Prediction of Crop Yield Using Machine Learning

Dona Anna Varghese
PG Scholar,
Department of Computer Applications
Amal Jyothi College of Engineering,
Kanjirappally, Kerala
donaannavarghese@mca.ajce.in

Shelly Shiju George
Assistant Professor,
Department of Computer Applications
Amal Jyothi College of Engineering,
Kanjirappally, Kerala
shellyshijugeorge@amaljyothi.ac.in

***Abstract*—Major source of India's population depends on agriculture. Researchers have been working to improve agricultural production prediction using different methods and techniques but all these methods have certain limitations. An important tool for crop yield prediction is Machine Learning. Machine Learning has turned out to be productive in data mining and agricultural studies. Some of the factors behind the reduced rate of crop production are climate and its unpredictability therefore crop yield prediction using machine learning help in increasing crop yield and production. The system takes various datasets consisting of soil moisture, physiological features of the crop, humidity, and temperature and uses various machine learning algorithms to train the model for predicting the yield. This study explains how supervised machine-learning techniques can be used to predict the yield of crops. Kaggle, a public data-gathering platform was utilized to obtain the data for this Project. With the help of Jupyter Notebook, Random Forest Regressor is used to train the model to get accurate predictions.**

Keywords: Random Forest Regressor, Ensemble Learning, Regression

I. INTRODUCTION

Predicting the crop's yield is a problematic but interesting topic. Future routes of price, security, food, and cropland extension are closely associated with forthcoming average crop yields in the world's major agricultural regions. Studies say that by 2030, worldwide cereal request for food is likely to be whole 2.8 billion(B) tonnes per year. The improvement in crop yield is the result of innovations in the agriculture sector, the use of organic fertilizer, better-improved tools, new methods of farming, and good-quality seeds. One of the interesting problems in agriculture is crop yield prediction and many models have been projected and authorized so far. This problem needs to use numerous datasets meanwhile crop yield depends on many factors such as weather, meteorological conditions, soil, use of manure, and seed variety[1]. This shows that crop yield prediction is not an irrelevant task rather it consists of difficult stages. Even though the crop yield prediction model can predict the concrete yield, an improved yield prediction is mandatory. Agriculture is one of the important sectors of the Indian economy as the demand for crops is increasing and to get more yield different technologies are being used these days. One such technology that can be used to predict the yield of crops is Machine Learning and thus producing more crops each year. Machine learning can provide improved yield prediction based on diverse features. The predictions are made using random forest regression. Random Forest Regression is a group of machine-learning techniques used to predict value across a specific range. By using Random Forest Regressor we arrive at an accurate value of yield prediction based on the gathering of a large number of past data. This makes it possible to surge the actual throughput of the agriculture sector.

II. RELATED WORKS

There are many machine learning techniques used to predict the yield of crops. Machine learning algorithms like decision trees, support vector machine, linear regression, etc are used.

Mishra et. al. [2] have discussed many machine learning methods that can be applied theoretically in various predicting fields. Their work, however, does not use any algorithms, so it is unable to provide a comprehensive understanding of the suggested work.

Manjula E [3] has tested several data mining approaches in order to estimate agricultural production for a future year. The report gave a brief review of crop yield prediction using data mining approaches based on association rules for chosen locations. Data mining is an emerging field in crop yield research.

Venugopal et. al [4] This paper uses machine learning approaches for the prediction of crops and the calculation of their yield. Random Forest classifier was used for the crop prediction for the chosen district. Executed a system for crop prediction from the collection of past data.

Klompenburg et. al [5] have deliberated on deep learning-based crop yield prediction. Studies in this paper have shown that models with additional features do not necessarily perform well in predicting yield. Models with more and fewer features should be evaluated in order to determine which one performs the best.

Sangeetha et. al [7] use machine learning which includes supervised learning models and the study
assesses the performance by using Random forest,
Polynomial Regression and Decision Tree algorithms.
The paper concludes that between all three algorithms Random forest gives a better yield prediction as related to other algorithms.

III. METHOD OF PREDICTION

In order to anticipate crop yield, data is crucial. The necessary dataset is taken through a public data-gathering platform called Kaggle, where information from various Indian states is gathered. There are two stages to the process:

DOI: 10.5281/zenodo.7961744

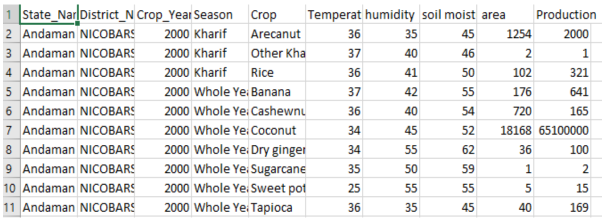
1. Training phase: The system is trained by fitting a model to the dataset's data using the selected algorithm.

2. Testing Phase: The system is tested, with inputs being fed into it and its functionality being examined.

The data utilized for training or testing must be appropriate as the precision is confirmed. Because the system is designed to forecast crop yield, an appropriate algorithm must be utilized that is more accurate. Finally, Random Forest Regressor is chosen as the suitable algorithm for prediction. The following are the main elements that have been gathered for analysis and could impact crop yield:

- State name
- District name
- Season
- Crop name
- Temperature
- Humidity
- Soil moisture
- Area
- Production

Figure 1: Sample Data



1	State_Nar	District_N	Crop_Year	Season	Crop	Temperat	humidity	soil moist	area	Production
2	Andaman	NICOBARS	2000	Kharif	Arecanut	36	35	45	1254	2000
3	Andaman	NICOBARS	2000	Kharif	Other Kha	37	40	46	2	1
4	Andaman	NICOBARS	2000	Kharif	Rice	36	41	50	102	321
5	Andaman	NICOBARS	2000	Whole Ye:	Banana	37	42	55	176	641
6	Andaman	NICOBARS	2000	Whole Ye:	Cashewnu	36	40	54	720	165
7	Andaman	NICOBARS	2000	Whole Ye:	Coconut	34	45	52	18168	65100000
8	Andaman	NICOBARS	2000	Whole Ye:	Dry ginger	34	55	62	36	100
9	Andaman	NICOBARS	2000	Whole Ye:	Sugarcane	35	50	59	1	2
10	Andaman	NICOBARS	2000	Whole Ye:	Sweet pot	25	55	55	5	15
11	Andaman	NICOBARS	2000	Whole Ye:	Tapioca	36	35	45	40	169

Techniques Used

- Random Forest Regressor

It is an ensemble learning method and is the process of using many models, trained over the same data, averaging the results of each model eventually finding a more powerful predictive result. The random forest has a lower generalization error than a single decision tree due to its randomness, decreasing the models' variance.

To predict values across a certain range machine learning technique called Regression is used. When carrying out regression using Random Forest, the forest selects the mean of the results. There are four steps in random forest regressor as follows:

1) Choose K data points at random from the training set.

2) Create a decision tree for these K data points.

3) Repeat steps 1 and 2 for as many N trees as you wish to build.

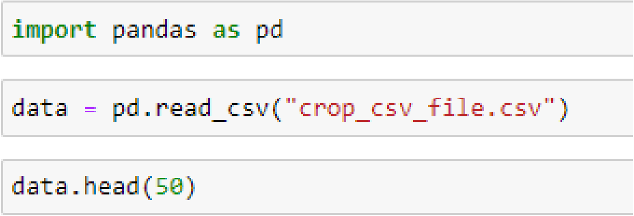
4) Make each of the Ntree trees forecast the value of Y for the data point in the question for a new data point, and then give the new data point the mean of all the expected Y values.

The sklearn module is used for training the random forest regression model, precisely the Random Forest Regressor function.

IV. BUILD MODEL

The primary stage in predicting crop yield is model building. Algorithms are used by the user when developing the model.

1) Import the libraries required and then import and print the dataset.

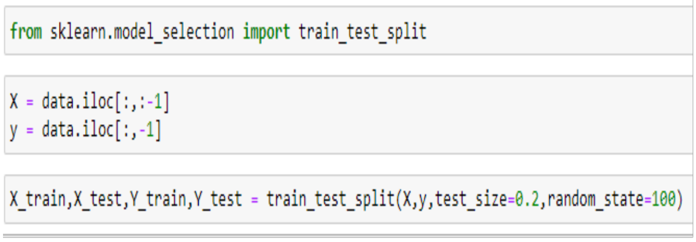


```
import pandas as pd
```

```
data = pd.read_csv("crop_csv_file.csv")
```

```
data.head(50)
```

2) Data preprocessing

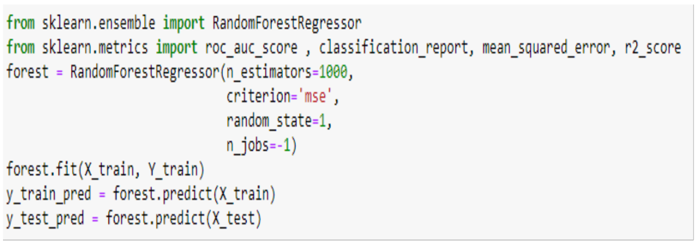


```
from sklearn.model_selection import train_test_split
```

```
X = data.iloc[:,:-1]
y = data.iloc[:,-1]
```

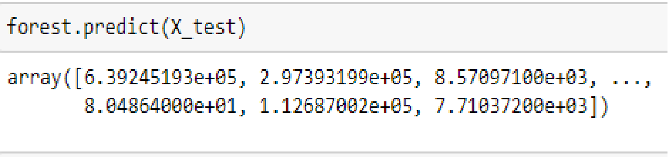
```
X_train,X_test,Y_train,Y_test = train_test_split(X,y,test_size=0.2,random_state=100)
```

3) Fit Random forest regressor to the dataset



```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import roc_auc_score , classification_report, mean_squared_error, r2_score
forest = RandomForestRegressor(n_estimators=1000,
                               criterion='mse',
                               random_state=1,
                               n_jobs=-1)
forest.fit(X_train, Y_train)
y_train_pred = forest.predict(X_train)
y_test_pred = forest.predict(X_test)
```

A test prediction is just made as follows:



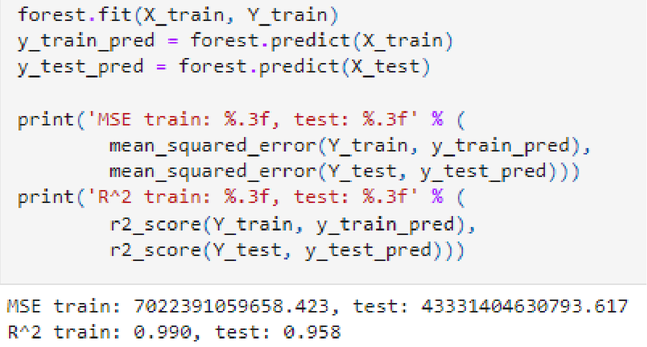
```
forest.predict(X_test)
```

```
array([6.39245193e+05, 2.97393199e+05, 8.57097100e+03, ...,
       8.04864000e+01, 1.12687002e+05, 7.71037200e+03])
```

V. RESULTS

The output is produced after fitting the data into the Random Forest Regressor algorithm. Before that, the dataset is first loaded after which data within the dataset are trained and tested.

DOI: 10.5281/zenodo.7961744

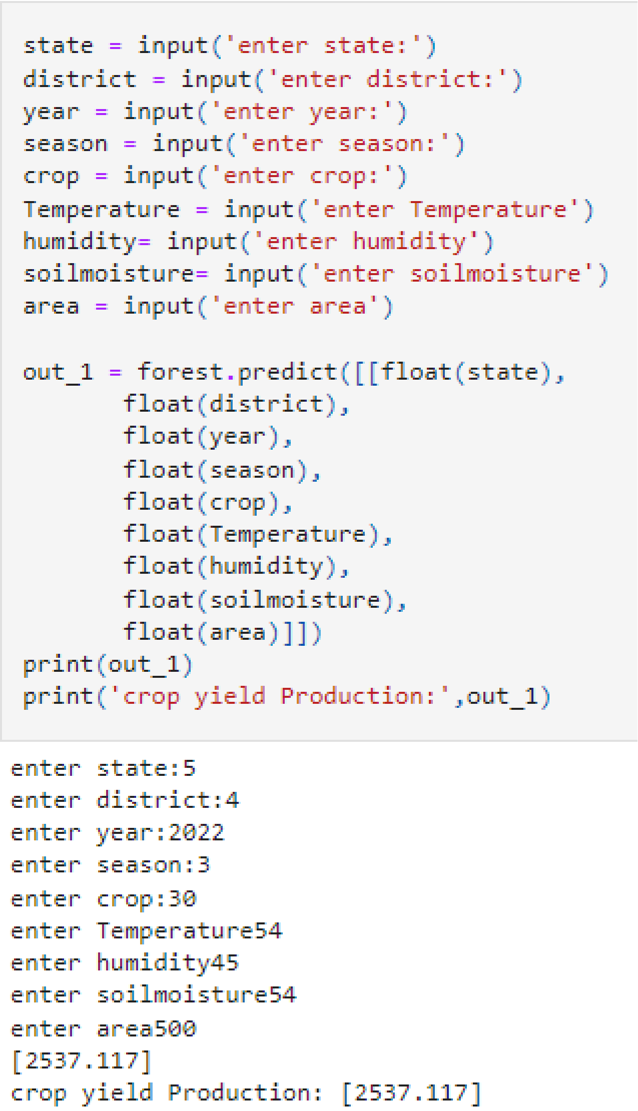


```
forest.fit(X_train, Y_train)
y_train_pred = forest.predict(X_train)
y_test_pred = forest.predict(X_test)

print('MSE train: %.3f, test: %.3f' % (
        mean_squared_error(Y_train, y_train_pred),
        mean_squared_error(Y_test, y_test_pred)))
print('R^2 train: %.3f, test: %.3f' % (
        r2_score(Y_train, y_train_pred),
        r2_score(Y_test, y_test_pred)))
```

```
MSE train: 7022391059658.423, test: 43331404630793.617
R^2 train: 0.990, test: 0.958
```

Then the yield of a crop at a particular place is predicted by getting the state name, district, year, season, name of crop, temperature, humidity, soil moisture, and area of cultivation.



```
state = input('enter state:')
district = input('enter district:')
year = input('enter year:')
season = input('enter season:')
crop = input('enter crop:')
Temperature = input('enter Temperature')
humidity= input('enter humidity')
soilmoisture= input('enter soilmoisture')
area = input('enter area')

out_1 = forest.predict([[float(state),
       float(district),
       float(year),
       float(season),
       float(crop),
       float(Temperature),
       float(humidity),
       float(soilmoisture),
       float(area)]])
print(out_1)
print('crop yield Production:',out_1)
```

```
enter state:5
enter district:4
enter year:2022
enter season:3
enter crop:30
enter Temperature54
enter humidity45
enter soilmoisture54
enter area500
[2537.117]
crop yield Production: [2537.117]
```

The entered values are evaluated based on the previous dataset and the final output which is the expected yield of the particular crop during that year is predicted. The random forest regressor has an accuracy of about 92.814 and has resulted in the exact prediction of yield for the selected district in the selected year.

VI. CONCLUSION

An essential area that contributes to a nation's expanding economy is agriculture. Therefore, it's critical that farmers are able to produce the highest possible agricultural production. To boost agricultural productivity, many machine learning approaches are used in agriculture. Random forest regression, one of many machine learning techniques, was utilized to estimate agricultural yields accurately. Two crucial elements in the prediction process were data preparation and collection. In this system, techniques for cleaning, normalizing, and standardizing data were devised so that machine learning algorithms could avoid unwanted noise. A mechanism was put in place to forecast crop yields based on the gathering of historical data. Farmers will use this forecast to assist them to choose which crop to plant.

REFERENCES

[1] X. Xu, P. Gao, X. Zhu, W. Guo, J. Ding, C. Li, “Design of an integrated climatic assessment indicator (ICAI)for wheat production”, Ecological Indicators, 2019, Vol. 101, Pages. 943-953

[2] Subhadra Mishra, Debahuti Mishra, and Gour Hari Santra, “Applications of machine learning techniques in agricultural crop prediction: a review.”, Indian Journal of Science and Technology, 2016, Vol. 9, Pages. 1-14

[3] Manjula. E, “A model for the prediction of crop yield”, International Journal of Computational Intelligence and Informatics, 2017, Vol. 6: No. 4

[4] Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Vinu Williams, “Crop yield prediction using machine learning algorithms”, National Conference on Novel & Challenging Issues and Recent Innovations in Engineering and Information Sciences, 2021, Vol. 9, Issue 13

[5] Thomas Van Klompenburg, Ayalaw Kassahun, Cagatey Catal, “Crop yield prediction using machine learning: A systematic literature review”, Computers and Electronics in Agriculture, 2020, Vol. 177

[6] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi, "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 736-741

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

[8] S.H. Bhojani, N. Bhatt, “Wheat crop yield prediction using new activation functions in neural networks”, Neural Computing and Applications, 2020, pp.1-11