Crop Prediction using Random Forest Algorithm

Ashna Karim
PG Scholar
Department Of Computer Application
Amal Jyothi College Of Engineering
Kanjirappally, India
ashnakarim2024@mca.ajce.in

Jetty Benjamin
Assistant Professor
Department Of Computer Application
Amal Jyothi College Of Engineering
Kanjirappally, India
jettybenjamin@amaljyothi.ac.in

***Abstract*—India's population depends heavily on agriculture. A lot of farmers in India follow their intuition when deciding which crop to plant in a particular season. The most common issue that farmers face is when they do not select the appropriate crop according to the needs of their property and its surroundings. Consequently, production is impacted. Researchers have been trying to agricultural production forecasting utilizing multiple approaches and methodologies and have comparative analysis on such algorithms here using decision tree, logistic regression and random forest classifier. Machine learning approaches, in particular Random Forest classifiers, have become beneficial tools for precise and reliable crop prediction. To create a predictive model, this study analyzes historical data on soil characteristics and climate trends. As a reliable foundation for crop prediction, the Random Forest method is used to effectively record complicated linkages and interactions among different data. The outcomes show the potential for crop prediction powered by machine learning to improve agricultural output, optimize resource allocation, and support ethical farming methods.**

Keywords—Crop prediction, Machine Learning, Random Forest Classifier, Decision tree, Logistic regression

I. INTRODUCTION

The state's economy frequently depends greatly on the performance of agriculture, which serves as the foundation of several regions. Fertile land has been put to use for commercial purposes recently due to the rising demand for agricultural land brought on by industrialization, infrastructure growth, urbanization and other triggers. This is having an impact on the overall agricultural output in addition to reducing the amount of cultivable land. It is crucial to maximize agricultural output by accurate forecasts when the amount of arable land available decreases, especially in light of how unpredictable weather patterns can be.

Machine learning has become an efficient method for predicting crops as a remedy to this problem. By utilizing machine learning, we can increase agricultural output and aid farmers in selecting crops wisely, hence increasing yield in a certain area. This method uses a number of variables, such as soil properties(Nitrogen, Phosphorus, Potassium, pH), humidity, temperature, and rainfall, to generate accurate crop production projections. For classification problems, Random Forest is a strong ensemble machine learning method that is frequently utilized. It is renowned for its capacity to handle both small and big datasets with high-dimensional feature spaces, and is particularly useful in scenarios where precise classification is required.

There are other methods can be used to get predictions such as Decision Tree, Logistic Regression etc. Here using a Random Forest Classifier to predict crops which is based on the comparative evaluation of accuracy of these algorithms, the machine learning model is trained to categorize various crop types based on the input variables including soil characteristics, weather trends, and environmental factors. This enables the agriculture sector's actual productivity to rise as well.

II. LITERATURE REVIEW

Data mining approaches have been proposed by S. Veenadhari, Dr. Bharat Misra, and Dr. CD Singh.[1] It is a straightforward website that predicts the crop based on the user's preferences by providing local climatic information. Decision trees are used in this situation. They considered wheat, paddy, maize, soybean in a particular district and have 75% accuracy. The variables used here are rainfall, temperature, the cloud cover, wet day frequency and additionally they gathered the 20 years of data about the yield of crops from various additional resources. Sorting of an attribute based on data from each of them. The analysis of relevant qualities can be aided by using this score.

Girish L, Gangadhar S, Bharath T R, Balaji K S, and Abhishek K T [2] use a machine learning mechanism to analyze crop yield and rainfall. They addressed a variety of machine learning methodologies for forecasting agricultural production and rainfall in this study, as well as the effectiveness of several machine learning algorithms such as liner regression, SVM, KNN method & decision tree. They conclude that SVM is the most successful strategy for forecasting rainfall.

A specialized framework for crop prediction taking into consideration aspects like area-wise month-to-month precipitation and harvest productivity from 2000 to 2014 in Maharashtra was proposed S. Devadhe, A. Kausadikar, P. Daphal, A. Joshi,[3] Calculations involving linear regression, decision trees, and random forests are used to forecast harvest rates of production, which may help with deciding best to organize sporadic harvests for planting. However, because only two or three borders were considered, the review was only permitted to occasionally decide harvests. When compared to a pure relapse promotion choice tree, outcomes of the Random Forest simulation showed better forecasts.

Machine learning has been applied by Sangeetha, Shruthi, [4] and comprises supervised learning models. Random forest, Polynomial Regression, and Decision Tree approaches are used to evaluate performance. The article comes to the conclusion that Random Forest leads the other two algorithms in terms of yield prediction.

A Crop Selection Method Using Machine Learning for Increasing Crop Production Rate has been discovered by Rakesh Kumar, M.P. Singh, Prabhat, and J.P. Singh [5]. In their research, they propose a strategy called CSM for deciding the sequence of harvests to be produced over the course of the planting season. This season's crops may yield more significantly as a result of the CSM methodology. The suggested approach deals with crop selection based on anticipated yield rates affected by a variety of elements, including weather, soil characteristics, the density of water in that area, and crop type.

In order to provide ranchers, choose the best harvest for a particular plot of land, Banavlikar, A. Mahir, M. Budukh, S. Dhodapkar [6] developed an extensive, precise, and robust yield suggestion framework that uses neural networks. Temperature and soil are considered, and sensors for soil moisture, mugginess, and temperature are given to measure the amount of vapor in the surrounding air and the amount of water in the soil.

III. METHODOLOGY

The aim of this study is to build a machine learning model using the available dataset, can accurately forecast a crop's ability to grow in accordance with the input parameters supplied by the user. The dataset includes soil attributes and weather parameters has to be exposed to preprocessing. The process includes training phase and testing phase.

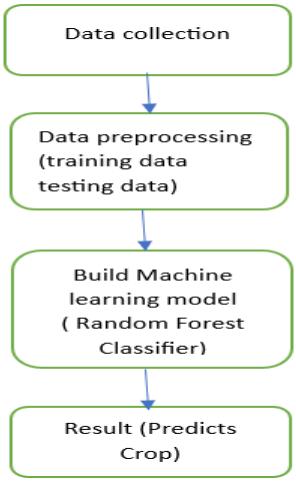

Fig 1. Methodology

A. Data Collection

The suggested model forecasts the crop based on the selected data sets from specific region. The dataset is taken from public platform and have 2200 rows and 8 columns. This includes soil attributes (Nitrogen, Phosphorus, Potassium, pH) and humidity, rainfall and temperature.

B. Data Preprocessing

Data preprocessing is an important phase in the machine learning procedure that involves cleaning, transforming, and organizing initial data for the training of models and analysis. And here also creates the training model. The performance of your machine learning models can be strongly impacted by the accuracy of your data and how effectively you preprocess it.

C. Machine Learning

The aim is to predict best crop for the particular data of a chosen region. This depends on the efficiency of machine learning model. In this study the algorithm is choosing based on the comparative evaluation of accuracy. For that here using decision tree, logistic regression and random forest classifier.

1.Decision tree

A decision tree is an effective tool in supervised learning algorithms that can be applied for both regression and classification programs. It generates a flowchart-like tree structure, with each internal node representing a test on an attribute, each branch reflecting a test outcome, and every terminal node, or leaf node, carrying a class label. It builds by iteratively splitting the training data into subsets depending on attribute values until a stopping requirement, such as the deepest point of the tree or the smallest quantity of samples required for splitting a node, is met.

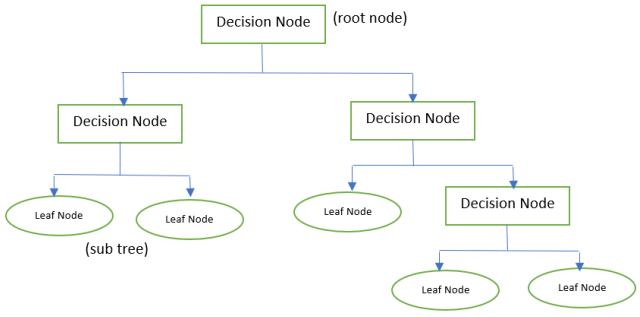

Fig 2. Decision tree

2. Logistic Regression

Logistic regression is a method of supervised machine learning that is primarily utilized for problems with classification, with the purpose of forecasting the likelihood that an instance belongs to a specified class. Its term is logistic regression, and it is utilized for classification methods. It is called regression for the reason that it utilizes the outcome of a linear regression value as input and estimates the likelihood for the selected class using a sigmoid function. The distinction among linear regression and logistic

regression is the fact that the outcome of linear regression is a continuous value that can be anything, but logistic regression forecasts whether or not an instance belongs to a particular category.

3.Random Forest Classifier

A common ensemble learning technique in machine learning for the classification of data is the Random Forest Classifier. It is a decision tree algorithm extension that is renowned for its capacity to create precise and reliable classification models. A Random Forest is a classification made up of a collection of classifiers organized into trees. The independent random vectors are dispersed across the forest indistinguishably, and each tree selects the most common class and the key benefits are improved accuracy, robustness against outliers, speed compared to bagging and boosting, and simplicity and ease of parallel processing.
First, the random forest is formed by combining N decision trees, and predictions are made for each tree built in the beginning stages.

This works in a way,

1. Choose K data points at random from the training set.
2. Create decision trees for each data point that has been selected.
3. Repeat the procedures and select N number of decision tress to create.
4. Find the predictions for any new data points in each decision tree, then place those new data points in the category with the most support.

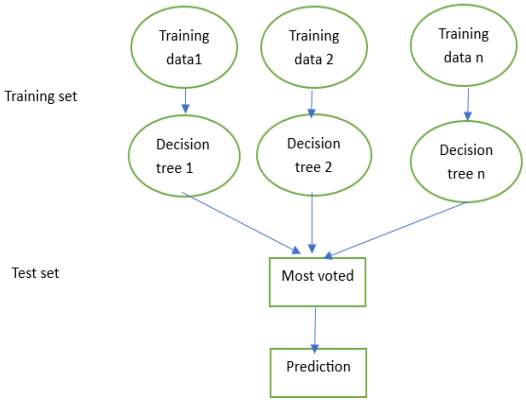

Fig 3. Random forest classifier

D. Model Evaluation

Model evaluation is the procedure of analyzing the efficiency of a machine learning model using multiple evaluation measures. In this study, evaluating different algorithms such as decision tree, logistic regression and random forest classifier.

```
Decision Tree --> 0.9
Logistic Regression --> 0.9522727272727273
RF --> 0.990909090909091
```

Fig 4. Accuracy comparison

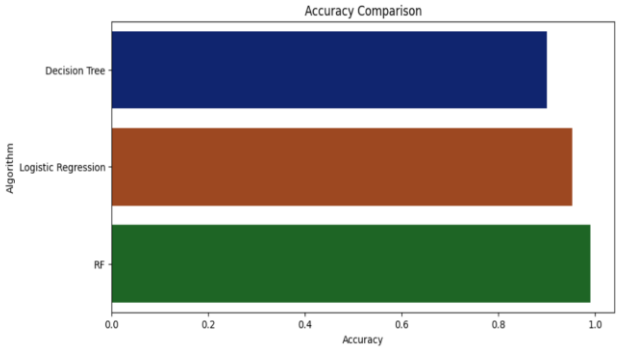

Fig 5. Accuracy comparison graph

From this evaluation result, the best method for implementation is random forest classifier which will give more accurate result that is predicts the crop efficiently than other methods.

IV. Implementation

Building model is the major step in crop prediction process which is achieved by following algorithm steps.

N	P	K	temperatu	humidity	ph	rainfall	label
90	42	43	20.87974	82.00274	6.502985	202.9355	rice
85	58	41	21.77046	80.31964	7.038096	226.6555	rice
60	55	44	23.00446	82.32076	7.840207	263.9642	rice
74	35	40	26.4911	80.15836	6.980401	242.864	rice
78	42	42	20.13017	81.60487	7.628473	262.7173	rice
69	37	42	23.05805	83.37012	7.073454	251.055	rice
69	55	38	22.70884	82.63941	5.700806	271.3249	rice
94	53	40	20.27774	82.89409	5.718627	241.9742	rice
89	54	38	24.51588	83.53522	6.685346	230.4462	rice
68	58	38	23.22397	83.03323	6.336254	221.2092	rice
91	53	40	26.52724	81.41754	5.386168	264.6149	rice
90	46	42	23.97898	81.45062	7.502834	250.0832	rice
78	58	44	26.8008	80.88685	5.108682	284.4365	rice
93	56	36	24.01498	82.05687	6.984354	185.2773	rice
94	50	37	25.66585	80.66385	6.94802	209.587	rice
60	48	39	24.28209	80.30026	7.042299	231.0863	rice
85	38	41	21.58712	82.78837	6.249051	276.6552	rice
91	35	39	23.79392	80.41818	6.97086	206.2612	rice
77	38	36	21.86525	80.1923	5.953933	224.555	rice

Fig 6. Data set used for predicting crop

1.Import the required packages and print data

```
from __future__ import print_function
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import classification_report
from sklearn import metrics
from sklearn import tree
import warnings
warnings.filterwarnings('ignore')
```

Fig 7. packages

```
[ ]  crop = pd.read_csv('/content/crop_recommendations.csv')

[ ]  crop.head()
```

Fig 8. Print data

2. Data preprocessing

split the dataset into training and testing datasets.

```
from sklearn.model_selection import train_test_split
Xtrain, Xtest, Ytrain, Ytest = train_test_split(features,target,test_size = 0.2,random_state =2)
```

Fig 9. Data preprocess

3. Initializing random forest classifier and calculating the accuracy

```
#RANDOM FOREST
from sklearn.ensemble import RandomForestClassifier

RF = RandomForestClassifier(n_estimators=20, random_state=0)
RF.fit(Xtrain,Ytrain)

predicted_values = RF.predict(Xtest)

x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('RF')
print("RF's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

Fig 10. Initializing random forest classifier

And test prediction is,

```
score = cross_val_score(RF,features,target,cv=5)
score
```

```
array([0.99772727, 0.99545455, 0.99772727, 0.99318182, 0.98863636])
```

Fig 11. Test prediction of random forest classifier

4.Print the correlation heat map

```
sns.heatmap(crop.corr(),annot=True)
```

```
<Axes: >
```

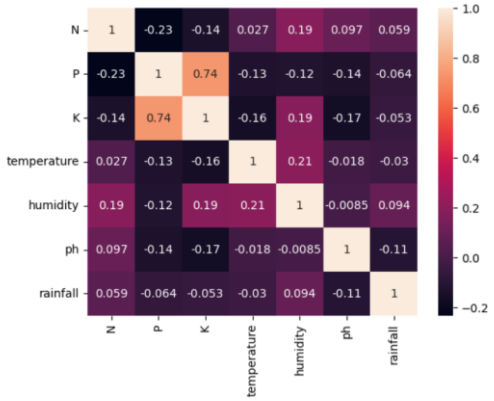

Fig 12. Heat map

V. Result

The result demonstrates that the crop forecast is based on the provided dataset. With great accuracy, Random Forest classifier deliver an accurate answer. User has to enter details as in Fig [1] such as soil attributes, humidity, rainfall, temperature. Dataset will be loaded, trained and tested and gives an accurate result.

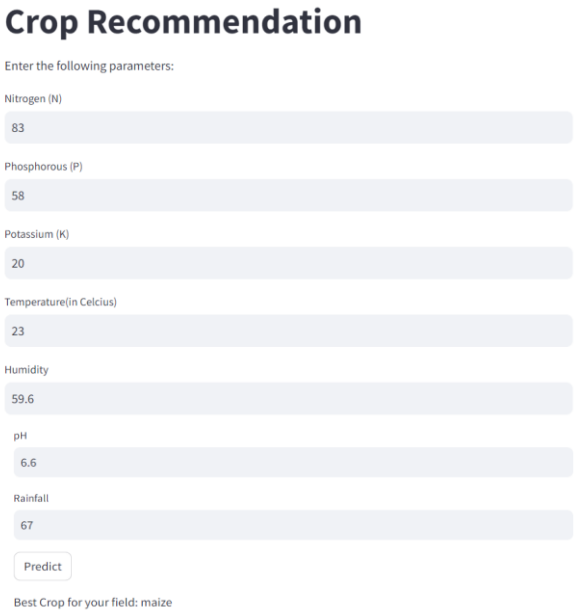

Fig 13. User Interface and result

VI. Conclusion

A crop recommendation system that considers the crop recommendation dataset with regard to the selected crops has been developed. After preprocessing the crop recommendation dataset, by using the ensemble technique this classifies the specific crops which is obtained by providing soil attributes such as Nitrogen, Phosphorus, Potassium, pH and humidity, rainfall, temperature. Random forest classifier gives accurate result and which has the high accuracy than other algorithms (99.09%). This helps the farmers to choose best crop to plant. By adding more crops this study can be extended in future.

VII. References

[1] S. Veenadhari, Dr Bharat Misra, Dr CSingh.2019."Machine learning approach for forecasting crop yield based on climatic parameters.".978-1-4799-2352- 6/14/$31.00 ©2014 IEEE.

[2] Girish L, Gangadhar S, Bharath T R, Balaji K S, Abhishek K T "Crop Yield and Rainfall Prediction in Tumakuru District using Machine

[3] S. Devadhe, A. Kausadikar, P. Daphal, A. Joshi, A. 2017, "Expert System for Crop selection. International Journal of Scientific Research in Science and Technology, Volume 3(3), pp. 436-438, 2017.

[4] Sangeetha, Shruthi, "Design and Implementation of Crop Yield Prediction Model in agriculture.", 2020, International Journal of Scientific & Technology Research, Vol. 8, Issue 04

[5] Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh (2015), International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICST)

[6] T. Banavlikar, A. Mahir, M. Budukh, S. Dhodapkar, "Crop recommendation system using Neural Networks," International Research Journal of Engineering and Technology (IRJET), vol. 5(5), pp.1475-1480, 2018.