Empirical Evaluation and Prediction of Crop Yield Process Using Machine Learning Methodology

Dr. Vineetha K.R, Associate Professor, Department of MCA, Nehru College of Engineering And Research Centre, vpvprakash@gmail.com

Hiba mol C, MCA student scholar, Department of MCA, Nehru College of Engineering and Research Centre, hibaanees012@gmail.com

ABSTRACT - Agriculture is the most important thing for survival. The perspective of machine learning, or ML, may be a crucial onein the search for a viable and practical solution to the crop yield issue. Given the current method, which uses satellite imagery, climate-smart pest management, and manual counting, the results aren't particularly accurate. The primary objective of this study is to forecast agricultural production using a variety of machine learning methods. Random Forest has the highest accuracy among the classifier models used in this study, followed by Logistic Regression and Naive Bayes. The predictions made by machine learning algorithms will help farmers select the crop that will yield themost by taking into account factors like temperature, rainfall, area, and other characteristics. This bridges the gap that exists between the technological and agricultural sectors

KEYWORDS: Crop Yield Prediction; Logistic Regression; Naïve Bayes; Random Forest; Weather Api

I. INTRODUCTION

Agriculture has been the primary activity of every society and civilization throughout human history, ever since it was invented. Not only does it play a significant role in the expanding economy, but it is also essential to our continued existence. Additionally, it is a crucial sector for the future of humanity and the Indian economy. Additionally, it contributes significantly to employment. The demand for production has significantly increasedover time. In order to make a lot of stuff, people use technology in totally wrong ways. New hybrid varieties are created on a daily basis. However, these varieties lack the essential components of a naturally grown crop. The soil is degraded by these synthetic methods. All of this worsens the state of the environment even more. Most ofthese unnatural strategies are employed to avoid losses.

However, when agricultural growers have access to accurate crop production data, the loss decreases. Machinelearning is a method that is rapidly expanding and supports decision-making in all fields to provide the most useful applications. Models should be examined prior to deployment for the majority of modern devices. The basic idea is to use models from machine learning to increase agricultural sector throughput. Because the numberof parameters was higher in comparison, another factor that influences the prediction is the amount of knowledgeimparted during the training period. The primary focus would be on precision agriculture, which places quality above adverse environmental factors. Numerous classifiers for machine learning, such as Logistic Regression, Naive Bayes, Random Forest, and others are used to establish a pattern so that accurate forecasts can be made and stand on the erratic patterns of temperature and rainfall. We concluded from our investigation of the aforementioned machine learning classifiers that the Random Forest approach provides the highest level of accuracy. Based on the collection of data from the past, the system forecasts crops. The data used to provide the information come from historical weather, temperature, and other variables data. A list of crops with predicted yield values that match the inputted data is displayed by our application after it runs an algorithm.

II. LITERATURE SURVEY

Aruvansh Nigam, Saksham Garg, and Archit Agrawal[1] conducted experiments using the Indian government dataset, and found that the Random Forest machine learning method provides the best yield forecast accuracy. Simple Recurrent Neural Network, a sequential model, is more effective at predicting rainfall than LSTM is at predicting temperature. For the purpose of yield forecast, the article combines variables such as rainfall, temperature, season, area, etc. When all parameters are considered, the results show that Random Forest is the best classifier.

Leo Brieman [2] is specializing in the strength, accuracy, and correlation of the random forest method. The random forest algorithm builds decision trees using several data samples, predicts the data from each subset, andthen determines the best solution for the system through voting. The data was trained in Random Forest using the bagging approach. The randomness must reduce correlation while retaining strength in order to increase accuracy.

 DOI : 10.35291/2454-9150.2023.0034

It is possible to develop a tool for farmers that will help to solve numerous issues facing the industry [3]. Farmers can use this tool to execute single or numerous tests by entering information about their crops, the time of year, and their location. The user can select a method and mine the outputs as soon as one gives the input. You can see the crop's yield rate in the outputs. The datasets contain the results of the previous year's data and have been converted into a format that is supported. Naive Bayes and KNN are the machine learning models that are employed.

Mishra [4] has provided a theoretical overview of a number of machine learning approaches that can be used indifferent forecasting contexts. Yet because they don't use any algorithms in their work, they can't really say if the suggested work is viable or not.

According to Dr. Y. Jeevan Nagendra Kumar [5], supervised learning allowsmachine learning algorithms to forecast an objective or outcome. This study focuses on supervised learning methods for predicting crop yields. It must create an acceptable function using a set of variables that may map the input variable to the desired output in order to obtain the outputs that are required. According to the paper, crop predictions may be made using the Random Forest ML method, which achieves the best accuracy value while taking into account the fewest number of models.

Information regarding crops over the past 10 years was acquired for the dataset from a number of places, including government websites. With the use of components such as soil sensors, a DHT11 sensor for temperature and humidity, and an Arduino Uno with an Atmega CPU, an Internet of Things device was set up to gather atmospheric data. Boosting method, which employs weak rule in an iterative process to deliver increased accuracy, was used to further improve Naive Bayes, a supervised learning algorithm with an accuracy of 97% [18].

III. METHODOLOGY

A. Pre-processing:

The raw data are transformed into a clean data set through a process known as data pre-processing. The data come from a variety of sources, but because they are collected in their raw form, it is not possible to analyze them. Using a variety of methods, such as substituting missing values and null values, we can transform data into a format that is easy to understand. The final step in the data pre-processing process is to divide the data intotraining and testing sets. The data typically tend to be unevenly distributed because training the model typicallyrequires as many datapoints as possible. The initial dataset used to instruct machine learning algorithms on howto learn and make accurate predictions is the training dataset, which in this instance makes up 80% of the dataset.

B. Factors affecting Crop Yield and Production:

Any crop's productivity and yield are affected by many different factors. These are basically the characteristics that make it easier to estimate how much a crop will produce each year. In this study, wetake into account temperature, rainfall, area, humidity, and wind speed

	A	B	C	D	E	F	G	H	I	J	K
1	State_N	District_	Crop_Ye	Season	Crop	Area	Production	Rainfall	Temper	Humidit	Windspeed
2	Kerala	ALAPPU	1997	Whole Y	Arecanu	2253	1518	271	24.54	79.64	1.88
3	Kerala	ALAPPU	1999	Whole Y	Arecanu	2308	1043	242.9	23.97	80.66	2.12
4	Kerala	ALAPPU	2004	Whole Y	Arecanu	2376	1006	240.5	24.28	79.87	2.05
5	Kerala	ALAPPU	2007	Whole Y	Arecanu	1696	687	290.8	24.35	79.08	1.97
6	Kerala	ALAPPU	2008	Whole Y	Arecanu	1577	955	210.4	23.98	81.34	1.87
7	Kerala	ALAPPU	2011	Whole Y	Arecanu	1615.4	659.29	252.9	24.06	80.86	1.99
8	Kerala	ERNAKU	1998	Whole Y	Arecanu	3604	1941	262.6	24.78	79.9	2.15
9	Kerala	ERNAKU	2003	Whole Y	Arecanu	5275	3813	199.6	24.48	80.6	1.89
10	Kerala	ERNAKU	2007	Whole Y	Arecanu	5207	6395	290.8	24.35	79.08	1.97
11	Kerala	ERNAKU	2010	Whole Y	Arecanu	4549.9	4889.9	261	24.54	80.84	1.99
12	Kerala	ERNAKU	2014	Whole Y	Arecanu	4133	4533	253.9	24.66	79.45	1.93
13	Kerala	IDUKKI	2005	Whole Y	Arecanu	4009	4669	252.6	24.34	82.23	2.03

Figure. 1. Pre-processed data

C. Comparison and Selection of Machine Learning Algorithm:

Before selecting the algorithm that is most suitable for this particular dataset, we must first evaluate and contrast potential algorithms. Machine learning is the most effective approach for practically resolving the crop productionissue. For agricultural yield forecasting, a variety of machine learning techniques are utilized. This paper includes the following selection and accuracy comparison machine learning techniques:

- ***Logistic regression:-*** Logistic regression is a classification approach for supervised learning that is used to forecast thelikelihood of a target variable. As the dependent variable's nature is dichotomous, there are only two viable classes. The accuracy of the logistic regression approach when applied to our dataset is 87.8%.
- ***Naive Bayes :-*** The Naive Bayes classifier makes the assumption that a certain feature's presence in a class has no bearingon the presence of any other feature. Simple to construct and especially helpful for very big data sets is the naive Bayesmodel. Even with being straightforward, Naive Bayes is known to perform better than even the most complex classification techniques. It offers a 91.50% accuracy.
- ***Random Forest*:-** Crop development in relation to current climatic conditions and biophysical change can be examined using Random Forest. The random forest algorithm builds decision trees using several data samples, predicts the data from each subset, and then determines which answer is best for the system through user voting. The bagging approach isused by Random Forest to train the data, increasing the accuracy of the outcome. RF offers a 92.81% accuracy for our data.

 DOI : 10.35291/2454-9150.2023.0034

It is evident that Random Forest provides greater accuracy than the other two algorithms out of the three.

D. Random Forest

In order for each tree to depend on the values of a random subset sampled independently and with the same distribution for all the trees in the forest, random forests are the collection of tree predictors. The bagging method was utilized by Random Forest to train the data, increasing the accuracy of the outcome. The Random Forest approach, which provides accuracy for model-based predictions and the actual results of predictions in the dataset, was utilized to achieve high accuracy. Analyzed is the model's anticipated accuracy, which is 91.34%. The random forest model's flowchart for predicting crop yield is shown in above figure.

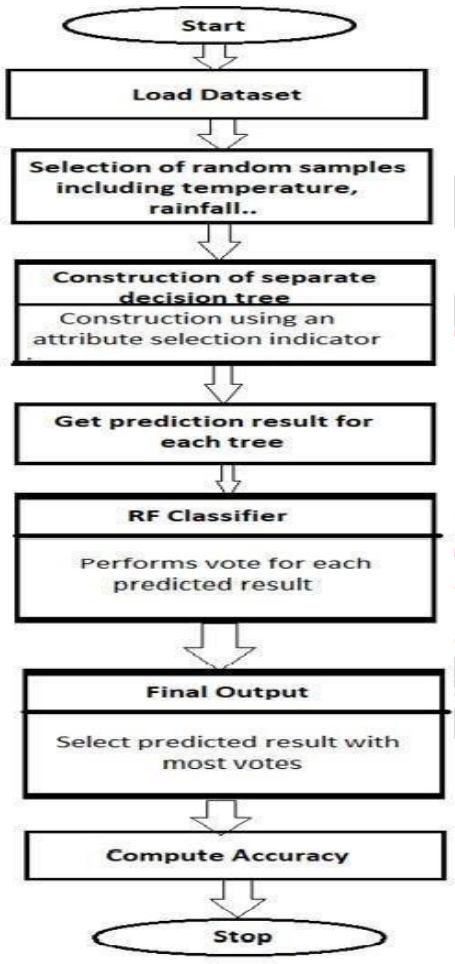

Figure 2: Flow Diagram of Crop yield Prediction Model.

A. System Architecture

The system architecture depicted in Fig. is centered on the weather API. 3, from which we obtain data on a variety of variables, including temperature, humidity, and rainfall. The data that was retrieved from the API is received by the server module. The data are stored in the database on the server. The user can provide details like their location, neighborhood, etc. through the app for mobile devices. The user can create an account on the mobile app by completing a single registration, and all of the information entered is sent to the server. To estimate crop yields, the server-installed trained Random Forest model looks at all the input and fetched data and finds the yield of the named crop in the specified location.

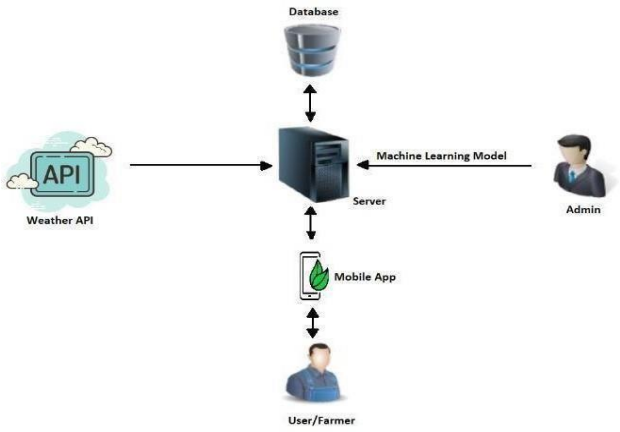

Figure .3. System Architecture

D. Proposed System

The crop's name and yield can be predicted using a mobile app that we have suggested. Numerous variables, including temperature, humidity, wind speed, rainfall, define the crop's name, and output is influenced by production and area. Predictions are made in this study using the Random Forest classifier. For crop forecasting, it will achieve the highest levels of accuracy.

E. System Analysis

- Python 3.8.5 (Jupyter Notebook): The framework for machine learning analysis is written in Python. Jupyter Notebooks provides the required output and illustrates the analysis process.
- An application programming interface called Weather API (Open Weather Map) is used to obtain a location'scurrent weather information. Current weather information required for crop prediction is displayed in the generated API key.
- The official integrated development environment (IDE) for creating Android applications is Android Studio (version 3.4.1). Java serves as the frontend design framework in this essay.

The connection between the IDE and app is made using the USB debugging mechanism.

- Python Flask Framework, 2.0.1, for example A Python microframework is called Flask. Flask is built using theJinja2 template engine and the WSGI (Web Server Gateway Interface) tools. In this study, the back-end framework for creating the application is flask. The developer can create applications without writing the low- level code necessary for protocols, thread management, etc. thanks to a collection of modules and libraries.

 DOI : 10.35291/2454-9150.2023.0034

- Heroku: Heroku is a container-based cloud platform that lets programmers create, launch, and manage only cloud-based apps. For the server portion of this essay, Heroku is used. Once a Heroku account has been created,it can be linked to a GitHub repository before being deployed

IV. RESULTS AND DISCUSSIONS

This paper uses machine learning approaches to improve crop output. The method that produces high accuracy forecasted the yield of the correct crop. Using input libraries like Scikit-Learn, Numpy, Keras, and Pandas, Python 3.8.5 (Jupyter Notebook) is used to implement the machine learning algorithms. An Android application that was created questioned the outcomes of the machine learning analysis. The crop name and accompanying yield were displayed through an Android app built with Flutter.

A. Datasets Used

The datasets were retrieved from many official government websites, including:

- data.gov.in-Details regarding area, production, crop name[8].
- indianwaterportal.org -Depicts rainfall details[9].
- power.larc.nasa.in –Temperature, humidity, wind speed details[10].

Whole dataset has 4261 occurrences. It contains information on fourteen districts in Kerala, such as crop name, area, productivity,temperature, rainfall, humidity, and wind speed. The pre-processing of the data produced the precise dataset that was required. Figure 4 illustrates a heat map that was used to represent the many features present.

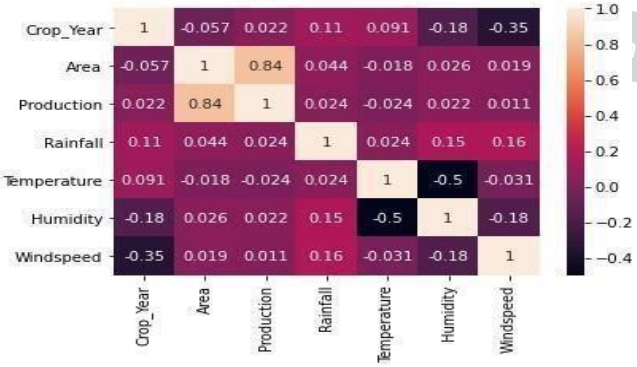

Figure. 4. Heat Map

B. Classifiers Used

Logistic Regression, Random Forest, and Nave Bayes are examples of machine learning classifiers that were used foraccuracy comparison and prediction. On the dataset, these three classifiers were trained.

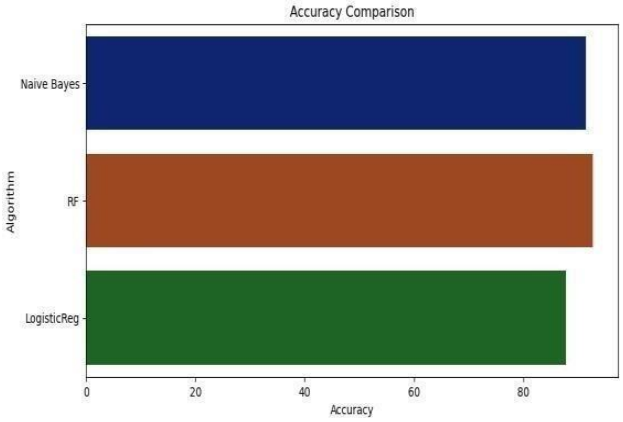

Figure. 5. Comparison Plot

and a comparison graph was drawn to highlight how well the models performed. The models' performance is shownin Fig. 5. The most accurate classifier out of the three utilised was Random Forest.

C. Weather API Used

Using the weather API allowed access to the necessary location's most recent weather information. "OpenWeatherMap" was the API used to retrieve the weather data.

```
Enter the city name: kottayam
-------------------------------------------------------------
Weather Stats for - KOTTAYAM  || 16 May 2021 | 01:03:56 PM
-------------------------------------------------------------
Current temperature is: 26.00 deg C
Current weather desc  : moderate rain
Current Humidity      : 94 %
Current wind speed    : 1.54 kmph
Current rainfall      : 90
```

Figure.6.API Output

With the obtained API key, the district name was entered to access the required metrological parameters, such as near- surface elements like temperature, wind speed, humidity, and precipitation. The current weather description for the entered location is shown in Fig. 6.

D. Crop Name Prediction

The crop that can be cultivated in a certain district at a given period was predicted using the Random Forest Classifier, which had the highest accuracy.

ALGORITHM	ACCURACY
RANDOM FOREST	92.81407991690006
NAÏVE BAYES	91.49621790098573
LOGISTIC REGRESSION	87.82982929223341

Table I: Accuracy Table

Based on production, area, temperature, humidity, rainfall, and wind speed, the names of Kerala's abundantly growing crops were chosen, and their yields were calculated. On the pre-processed dataset, the Random Forest classifier was

 DOI : 10.35291/2454-9150.2023.0034

trained. Using API-accessible instant weather data for the selected district, predictions were made. The appropriate crop prediction was generated by the trained model for the selected district.

E. Crop Yield Calculation

The crop predicted by the Random Forest Classifier was mapped to its actual production. The crop yield was then calculated by dividing the user-entered area by the amount of production [1]. Using the estimated yield foreach crop's name, farmers can select the optimal planting time for the highest yield.

Yield= Production/Area

V. CONCLUSION AND FUTURE WORK

The primary focus of this research is the application of machine learning algorithms to crop forecasting and yield estimation. There are a variety of machine learning methods used to calculate accuracy. The Random Forest classifierwas utilized in the crop prediction for the selected district. Developed a method for anticipating crops using data from the past. Farmers can choose which crop to plant in the field with more ease using the suggested method. This workis done to learn more about the crops that can be used to harvest things in a way that is both efficient and useful. Accurate crop forecasts for many districts will be beneficial to farmers in many districts in Kerala. Consequently, crop yield rates are maximized to the benefit of our Indian economy.

It can be implemented and interfaced with a versatile and multi-skilled application, and it has a significant potential extension. The farmers would receive clear information on their mobile devices on the highest crop output as a result of their need for education. The development of an application that farmerscan use as an app and the translation of the entire system into their regional language would be the focus of our futurework . The following functionality can be added to the work to make it more complete. By sharing images of farms, this application can be created to assist farmers. Image processing is used to detect crop illnesses, and the user is provided with pesticides based on disease images. Farms should use a smart irrigation system to increase productivity. With this, the task may be controlled at that same moment without losing any money, even if the rancher is at home. The agricultural sector will experience significant advancements that will further aid farmers in crop output.

REFERENCES

[1] Aruvansh Nigam, Saksham Garg, Archit Agrawal "Crop Yield Prediction using ML Algorithms",2019

[2] Leo Brieman, "Random Forests", 2001

[3] Medar R, Rajpurohit V S and Shweta S 2019 Crop yield prediction using machine learning techniques IEEE 5th International Conference for Convergence in Technology (I2CT) pp 1-5 doi: 10.1109/I2CT45611.2019.9033611. Mishra, S., Mishra, D., Santra, G. H. "Applications of machine learning techniques in agricultural crop production",2016

[4] Dr. Y Jeevan Kumar, "Supervised Learning Approach for Crop Production",2020

[5] Ramesh Medar, Vijay S, Shweta, "Crop Yield Prediction using Machine Learning Techniques", 2019

[6] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa, "Machine Learning Methodologies for Paddy Yield Estimation in India: A CASE STUDY", 2019

[7] Sangeeta, Shruthi G, "Design And Implementation Of Crop Yield Prediction Model InAgriculture",2020

[8] https://www.data.gov.in

[10]https://power.larc.nasa.gov/data-access-viewer/

[11]https://en.wikipedia.org/wiki/Agriculture

[12]https://www.ibm.com/weather

[13]https;//flutter.dev

[14]https://openweathermap.org

[15] https;//builtin.com/data-science/random-forest-algorithm

[16] https://tutorialspoint/machine-learning/logistic-regression

[17] http://scikit-learn.org/modules/naive-bayes

[18] Kalimuthu M, Vaishnavi P and Kishore M 2020 Crop prediction using machine learning 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) pp 926-32 doi: 10.1109/ICSSIT48917.2020.9214190.

 DOI : 10.35291/2454-9150.2023.0034