Crop Recommendation System using hybrid of KNN and Random Forest Classifier

Mrs.P.Gajalakshmi[1], Aruna Cathciyal. G[2], Sri Amirtha. K[3], Viji. D[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Pondicherry 605107,India
[2,3,4]B.Tech. Students, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Pondicherry 605107,India

Abstract

Machine learning and its rapid advancement have significantly improved the way we interact with computers. We can find applications of machine learning in almost every field, like the IT industry, medicine, agriculture, etc. The idea of imparting machine learning to agriculture rose decades ago, and, as a result, many improvements were made in the field of agriculture. Various models are developed to predict the crop and yield using machine learning algorithms like decision trees, but the main problem with using algorithms like decision trees is that they do not provide the desired accuracy, which may lead to incorrect predictions. This paper proposes a user-friendly crop recommendation andyield prediction system. The user provides the following as input: state name, district name, soil type, and season. To recommend the crop and predict the yield of the crop,a combination of K-nearest neighbor (KNN) and randomforest (RM) is used. The K-nearest neighbor algorithm showed 98% accuracy, and the Random Forest algorithmshowed 96% accuracy.

Keywords: Crop recommendation, yield prediction, Machine learning, KNN, Random Forest

I. INTRODUCTION

From ancient period, agriculture is considered as the mainand the foremost culture practiced in India. Ancient people cultivate the crops in their own land and so they have been accommodated to their needs. Therefore, the natural crops arecultivated and have been used by many creatures such as human beings, animals and birds. The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention peopleare been concentrated on cultivating artificial products that ishybrid products where there leads to an unhealthy life. Nowadays, modern people don’t have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climaticconditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food.

By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economic growthin the field of agriculture. There are multiple ways to increaseand improve the crop yield and the quality of the crops. Datamining also useful for predicting the crop yield production.

Crop yield prediction is an important agricultural problem. Each and Every farmer is always tries to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depending on weather conditions, pests and planningof harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. Unfortunately, there areonly very few information that are available regarding whichcrop is harvest in which season and which crop will provide you the maximum yield.

This paper, proposes a model which addresses these issues. The aim of this proposed system is to help and guide farmersin order to maximize the yield of the crop and also suggest which crop is suitable for which season which will again improves the yield of the crop. The proposed system recommends crop based on environmental condition and nutrition content of the soil, and benefit to maximize the cropyield that will help to meet the increasing demand for the country's food supplies. The proposed model recommends the crop and predict the yield of the crop by taking into consideration various factors like temperature, humidity level, season, area, soil etc.

II. LITERATURE REVIEW

In the following sections, we present the works that havebeen performed using various existing algorithms for crop yield prediction.

[1] **Nishant, Potnuru Sai, Pinapa Sai Venkat, Bollu Lakshmi Avinash, and B. Jabber** proposed a model for Crop Yield Prediction based on Indian Agriculture using Machine Learning. This paper predicts the yield of almostallkinds of crops that are planted in India. This paper uses advanced regression techniques like Kernel Ridge, Lasso, andENet algorithms to predict the yield and uses the concept of Stacking Regression further we used stacking of these models to minimize the error and to obtain betterpredictions.

[2] **Priya P, Muthaiah U, and Balamurugan M** proposed a model for predicting yield of the crop using a Machine learning algorithm. This paper involves predicting the yield of the crop from available historical available data like weather parameters, soil parameters, and historic crop yield. This paper focuses on predicting the yield of the crop based on the existing data by using the Random Forest algorithm. Real data from Tamil Nadu were used for buildingthe models and the models were tested with samples. The prediction will help the farmer to predict the yield of the cropbefore cultivating it in the agriculture field. To predict the crop yield in the future accurately Random Forest, a most powerful and popular supervised machine learning algorithmis used.

[3] **S. Pavani and Augusta sophy Beulet P** proposed the Heuristic Prediction of Crop Yield using a Machine Learning Technique which uses Machine learning-based solutions developed to solve the difficulties faced by the farmers being discussed in this work. The proposed model predicts plant eld and gives reasonable crop yield suggestionsfor particular districts in Telangana. The real- time environmental parameters of Telangana District like soil moisture, temperature, rainfall, humidity are collected and crop yield is being predicted using KNN Algorithm.

[4] **Kale, Shivani. S, and Preeti S. Patil** published "A Machine Learning Approach to Predict Crop Yield and Success Rate". The model was developed using a Multilayer perceptron neural network. Initially, the result was obtained considering the optimizer RMS prop with an accuracy of 45
%, later it was enhanced to 90% by increasing layers, adjusting weight, and bias, and changing the optimizer to Adam. This research describes the development of a differentcrop yield prediction model with ANN, with 3Layer Neural Network. The ANN model develops a formulato ascertain therelationship using a large number of input and output examples, to establish a model for yield predictions an Activation function: Rectified Linear activation unit (Relu) isused. The backward and forward propagation techniques were used.

[5] In this paper, **Manjula E, and Djodiltachoumy S** have presented "A model for prediction of crop yield". This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on the creation of a prediction model which may be used to the future prediction of crop yield. This paper presents a brief analysis of crop yield prediction using data mining techniques based on association rules for the selected region
i.e. district of Tamil Nadu in India. The experimental results show that the proposed work efficiently predicts crop yield production.

[6] **Veenadhari S, Misra B, Singh CD** proposed a Machine learning approach for forecasting crop yield based on climatic parameters. The study was aimed to develop a website for finding out the influence of climatic parameters on crop production in selected districts of Madhya Pradesh. They developed a user-friendly website and the accuracy of predictions are above 75 per cent in all the crops and districtsselected in the study indicating higher accuracy of prediction. The user-friendly web page developed for predicting crop yield can be used by any user their choice ofcrop by providing climatic data of that place. C4.5 algorithmwas used.

III. EXISTING SYSTEM

The existing system proposes a viable and user-friendly yield prediction system for farmers. The proposed system a user-friendly interface that will recommend crop based on user inputs like season, area, district etc. In the existing system they have proposed an crop recommendation system which will recommend crop using a machine learning algorithm. The output provided by the machine learningalgorithm will be showed as the input to the user

To predict the crop yield, selected Machine Learning algorithms such decision tree is used which is a classifier algorithm. The input provided by the user will be fed into themachine learning algorithm in this case decision tree which will recommend the crop by analyzing various inputs provided by the user. The decision tree algorithm showed accuracy of 95%

Additionally, In the existing system they have proposed an mobile application version that can be used in any smartphone which makes sure that in-order to use their application the user do not need to own any fancy device likelaptop any with a basic smart phone can utilize their application

The major contributions of the paper are enlisted below,

Prediction of the crop yield for specific regions by executingvarious Machine Learning algorithms, with a comparison of error rate and accuracy.

A user-friendly mobile application to recommend the most profitable crop.

An economical crop recommendation system that can be accessed by anyone who has access to internet

A recommender system that uses machine learning algorithmlike decision tree.

Issues in the existing system

The existing system deals with proposing an automatic crop recommendation system using machine learning techniques like decision tree.

One of the main issues in the existing system is that it canpredict only for a particular state in India and also the modeldoes not take into consideration various factors like nutrient content of the soil, humidity etc.

The model only aims in recommending crops and does not give the crop yield under that climatic condition.

PROPOSED SYSTEM

The task crop recommendation system is of greater importance in the field of agriculture. Therefore, developing a crop recommendation system has gained much interest in the field of research nowadays.

The task of this crop recommendation system is to recommend crop type and yield of the crop for various statesof India based on a variety of factors like nitrogen level of thesoil, phosphorous level of the soil, potassium level of the soil,season, and district. The current paper focus on Machine learning which includes supervised learning models like KNN and Random Forest. To recommend the crop a K- Nearest Neighbor algorithm is used, the output from the KNNis fed to the Random Forest which will provide us the desiredoutcome. System architecture is one which defines the Conceptual model in different structures and multiple views of the system. Fig. 1 shows architecture design of proposed system of the project.

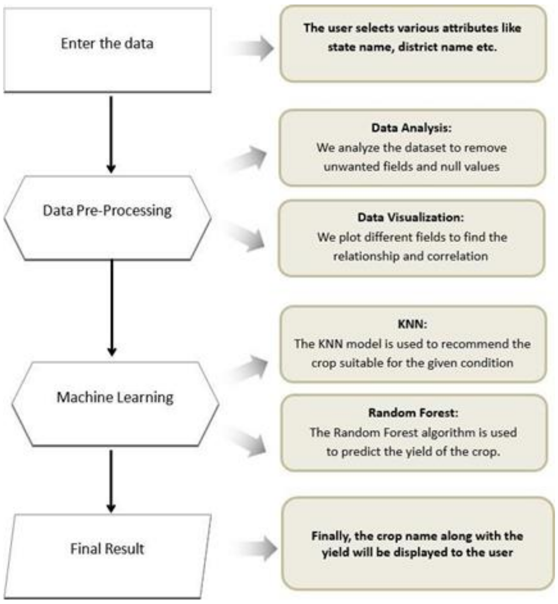

Enter the data
The user selects various attributes like state name, district name etc.
Data Pre-Processing
Data Analysis:
We analyze the dataset to remove unwanted fields and null values
Data Visualization:
We plot different fields to find the relationship and correlation
Machine Learning
KNN:
The KNN model is used to recommend the crop suitable for the given condition
Random Forest:
The Random Forest algorithm is used to predict the yield of the crop.
Final Result
Finally, the crop name along with the yield will be displayed to the user

Fig. 1. Process flow Architecture

Above architecture clearly explains about how the components of the system communicate among themselves starting from data entry by user. This proposed framework isable to finding out the crop and its yield. This model gives clear picture of huge amount of data capture and preprocessing of data to remove the unwanted data such as NULL etc presented in it. During preprocessing step, we splitthe dataset into training and testing dataset. Train dataset to detect the crop yield present in the dataset using appropriate supervised learning algorithms. Apply the machine learning techniques which are helpful for finding crop and its yield forany new data occurred in the data. After this data acquisitionsuitable machine learning algorithm must be applied to compute efficiency and capability of the model, here we haveapplied various machine learning algorithms like KNN and Random Forest.

Packages that we have used while developing the ML modelare:
Scikit: *Used for data pre-processing*
Pandas: *For performing certain operations on data*
Matplotlib: *For Data Visualization part*
Dataset:
This machine learning model is built using benchmark dataset. We have collected our benchmark dataset from a wellreputed website called Kaggle. We have collected two data set one for recommending the crop and one for predicting thecrop.

Attributes or variables in dataset are,

a) *State name*
b) *District name*
c) *Season*
d) *Crop*
e) *Area*

f) *Production*
g) *Nitrogen level*
h) *Phosphorous level*
i) *Potassium level*
j) *Rainfall*
k) *Humidity*
l) *Temperature*

4.1 Data Preprocessing

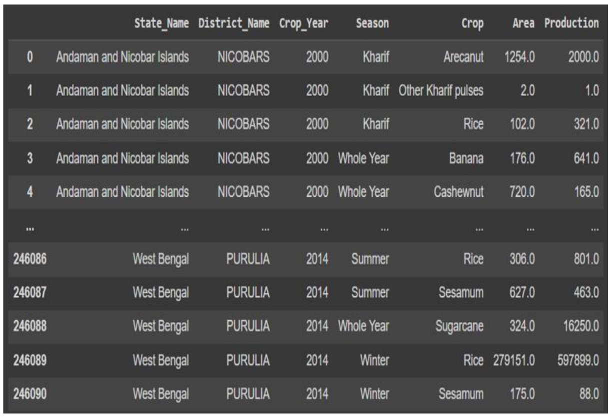
Data preprocessing is a process of preparing the raw dataand making it suitable for a machine learning model. It is thefirst and crucial step while creating a machine learning model.
Data preprocessing involves data analysis and data visualization.

Data Analysis
In Data analysis the main focus falls on removing unnecessary fields and null values from the dataset in order to infer meaningful insights from the results. thereby improving the efficiency of the model.

This helps make sure that data is evenly distributed, and the ordering does not affect the learning process. Cleaning thedata to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.

Basically, there are two types of data analysis techniques.There are two primary methods for data analysis These data analysis techniques can be used independently or incombination with the other. They are:

i) Qualitative analysis
ii) Quantitative analysis



	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0
...	...	...	...	...	...	...	...
246086	West Bengal	PURULIA	2014	Summer	Rice	306.0	801.0
246087	West Bengal	PURULIA	2014	Summer	Sesamum	627.0	463.0
246088	West Bengal	PURULIA	2014	Whole Year	Sugarcane	324.0	16250.0
246089	West Bengal	PURULIA	2014	Winter	Rice	279151.0	597899.0
246090	West Bengal	PURULIA	2014	Winter	Sesamum	175.0	88.0

Table 1. Attributes for data analysis

The above table shows all the fields that are present in thedataset. It contains attributes such as state name, district name, crop year, season, area, production. The total number of rows present in the dataset are 246090 rows and total number of columns present in the dataset are 7 columns.

Data Visualization

Data visualization is the process of visualizing the attributes in a given dataset. Data visualization techniques like scatter plots, and heat maps help us to identify which attributes have the best effect on the dependent variable.

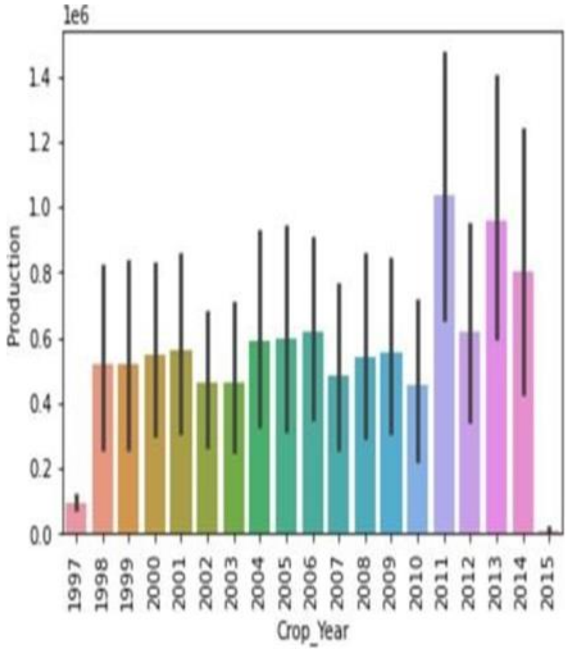


Fig 2. Data visualization for crop production in year

The above figure represents the Data visualization for crop production in year which is used for identifying dependent variable in the dataset.

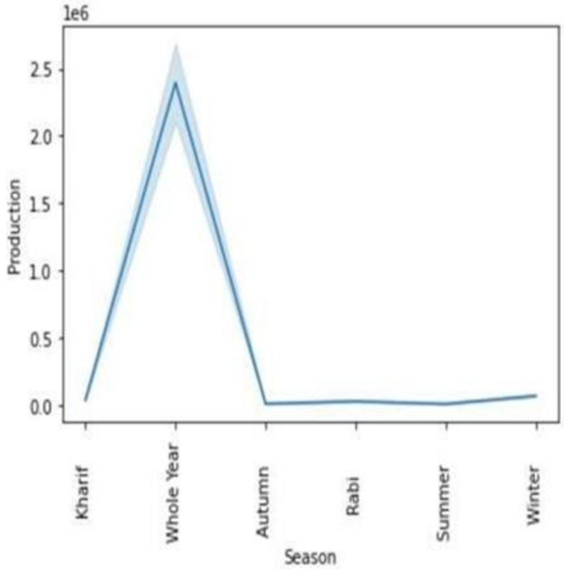


Fig 3. Data visualization for crop production in season

4.2 Performance Evaluation

There are various metrics which we can use to evaluate the performance of ML algorithms, classification as well as regression algorithms. We must carefully choose the metricsfor evaluating. In this project we used two classification algorithms such as K-Nearest Neighbor and Random Forest.

K-Nearest Neighbor

The K-Nearest Neighbor algorithm is a classificationalgorithm takes a bunch of labeled points and uses them to learn how to label other points. K-NN is a non-parametric algorithm, which means it does not make any assumption onunderlying data. It is also called a lazy learner algorithm K- Nearest Neighbor classifier helps us to build a model when we are given a dataset with predefined labels. K-Nearest Neighbor is a classification algorithm that takes a bunch of labeled Points and uses them to learn how to label other points. It classifies cases based on their similarity to other cases. In KNN, data points that are near each other are said tobe neighbors. KNN paradigm: Similar cases with the same class labels are near each other. Thus, the distance between two cases is the measure of dissimilarity.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

In our project, the KNN machine learning model is used for recommending crop for certain area based on the input given by the user. User gives information like state name, district name, potassium level, phosphorous level, nitrogen level, season and climatic parameters.

Then the extracted data is fed to the machine learningmodel i.e. KNN. The KNN model perform certain analysis and gives output.

Here we achieved the testing accuracy of 98 percent.

Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and is based on themajority votes of predictions, and Random Forest is a Supervised Machine Learning Algorithm that is used widelyin Classification and Regression problems. It builds decisiontrees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random ForestAlgorithm is that it can handle the data set containing variables as in the case of regression and categorical variablesas in the case of classification. It performs better results for classification problems. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the firstphase.

In our project, the random forest model is used to predict the yield of the recommended crop. The output from the KNNis fed to the random forest model for predicting the yield of the crop.

Here we achieved the testing accuracy of 96 percent.

IV. CONCLUSION

In this paper, we propose an automatic croprecommendation system that recommends crops and predicts the yield of the crop machine learning technique. In the existing the model is trained to predict only for a particular state in India and also the model does not take into consideration various factors like nutrient content of thesoil, humidity, etc. The model only aims is recommending crops and does not give the crop yield under that climatic condition.Whereas the solution we have proposed aims in recommending crop type and yield of the crop for various states of India based on a variety of factors like nitrogen levelof soil, phosphorous level of the soil, potassium level of the soil, season, and district. To recommend the crop and yield ofthe crop a K-Nearest Neighbor algorithm is used, and the output from the KNN is fed to the Random Forest which will provide us the desired outcome. The accuracy percentage of our crop recommendation system is 98% and the accuracy percentage of our yield prediction system is 96%.

REFERENCES

[1] Manjula E, Djodiltachoumy S, “A model for prediction of crop yield” International Journal of Computational Intelligence and Informatics, 2017 Mar;6(4):2349-6363.

[2] Sagar BM, Cauvery NK., “Agriculture Data Analytics inCrop Yield Estimation: A Critical Review”, Indonesian Journal of Electrical Engineering and Computer Science,2018 Dec;12(3):1087-93.

[3] Veenadhari S, Misra B, Singh CD, “Machine learning approach for forecasting crop yield based on climatic parameters”, In 2014 International Conference on Computer Communication and Informatics, 2014 Jan 3(pp. 1-5). IEEE.

[4] Ghadge R, Kulkarni J, More P,Priya RL, “Prediction of crop yield using machine learning”, Int. Res. J. Eng. Technol.(IRJET), 2018 Feb;5.

[5] Bang, Shivam, Rajat Bishnoi, Ankit Singh Chauhan,Akshay Kumar Dixit, and Indu Chawla. "Fuzzy logic based crop yield prediction using temperature and rainfall parameters predicted through ARMA, SARIMA, and ARMAX models." In 2019 Twelfth International Conferenceon Contemporary Computing (IC3), pp. 1-6.IEEE, 2019.

[6] Nigam, Aruvansh, Saksham Garg, Archit Agrawal, and Parul Agrawal. "Crop yield prediction using machinelearning algorithms." In 2019 Fifth International Conferenceon Image Information Processing (ICIIP), pp. 125-130. IEEE,2019.

[7] S. Pavani, Augusta Sophy Boulet P., “Heuristic Prediction of Crop Yield Using Machine Learning Technique”, International Journal of Engineering and Advanced Technology (IJEAT), December 2019, pp(135- 138)

[8] Nishant, Potnuru Sai, Pinapa Sai Venkat, Bollu Lakshmi Avinash, and B. Jabber. "Crop Yield Prediction based on Indian Agriculture using Machine Learning." In 2020 International Conference for Emerging Technology (INCET), pp. 1-4. IEEE, 2020.

[9] Kumar, Y. Jeevan Nagendra, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi. "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector." In 2020 5th International Conference onCommunication and Electronics Systems (ICCES), pp. 736- 741. IEEE, 2020.

[10] Johnson LK, Bloom JD, Dunning RD, Gunter CC, Boyette MD, Creamer NG, “Farmer harvest decisions and vegetable loss in primary production. Agricultural Systems”,2019 Nov 1;176:102672.