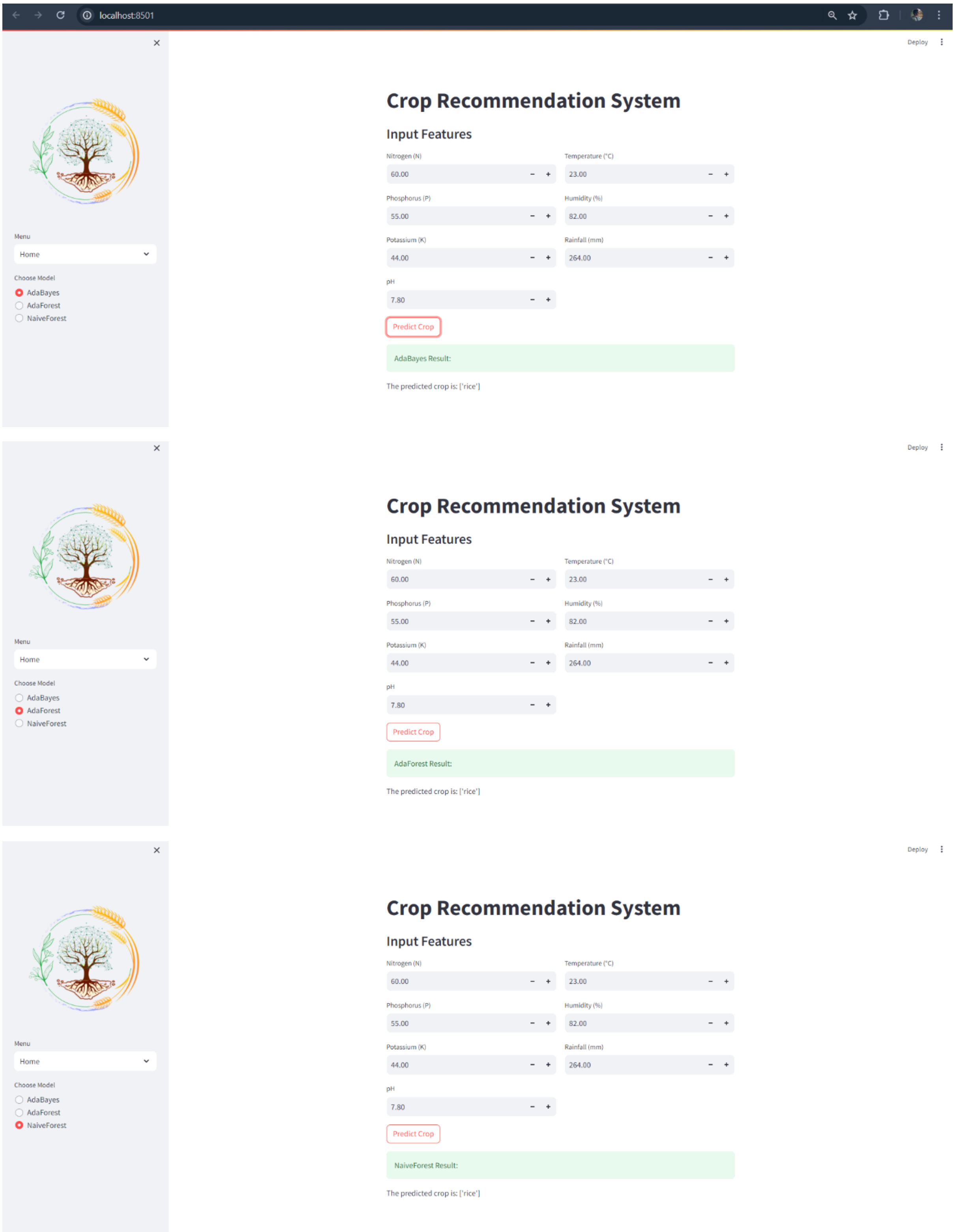
```
In [79]:
data = np.array([[60, 55, 44, 23, 82, 7.8, 264]])
prediction = NVF.predict(data)
print(prediction)

['rice']

In [80]:
data = np.array([[60, 55, 44, 23, 82, 7.8, 264]])
prediction = AdRf.predict(data)
print(prediction)

['rice']

In [81]:
data = np.array([[60, 55, 44, 23, 82, 7.8, 264]])
prediction = AdNb.predict(data)
print(prediction)

['rice']
```

Figure 12: Crop Prediction using Ensemble Models

6.5. Web Application Interface:

Streamlit is an open-source Python framework for data scientists and AI/ML engineers to deliver dynamic data apps with only a few lines of code. Build and deploy powerful data apps in minutes. Streamlit apps have a client-server structure. The Python backend of the app is the server. The frontend we view through a browser is the client. When we develop an app locally, our computer runs both the server and the client.

Here Streamlit is used along with Visual Studio Code to run the web application at localhost. In the application, three different ensemble models can be chosen separately to make predictions.



Menu: Home

Choose Model
● AdaBayes
○ AdaForest
○ NaiveForest

Crop Recommendation System

Input Features

Nitrogen (N): 60.00
Temperature (°C): 23.00
Phosphorus (P): 55.00
Humidity (%): 82.00
Potassium (K): 44.00
Rainfall (mm): 264.00
pH: 7.80

AdaBayes Result:

The predicted crop is: ['rice']

Menu: Home

Choose Model
○ AdaBayes
● AdaForest
○ NaiveForest

Crop Recommendation System

Input Features

Nitrogen (N): 60.00
Temperature (°C): 23.00
Phosphorus (P): 55.00
Humidity (%): 82.00
Potassium (K): 44.00
Rainfall (mm): 264.00
pH: 7.80

AdaForest Result:

The predicted crop is: ['rice']

Menu: Home

Choose Model
○ AdaBayes
○ AdaForest
● NaiveForest

Crop Recommendation System

Input Features

Nitrogen (N): 60.00
Temperature (°C): 23.00
Phosphorus (P): 55.00
Humidity (%): 82.00
Potassium (K): 44.00
Rainfall (mm): 264.00
pH: 7.80

NaiveForest Result:

The predicted crop is: ['rice']

Figure 13: ‘Crop Recommendation System’ – the interactive web application

6.6. Limitations of the Study:

The limited availability of comprehensive data for all significant crops within this domain of study restricts our research to focusing on a select few key crops and their specific conditions within a confined geographic region. This limitation necessitates a targeted approach, wherein we concentrate our efforts on the most prevalent and impactful crops for which data is available. By doing so, we aim to derive meaningful insights and conclusions that, while not universally applicable to all crops, can still provide valuable information about agricultural trends and practices in the region under study. This focused methodology allows for a deeper and more detailed analysis of the selected crops, potentially uncovering patterns and recommendations that could inform broader agricultural policies and strategies when more comprehensive data becomes available.

6.7. Suggestions for Future Research:

In the future, integrating real-time data for more accurate predictions in wide-scale regions and analysing other deep-learning algorithms will help us predict the crops for the correct parameters more accurately. Additionally, a crop recommendation system can be created in the form of a mobile application for ease of access to farmers which will help them in better understanding of the system and information.

7. Conclusion

In a nutshell, the proposed ensemble ML models named ‘Naïve Forest’, 'AdaBayes', 'AdaForest' in crop production appears to be a critical step in addressing the difficulties imposed by environmental unpredictability and assuring sustainable agricultural practices. This knowledge can empower stakeholders to make resourceful choices, optimize resource allocation, and contribute to sustainable land management. It will also ensure food security, promote sustainable farming practices, and equip farmers with the tools they need to navigate the complexities of a rapidly changing world. Our proposed ensemble models will serve as a foundation for future research endeavours in the realm of forecasting crops.

8. References:

1. Subhadra Mishra Gour Hari Santra and Debahuti Mishra. Applications of machine learning in the production of agricultural crops. Indian Journal of Science and Technology, 9(38), 2016.

2. G Muthukumarasamy N Balakrishnan. Crop production-ensemble machine learning model for prediction. International Journal of Computer Science and Software Engineering, 5(7):148– 153, 2016.

3. Mayank Champaneri, Darpan Chachpara, Chaitanya Chandvidkar, and Mansing Rathod. Crop yield prediction using machine learning. International Journal of Science and Research, 9(2), 2020.

4. A Mythili N Saranya. Classification of soil and crop suggestion using machine learning techniques. International Journal of Engineering Research & Technology, 9(2), 2020.

5. Baisali Ghosh. Crop yield prediction for cultivating alternative crops based on weather and soil conditions using machine learning algorithm. International Journal of Modern Developments in Engineering and Science, 2016.

6. Utkarsha P Narkhede and K P Adhiya. Evaluation of modified kmeans clustering algorithm in crop prediction. International Journal of Advanced Computer Research, 4(3), 2014.

7. Angu Raj, Dr. Thiyaneswaran Balashanmugam, J .Jayanthi, N Yoganathan, and P Srinivasan. Crop recommendation on analyzing soil using machine learning. Turkish Journal of Computer and Mathematics Education, 12:1784–1791, 2021.

8. M Hasan, MA Marjan, MP Uddin, MI Afjal, S Kardy, S.Ma, and Y Nam. Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. Front. Plant Sci., 14(1234555), 2023.

9. V K Lohit, L. Vijayalakshmi, G Brunda, M D Sanjay, and K T Rashmi. Crop yield prediction using machine learning. International Journal of Engineering Research & Technology, 10(12), 2022.

10. A. Kumar P. A, S. Chakraborty and O. R. Pooniwala. Intelligent crop recommendation system using machine learning. International Conference on Computing Methodologies and Communication, pages 843–848, 2021.

11. B. Misra S. Veenadhari and C. Singh. Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics, pages 1–5, 2014.

12. P. Mehta, H. Shah, V. Kori, V. Vikani, S. Shukla, and M. Shenoy. Survey of unsupervised machine learning algorithms on precision agricultural data. International Conference on Innovations in Information, Embedded and Communication Systems, pages 1–8, 2015.

13. U.Muthaiah & M.Balamurugan P.Priya. Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Sciences & Research Technology, 7(4):1–7, 2018.

14. S Djodiltachoumy E Manjula. A model for prediction of crop yield. International Journal of Computational Intelligence and Informatics, 6(4), 2017.

15. S. Athani, C. H. Tejeshwar, M. M. Patil, P. Patil, and R. Kulkarni. Soil moisture monitoring using iot enabled arduino sensors with neural networks for improving soil management for farmers and predict seasonal rainfall for planning future harvest in north karnataka — india. International Conference on I-SMAC, pages 43–48, 2017.

16. P. Kumar R. Kumar, M. P. Singh 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, pages 138–145, 2015.

17. Y.Radhika and M.Shashi. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering, 1(1):55–58, 2009.

18. Modeling and forecasting tasks of agriculture based on machine learning, Baratbek Sabitov, Asel Kartanova, Talant Kurmanbek uulu, Nazgul Seitkazieva, Ainura Dyikanova, Aida Orozobekova, E3S Web Conf. 380 01026 (2023), DOI: 10.1051/e3sconf/202338001026

19. Nihar, Ashmitha & R., Patel & Danodia, Abhishek. (2022). Machine-Learning-Based Regional Yield Forecasting for Sugarcane Crop in Uttar Pradesh, India. 10.1007/s12524-022-01549-0.

20. Alberto Torres-Barran, Alvaro Alonso, Jose R. Dorronsoro, Regression, Tree Ensembles for Wind Energy and Solar Radiation Prediction, Neurocomputing (2017), doi: 10.1016/j.neucom.2017.05.104

21. H. R. Seireg, Y. M. K. Omar, F. E. A. El-Samie, A. S. El-Fishawy and A. Elmahalawy, "Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction," in IEEE Access, vol. 10, pp. 64671-64687, 2022, doi: 10.1109/ACCESS.2022.3181970. keywords: {Crops; Predictive models; Meteorology; Data models; Computational modeling; Stacking; Feature extraction; Bayesian optimization; cascading technique; GBR; LGBM; Ridge; stacking technique; XGBoost; EMLA; wild blueberry yield}