Crop Prediction Using Machine Learning Algorithm

Rohit Joshi[1], Ojas Gurav[2], Rajeshwar Jadhav[3], Tanishk Shinde[4]

[1,2,3,4]Student Researcher, Department of Computer Science and Engineering, Vishwakarma University, Pune, India.

Abstract:

Many developing nations rely heavily on agriculture as their main source of income. The ongoing evolution of modern agriculture involves constant innovation in farming practices. Meeting the ever-changing demands of our planet and satisfying the expectations of merchants and consumers is significant challenges for farmers. Some of these challenges include coping with climate changes due to soil erosion and industrial emissions, addressing nutrient deficiencies in the soil (such as potassium, nitrogen, and phosphorus), which can hinder crop growth, and overcoming the tendency to cultivate the same crops repeatedly without experimenting with different varieties. The objective of the paper is to identify the optimal model for predicting crop outcomes, assisting farmers in choosing suitable crops based on climatic conditions and soil nutrient levels. This paper compare accuracy of different and using best of among these.

Keywords: machine learning, crop prediction, accuracy, random forest classifier.

1. Introduction:

Machine learning serves as a valuable tool for decision-making in predicting agricultural yields and determining optimal crop choices and activities throughout the growing season. Various machine learning methods have been employed in crop prediction studies to enhance accuracy and efficiency.

We're diving into the world of using fancy computer programs to predict which crops will do well on a farm. Imagine we have sensors in the soil that measure things like nutrients (N, P, K), pH levels, humidity, and temperature. By gathering and analyzing this info, we want to create a smart system that helps farmers decide which crops to plant. We're not just stopping there; we're also testing different computer models to see which one works best. One standout is the Random Forest model—it's like a digital superhero for handling lots of information and making accurate predictions. The big picture is to give farmers a handy tool that considers all sorts of factors, so they can make smarter choices about what to plant among provided historical crop dataset.

2. Literature Survey

Ashwani Kumar Kushwaha[2] discusses methods for predicting crop yield to enhance farmer profits and the agriculture sector's quality. The study employs big data, specifically soil and weather data, collected through the Hadoop platform and agro algorithms. The repository data is utilized to predict suitable crops for specific conditions, thereby improving crop quality.

Dahikar S[3] highlights the significance of crop prediction and proposes methods to enhance accuracy. The paper introduces a feed-forward backpropagation Artificial Neural Network approach to model and forecast crop yields in rural areas based on soil parameters (PH, nitrogen, potassium) and atmospheric parameters (rainfall, humidity).
Rahul Katarya[4] investigates diverse machine learning approaches to enhance crop yield. The paper explores the application of artificial intelligence, incorporating machine learning algorithms and big data analysis in precision agriculture. The author details the implementation of a crop recommendation system, employing methods such as K-Nearest Neighbors (KNN), Ensemble-based Models, and Neural Networks.
Dhanush Vishwakarma[5] utilizes the SVM algorithm for rainfall prediction and the Decision Tree algorithm for crop prediction. Input variables include N, P, K, pH, rainfall, and humidity. This integrated approach aims to provide accurate predictions for better agricultural planning.

3. Proposed Work :

The proposed system aims to forecast the optimal crop for a specific piece of land by considering soil composition and various weather parameters, including temperature, humidity, soil pH, and rainfall. by collecting these values by sensors and testing .

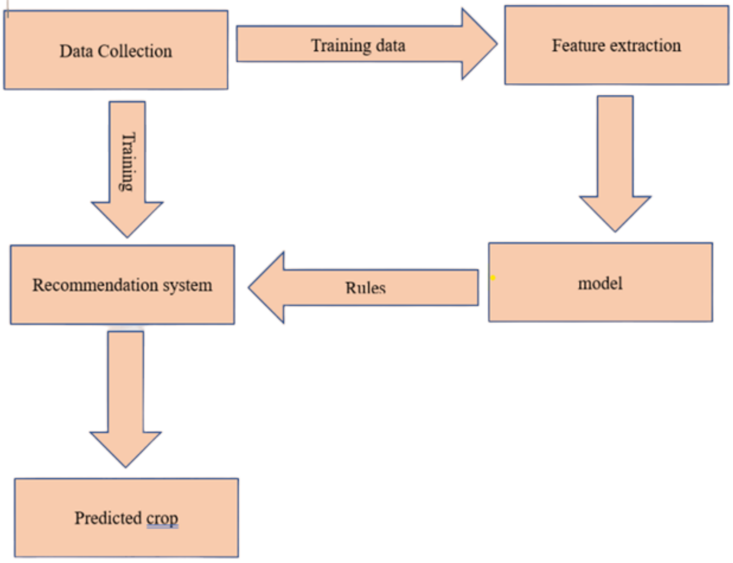


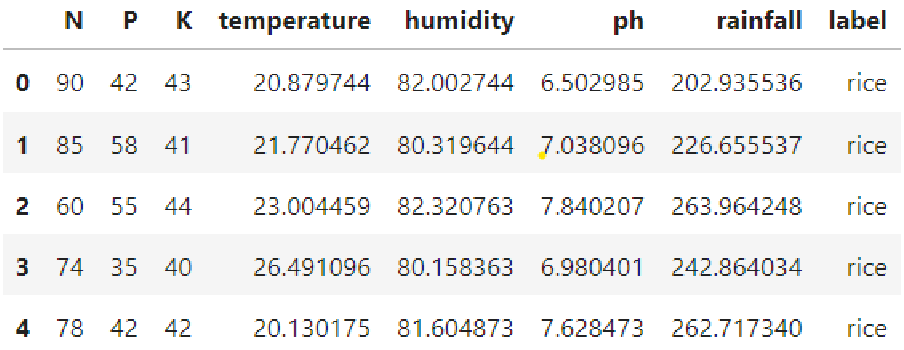
Figure 1. Steps Involved in the Methodology

3.1. Collecting Raw Data

The process of collecting and analyzing data from diverse sources is known as data collection. This practice facilitates the retrospective examination of events and supports the application of data analysis to identify recurrent patterns. The dataset utilized in the 'Crop Recommendation' project is sourced from the Kaggle platform. This dataset includes 22 distinct crops as class labels and comprises seven attributes:

1. Nitrogen Ratio (N): The proportion of nitrogen in the soil, a critical factor for plant growth.
2. Phosphorus Ratio (P): The proportion of phosphorus in the soil, an essential nutrient influencing plant development.
3. Potassium Ratio (K): The proportion of potassium in the soil, a key element for overall plant health.
4. Soil Temperature: The temperature in degrees Celsius of the surrounding environment, affecting biochemical reactions in plants.

5. Relative Humidity: The percentage of water vapor in the air relative to its maximum capacity at the given temperature.
6. Soil pH: The measurement of soil acidity or alkalinity, influencing nutrient availability for plants.
7. Rainfall: The amount of precipitation in millimeters, a crucial factor for crop irrigation.



	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Figure 2. Dataset Sample

For collecting humidity and temperature used DHT11 sensor and DS18B20 sensor respectively.
DHT11 specification:

1. Supply Voltage: 3.5V to 5.5V
2. Operating Current: 0.3mA (during measurement), 60uA (in standby)
3. Output: Serial Data
4. Humidity Range: 20% to 90%
5. Output Resolution: 16 bits
6. Accuracy: ±1°C and ±1%

DS18B20 specification:

1. Type: Programmable Digital Temperature Sensor
2. Operating Voltage: 3V to 5V
3. Temperature Range: -55°C to +125°C
4. Accuracy: ±0.5°C
5. Output Resolution: Programmable, 9-bit to 12-bit

3.2. Data Preprocessing

Data preprocessing is the crucial step of transforming raw data into a format suitable for analysis and machine learning algorithms. It enables analysts and data scientists to derive insights or predict outcomes. In this project, our data preprocessing primarily focuses on identifying and handling missing values. It's common for datasets to contain empty cells, null values, or specific characters like question marks, all of which may indicate missing data. Fortunately, the dataset utilized in this project is free from any missing values.

3.3. Train and Test Split

The dataset is divided into a training dataset and a testing dataset using the train_test_split() method from the scikit-learn module. Out of the 2200 data points in the dataset, 80% (1760 data points) constitute the training dataset, while the remaining 20% (440 data points) form the testing dataset.

3.4.1. Fitting the model

Model fitting involves adjusting the model's parameters to enhance accuracy. This process entails running the algorithm on labeled data to establish a machine learning model. The model's accuracy is then assessed by comparing its predictions against the known target variable. A well-fitted model can generalize well to new, unseen data..

3.4.2. Model used for system is Random Forests Classifier .

The chosen model for this system is the Random Forests Classifier. This ensemble learning method utilizes multiple decision tree classifiers to improve overall performance. The algorithm creates decision trees randomly using instances from the training set. Each decision tree provides predictions, and the final model prediction is determined through majority voting. The Random Forests Classifier is favored in machine learning due to its ability to handle overfitting issues, with increased accuracy achievable by incorporating more trees.

1. Randomly select K instances from the provided training dataset.
2. Build decision trees based on the chosen instances.
3. Specify the number of estimators (N) for the total trees to be created.
4. Repeat steps 1 and 2 for N iterations.
5. For a new instance, gather predictions from each estimator, and assign the category with the highest number of votes as the final prediction.

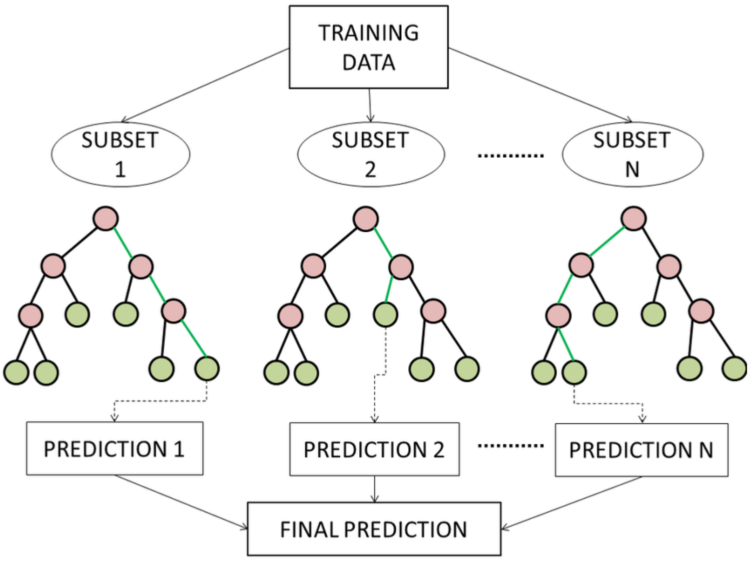


Figure 3. Random Forest Classifier

3.4.3. Testing on model

1. Confusion matrix

The confusion matrix displays the counts of true negatives, false negatives, true positives, and false positives.

2. MCC

MCC stands for Matthews Correlation Coefficient. It is a metric used to assess the quality of binary (two-class) and multiclass classifications. The Matthews Correlation Coefficient considers true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) to offer a well-balanced evaluation of the classifier's effectiveness.

$$\text{MCC} = \frac{\text{TP}*\text{TN} - \text{FP}*\text{FN}}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

3. F1 score

The F1 score is a weighted harmonic mean of precision and recall, ranging from 0.0 (indicating the worst performance) to 1.0 (representing the best performance). Unlike accuracy measurements, F1 scores tend to be lower, as they take into account both precision and recall in their computation.

$$\text{F1 score} = \frac{2 * \text{PR}}{(\text{P} + \text{R})}$$

Where P-Precision; R-Recall

4. Accuracy

Accuracy Model accuracy is calculated as the ratio of correct predictions to the total number of predictions. This metric is part of the metrics module.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

In the given sentence, TP represents True Positive, FP stands for False Positive, TN denotes True Negative, and FN represents False Negative.

4. Result and Analysis

We conducted experiments with various machine learning models, including the random forest classifier, decision tree, support vector machine, k-nearest neighbors (KNN), and a stacked model. Our dataset contains seven features: (i) Content ration of N (ii) Content ration of P, (iii) Content ration of K in the soil, (iv) Temperature expressed in degrees Celsius, (v) Relative humidity percentage, (vi) pH value, and (vii) Rainfall measured in millimeters.

The result obtained are represented using accuracy , F1 score, MCC are as follows:

4.1. random forest classifier

Model performance for Training set

- Accuracy: 100 %
- MCC: 1.0
- F1 score: 1.0

Model performance for Test set

- Accuracy: 99.09 %
- MCC: 0.9905
- F1 score: 0.9909

4.2 Decision tree

Model performance for Training set

- Accuracy: 51.19 %
- MCC: 0.5161
- F1 score: 0.4141

Model performance for Test set

- Accuracy: 0.45227272727272727
- MCC: 0.4627441443855779
- F1 score: 0.3684802868057341

4.3 support vector machine

Model performance for Training set

- Accuracy: 98.57 %
- MCC: 0.9851
- F1 score: 0.9858

Model performance for Test set

- Accuracy: 96.59 %
- MCC: 0.9644
- F1 score: 0.9658

4.4. KNN

Model performance for Training set

- Accuracy: 99.03 %
- MCC: 0.9898
- F1 score: 0.9903

Model performance for Test set

- Accuracy: 0.9704545454545455
- MCC: 0.9691826495580284
- F1 score: 0.970540270491023

4.5. stack model

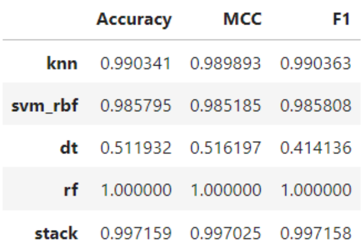
Model performance for Training set

- Accuracy: 0.9971590909090909
- MCC: 0.997025149427043
- F1 score: 0.9971578884960391

Model performance for Test set

- Accuracy: 0.9681818181818181
- MCC: 0.9668117578315487
- F1 score: 0.9685082026595072

4.6. Result for comparison between model



	Accuracy	MCC	F1
knn	0.990341	0.989893	0.990363
svm_rbf	0.985795	0.985185	0.985808
dt	0.511932	0.516197	0.414136
rf	1.000000	1.000000	1.000000
stack	0.997159	0.997025	0.997158

Figure 4. Comparison between models

4.7 Data analysis by visualization

4.7.1 pair plot representation of input variables .

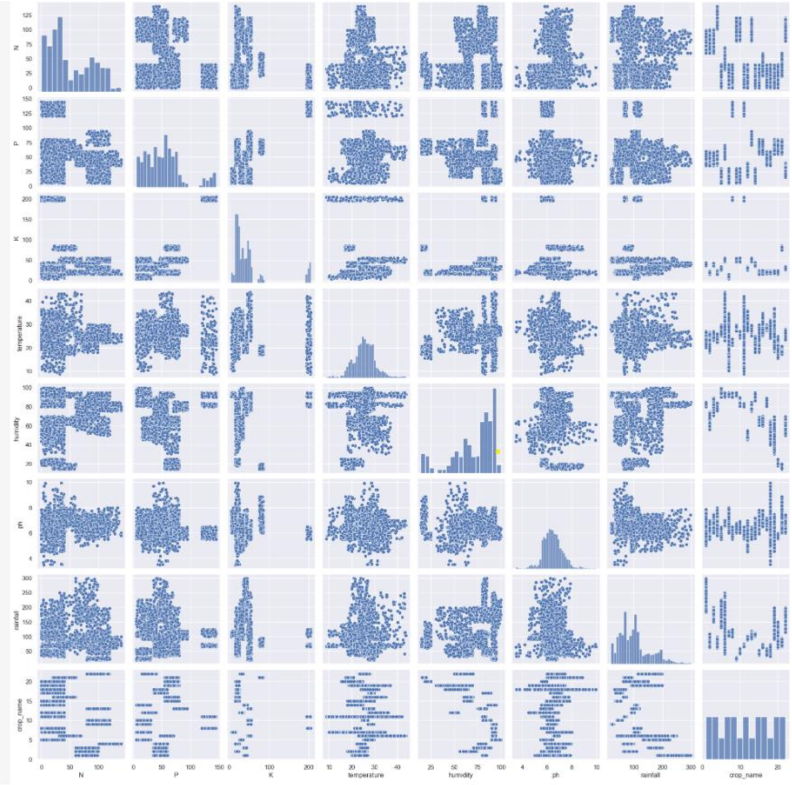


Figure 4.1 pair plot representation

4.7.2. How Ph value vary from crop to crop can be see in next figure.

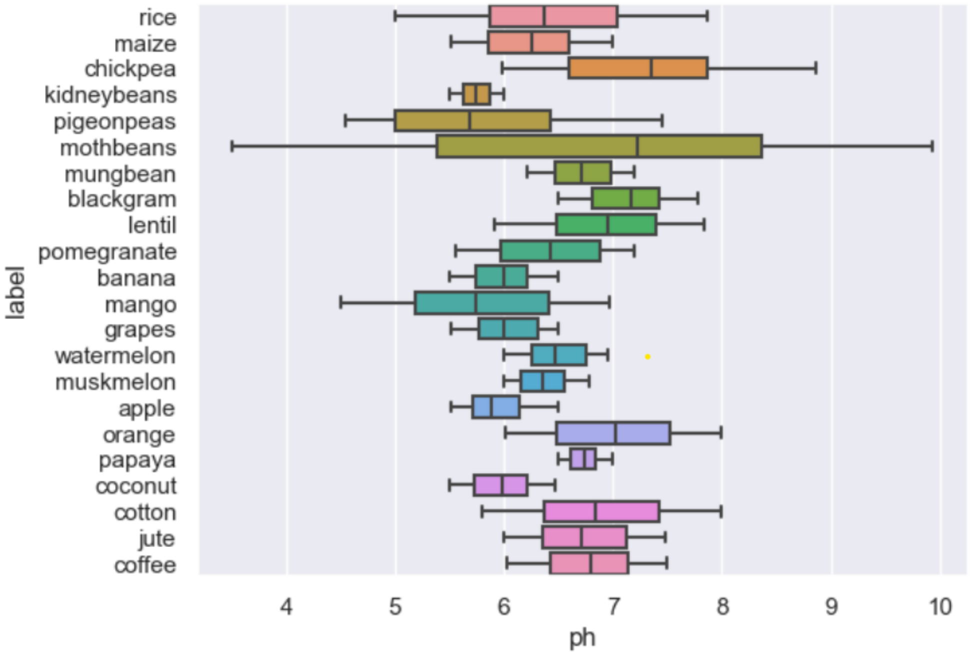
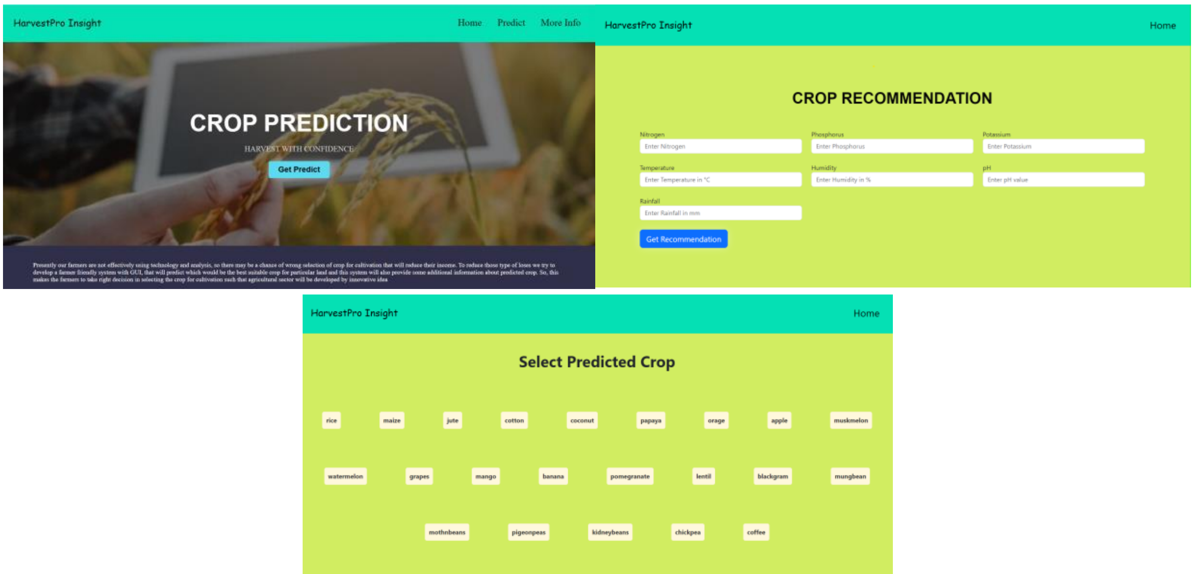


Figure 4.2 pH range representation

4.8 crop prediction user interface



Presently our farmers are not effectively using technology and analysis, so there may be a chance of wrong selection of crop for cultivation that will reduce their income. To reduce those type of loses we try to develop a farmer friendly system with GUI, that will predict which would be the best suitable crop for particular land and this system will also provide some additional information about predicted crop. So, this makes the farmers to take right decision in selecting the crop for cultivation such that agricultural sector will be developed by innovative idea

Figure 4.3 user interface of software

5. Conclusion:

Currently, our farmers are not making optimal use of technology and analysis, increasing the risk of selecting the wrong crops for cultivation and subsequently decreasing their income. In order to mitigate these potential losses, we are working on the development of a user-friendly system with a graphical user interface (GUI). This system aims to predict the most suitable crop for a specific piece of land and provide additional information about the recommended crop. By empowering farmers with accurate insights, we hope to enable them to make informed decisions in crop selection, contributing to the development of the agricultural sector through innovative ideas.

6. Reference :

1. Dinesh A. , “Crop prediction using machine learning”, Journal of Physics: Conference Series. 2161. 012033. 10.1088.
2. Ashwani K. , “crop yield prediction using agro algorithm in hatoop”, International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 5, No2, April 2015.
3. Rode V. ,”Agricultural crop yield prediction” ,International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering vol 2 Issue 1 pp 683-6.
4. Abhinav T. , “Impact of Machine Learning Techniques in Precision Agriculture”, International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, ICETCE48199.2020.9091741.
5. Nischitha K.,” Crop Prediction using Machine Learning Approaches”, International Journal Of Engineering Research & Technology, VOLUME 09, ISSUE 08 (AUGUST 2020).
6. Pavan P. , “Crop Prediction System using Machine Learning Algorithms”,International Journal of Enhineering Research & Technology. e-ISSN : 2395-0056 , Vol.7 Issue 02.