Crop Yield Prediction using Machine Learning Algorithms

Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Prof. Vinu Williams
Department of Computer Science and Engineering
College of Engineering, Kidangoor
Kottayam, India

Abstract— Agriculture is first and foremost factor which is important for survival. Machine learning (ML) could be a crucial perspective for acquiring real-world and operative solution for crop yield issue. Considering the present system including manual counting, climate smart pest management and satellite imagery, the result obtained aren't really accurate. This paper focuses mainly on predicting the yield of the crop by applying various machine learning techniques. The classifier models used here include Logistic Regression, Naïve Bayes and Random Forest, out of which the Random Forest provides maximum accuracy. The prediction made by machine learning algorithms will help the farmers to come to a decision which crop to grow to induce the most yield by considering factors like temperature, rainfall, area, etc. This bridges the gap between technology and agriculture sector.

Keywords—Crop_yield_prediction; logistic_regression; naïve bayes; random forest; weather_api

I. INTRODUCTION

Agriculture, since its invention and inception, be the prime and pre-eminent activity of every culture and civilization throughout the history of mankind. It is not only an enormous aspect of the growing economy, but it's essential for us to survive. It's also a crucial sector for Indian economy and also human future. It also contributes an outsized portion of employment. Because the time passes the requirement for production has been increased exponentially. So as to produce in mass quantity people are using technology in an exceedingly wrong way. New sorts of hybrid varieties are produced day by day. However, these varieties don't provide the essential contents as naturally produced crop. These unnatural techniques spoil the soil. It all ends up in further environmental harm. Most of these unnatural techniques are wont to avoid losses.

But when the producers of the crops know the accurate information on the crop yield it minimizes the loss. Machine learning, a fast-growing approach that's spreading out and helping every sector in making viable decisions to create the foremost of its applications. Most devices nowadays are facilitated by models being analyzed before deployment. The main concept is to increase the throughput of the agriculture sector with the Machine Learning models. Another factor that also affects the prediction is the amount of knowledge that's being given within the training period, as the number of parameters was higher comparatively. The core emphasis would be on precision agriculture, where quality is ensured over undesirable environmental factors. So as to perform accurate prediction and stand on the inconsistent trends in temperature and rainfall various machine learning classifiers like Logistic Regression, Naïve Bayes, Random Forest etc. are applied to urge a pattern. By applying the above machine learning classifiers, we came into a conclusion that Random Forest algorithm provides the foremost accurate value. System predicts crop prediction from the gathering of past data. Using past information on weather, temperature and a number of other factors the information is given. The Application which we developed, runs the algorithm and shows the list of crops suitable for entered data with predicted yield value.

II. LITERATURE SURVEY

Aruvansh Nigam, Saksham Garg, Archit Agrawal[1] conducted experiments on Indian government dataset and it's been established that Random Forest machine learning algorithm gives the best yield prediction accuracy. Sequential model that's Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction. The paper puts factors like rainfall, temperature, season, area etc. together for yield prediction. Results reveals that Random Forest is the best classifier when all parameters are combined.

Leo Brieman [2] **,** is specializing in the accuracy and strength & correlation of random forest algorithm. Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better the answer for the system. Random Forest used the bagging method to trained the data. To boost the accuracy, the randomness injected has to minimize the correlation ρ while maintaining strength.

Balamurugan [3], have implemented crop yield prediction by using only the random forest classifier. Various features like rainfall, temperature and season were taken into account to predict the crop yield. Other machine learning algorithms were not applied to the datasets. With the absence of other algorithms, comparison and quantification were missing thus unable to provide the apt algorithm.

Mishra [4], has theoretically described various machine learning techniques that can be applied in various forecasting areas. However, their work fails to implement any algorithms and thus cannot provide a clear insight into the practicality of the proposed work.

Dr. Y. Jeevan Nagendra Kumar [5], have concluded Machine Learning algorithms can predict a target/outcome by using Supervised Learning. This paper focuses on supervised learning techniques for crop yield prediction. To get the

specified outputs it needs to generate an appropriate function by set of some variables which can map the input variable to the aim output. The paper conveys that the predictions can be done by Random Forest ML algorithm which attain the crop prediction with best accurate value by considering least number of models.

III. METHODOLOGY

A. *Data Pre-Processing*

Data Preprocessing is a method that is used to convert the raw data into a clean data set. The data are gathered from different sources, it is collected in raw format which is not feasible for the analysis. By applying different techniques like replacing missing values and null values, we can transform data into an understandable format. The final step on data preprocessing is the splitting of training and testing data. The data usually tend to be split unequally because training the model usually requires as much data-points as possible. The training dataset is the initial dataset used to train ML algorithms to learn and produce right predictions (Here 80% of dataset is taken as training dataset). Fig.1. shows the few rows of the preprocessed data.

B. Factors affecting Crop Yield and Production

There are a lot of factors that affects the yield of any crop and its production. These are basically the features that help in predicting the production of any crop over the year. In this paper we include factors like Temperature, Rainfall, Area, Humidity and Windspeed (Fig.1 shows the attributes for the crop name prediction and its yield calculation).

	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

Fig. 1. Preprocessed data

C. Comparison and Selection of Machine Learning Algorithm

Before deciding on an algorithm to use, first we need to evaluate and compare, then choose the best one that fits this specific dataset. Machine Learning is the best technique which gives a better practical solution to crop yield problem. There are a lot of machine learning algorithms used for predicting the crop yield. In this paper we include the following machine learning algorithms for selection and accuracy comparison :

- .*Logistic Regression*:- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. When logistic regression algorithm applied on our dataset it provides an accuracy of 87.8%.
- *Naive Bayes:*- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. It provides an accuracy of 91.50%.
- *Random Forest*:- Random Forest has the ability to analyze crop growth related to the current climatic conditions and biophysical change. Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better solution for the system. Random Forest uses the bagging method to train the data which increases the accuracy of the result. For our data, RF provides an accuracy of 92.81%.

It is clear that among all the three algorithms, Random forest gives the better accuracy as compared to other algorithms.

D. Random Forest Model for Crop Prediction

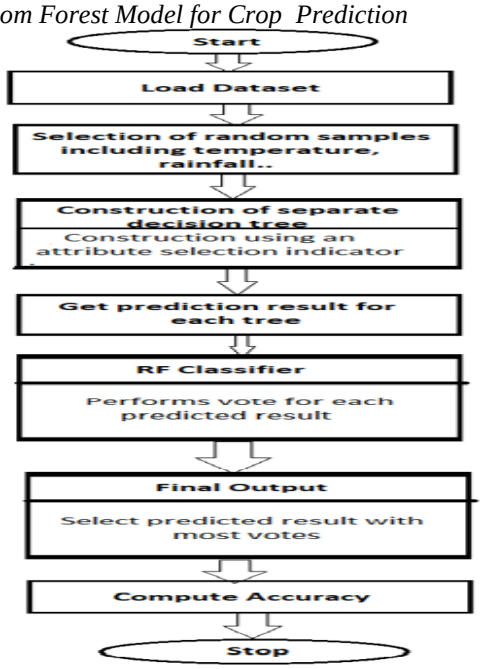

Fig. 2. Flowchart for Random Forest Model

Random forests are the aggregation of tree predictors in such a way that each tree depends on the values of a random subset sampled independently and with the same distribution for all trees in the forest. Random Forest used the bagging method to trained the data which increases the accuracy of the result. For getting high accuracy we used the Random Forest algorithm which gives accuracy which predicate by model and actual outcome of predication in the dataset. The predicted accuracy of the model is analyzed 91.34%. Fig.2 shows the flowchart of random forest model for crop yield prediction.

E. *System Architecture*

System architecture represented in the Fig.3 mainly consists of weather API where we fetch the data such as temperature, humidity, rainfall etc. The data fetched from the API are sent to the server module. The data gets stored on to the database on the server. Using the mobile application, the user can provide details like location, area, etc. The user can create an account on the mobile app by one-time registration

and all these entered data are sent to server. The trained Random forest model deployed on the server uses all the fetched and input data for crop yield prediction, finds the yield of predicted crop with its name in the particular area.

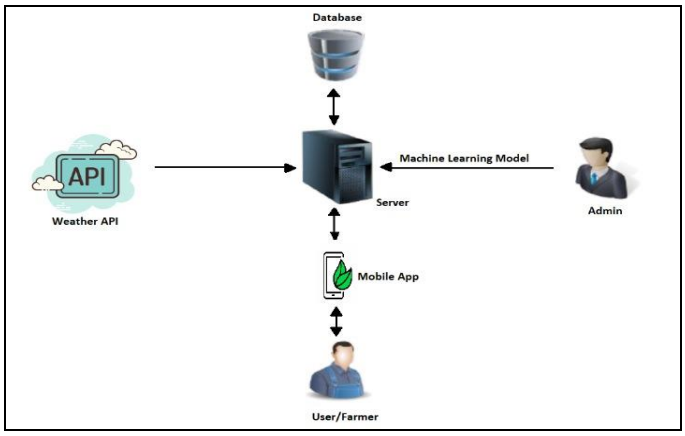

Fig. 3. System Architecture

F. *Proposed System*

Our proposed system system is a mobile application which predicts name of the crop as well as calculate its corresponding yield. Name of the crop is determined by several features like temperature, humidity, wind-speed, rainfall etc. and yield is determined by the area and production. In this paper, Random Forest classifier is used for prediction. It will attain the crop prediction with best accurate values.

G. *System Analysis*

- *Python 3.8.5(Jupyter Notebook):*Python is the coding language used as the platform for machine learning analysis. Jupyter Notebooks illustrates the analysis process and gives out the needed result.

- *Weather_API (Open Weather Map):* Weather API is an application programming interface used to access the current weather details of a location. The generated API key illustrates current weather forecast needed for crop prediction.

- *Android Studio* (Version 3.4.1): Android Studio is the official integrated development environment (IDE) for Android application development. This paper uses java as the framework for frontend designing. USB debugging method is used for the connection of IDE and app.

- *Python Flask Framework* (Version 2.0.1): Flask is a micro framework in python. Flask is based on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine. In this paper flask is used as the back-end framework for building the application. It is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc.

- *Heroku*: Heroku is the container-based cloud platform that allows developers to build, run & operate applications exclusively in the cloud. In this paper Heroku is used for server part. Once created an account in the Heroku we can connect it with the GitHub repository and then deploy.

IV. RESULTS AND DISCUSSIONS

This paper reinforces the crop production with the aid of machine learning techniques. The technique which results in high accuracy predicted the right crop with its yield. The machine learning algorithms are implemented on Python 3.8.5(Jupyter Notebook) having input libraries such as Scikit-Learn, Numpy, Keras, Pandas. Developed Android application queried the results of machine learning analysis. Flutter based Android app portrayed crop name and its corresponding yield.

A. *Datasets Used*

The datasets have been obtained from different official Government websites:

- 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].

Combined dataset has 4261 instances. It includes features like crop name, area, production, temperature, rainfall, humidity and wind speed of fourteen districts in Kerala. The data pre-processing phase resulted in needed accurate dataset. Fig. 4. shows a heat map used to portray the individual attributes contained in.

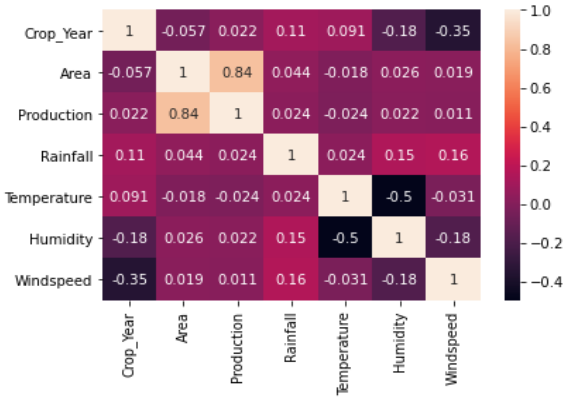

Fig. 4. Heat Map

B. *Classifiers Used*

Machine learning classifiers used for accuracy comparison and prediction were Logistic Regression, Random Forest and Naïve Bayes. These three classifiers were trained on the dataset

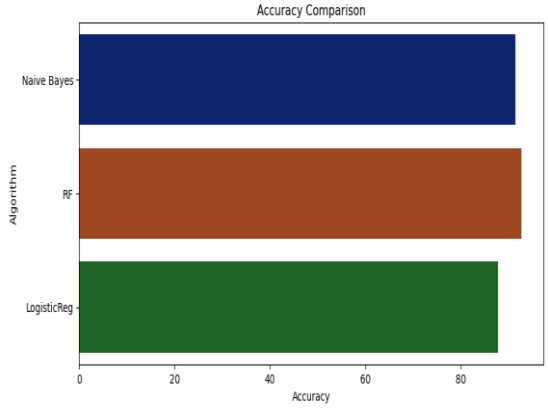

Fig. 5.Comparison Plot

and a comparison graph was plotted to showcase the performance of the models. Fig.5 showcase the performance of the models. Of the three classifiers used, Random Forest resulted in high accuracy.

C. Weather_ API Used

Weather _ API usage provided current weather data access for the required location. For retrieving the weather data used API

```
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
```

Fig. 6. API Output

was “OpenWeatherMap”. By entering the district name, needed metrological factors such as near surface elements which include temperature, wind speed, humidity, precipitation were accessed by using generated API key. Fig.6. depicts current weather description for entered location.

After the training of dataset, API data was given as input to illustrate the crop name with its yield.

D. Crop Name Prediction

Random Forest Classifier having the highest accuracy was used as the midway to predict the crop that can be grown on a selected district at the respective time.

ALGORITHM	ACCURACY
RANDOM FOREST	92.81407991690006
NAÏVE BAYES	91.49621790098573
LOGISTIC REGRESSION	87.82982929223341

Table I : Accuracy Table

Abundantly growing crops in Kerala were chosen and their name was predicted and yield was calculated on the basis of area, production, temperature, humidity, rainfall and wind speed. The preprocessed dataset was trained using Random Forest classifier. Chosen district’s instant weather data accessed from API was used for prediction. Trained model resulted in right crop prediction for the selected district.

E. Crop Yield Calculation

The crop which was predicted by the Random Forest Classifier was mapped to the production of predicted crop. Then the area entered by the user was divide from the production to get crop yield[1].

Yield= Production/Area

Crop name predicted with their respective yield helps farmers to decide correct time to grow the right crop to yield maximum result.

F. Android Application

An Android app has been developed to query the results of machine learning analysis. The app is compatible with Android OS version 7. The pages were written in Java language. The app has a simple, easy-to-use interface requiring only few taps to retrieve desired results. Just only giving the location and area of the field the Android app gives the name of right crop to grown there.
By accessing the user entered details, app will queries the machine learning analysis. Using the location, API will give out details of weather data. The retrieved weather data get acquired by machine learning classifier to predict the crop and calculate the yield. The output is then fetched by the server to portray the result in application.
The main activities in the application were account creation, detail_entry and results_fetch. The account_creation helps the user to actively interact with application interface. The user fill the field in home page to move onto the results activity. The retrieved data passed to machine learning model and crop name is predicted with calculated yield value.

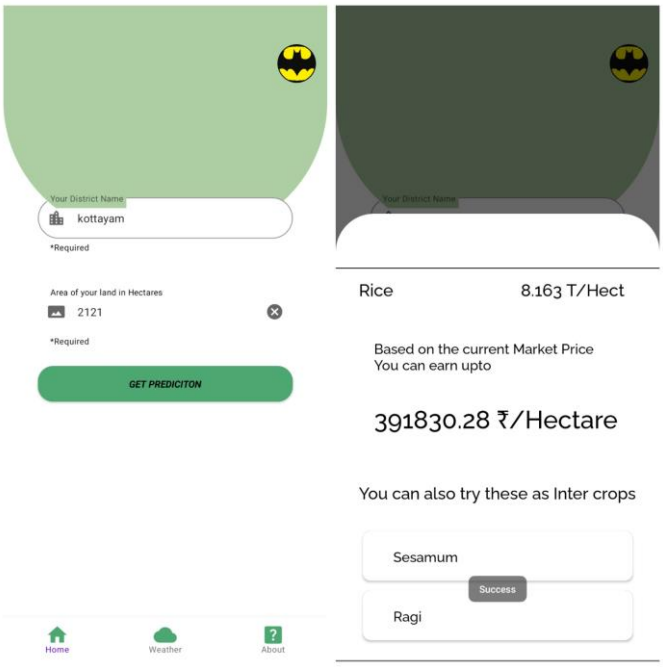

Fig. 7. Home Screen

V. CONCLUSION

This paper focuses on the prediction of crop and calculation of its yield with the help of machine learning techniques. Several machine learning methodologies used for the calculation of accuracy. Random Forest classifier was used for the crop prediction for chosen district. Implemented a system to crop prediction from the collection of past data. The proposed technique helps farmers in decision making of which crop to cultivate in the field. This work is employed to search out the gain knowledge about the crop that can be deployed to make an efficient and useful harvesting. The accurate prediction of different specified crops across different districts will help farmers of Kerala. This improves our Indian economy by maximizing the yield rate of crop production.

VI. FUTURE SCOPE

In coming years, can try applying data independent system. That is whatever be the format our system should work with same accuracy. Integrating soil details to the system is an advantage, as for the selection of crops knowledge on soil is also a parameter. Proper irrigation is also a needed feature crop cultivation. In reference to rainfall can depict whether extra water availability is needed or not. This research work can be enhanced to higher level by availing it to whole India.

REFERENCES

[1] Aruvansh Nigam, Saksham Garg, Archit Agrawal “Crop Yield Prediction using ML Algorithms “,2019
[2] Leo Brieman, “Random Forests”, 2001
[3] Priya, P., Muthaiah, U., Balamurugan, M.”Predicting Yield of the Crop Using Machine Learning Algorithm”,2015
[4] Mishra, S., Mishra, D., Santra, G. H.,“Applications of machine learning techniques in agricultural crop production”,2016
[5] Dr.Y Jeevan Kumar,”Supervised Learning Approach for Crop Production”,2020
[6] Ramesh Medar,Vijay S, Shweta, “Crop Yield Prediction using Machine Learning Techniques”, 2019
[7] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa,”Machine Learning Methodologies for Paddy Yield Estimation in India: A CASE STUDY”, 2019
[8] Sangeeta, Shruthi G, “Design And Implementation Of Crop Yield Prediction Model In Agriculture”,2020
[9] 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