Crop Recommendation System using KNN and Random Forest considering Indian Data set

Tapas Kumar Mishra*, Sambit Kumar Mishra, Kanaparthi Jeevan Sai, Bachu Sai Alekhya, Athukuri Rama Nishith
Department of Computer Science and Engineering, *SRM University-AP,* Amaravati, Andhra Pradesh India.
{*kmtapas, skmishra.nitrkl}@gmail.com, {kanaparthi_jeevan, bachu_sai, athukuri_rama}@srmap.edu.in.

2021 19th OITS International Conference on Information Technology (OCIT) | 978-1-6654-1664-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/OCIT53463.2021.00068

***Abstract*—The agriculture plays crucial role in the growth of the country's economy. In comparison to other countries, India has the highest production rate in agriculture. Agriculture when combined with technology can bring the finest results. Crop prediction is a highly complex trait determined by multiple factors such as Contents of Nitrogen, Phosphorous, Potassium, Rainfall, Temperature, Humidity, Ph level. Predicting the crop in advance would help the policymakers and farmers for taking appropriate measures for farming, marketing and storage. Thus, in this paper we propose crop selection using machine learning techniques such as K- Nearest Neighbour (KNN) and Random Forest. Both of the models are simulated comprehensively on Indian Data set and an analytical report has been presented. This model will help the farmers to know the type of the crop before cultivating onto the agricultural field and thus help them to make appropriate decisions.**

***Index Terms*—Crop Selection, Machine Learning, Indian Agriculture, Prediction**

I. Introduction

Agriculture is the backbone of Indian economy. In India, most of the crops depend on the weather conditions. However, the soil quality also plays a major role for the productivity of a particular crop. For example, rice cultivation mainly depends on the rainfall. Now-a-days all the seasonal moments are not the same as the previous. We could not even predict whether there would be any floods or any water scarcity in the future. In addition to this the farmers are not strong enough at technology that they can predict the crop production for any particular crop if it is chosen to be farmed. But it is inevitable that soil health status can be used to recommend a crop type to be farmed for the next season. So in order to maximize the crop production, prediction of various aspects of crop are required based on the weather conditions in the locality. Yield prediction is an important agricultural problem. Usually, the farmers used to predict their yield from the previous year's yield. As it is discussed earlier, we could not predict the yield based on last year's outcome due to many factors like crop stress, soil impurity, floods, pesticides, pests and diseases. Here we are going to use some existing mathematical models. As farmers are growing the hybrid products that the soil generally is not supportable but they are using pesticides and growing those. So the quality of the soil decreases. And hence we could not predict for those crops. Due to these abundant inventions people are concentrated on cultivating hybrid crops where they lead to an unhealthy life. Now-a-days, modern people can take the help of technology in various dimensions to grow crops. Thus, this paper aims at predicting the suitable crop type by getting the contents of Nitrogen, Phosphorous, Potassium of the land as an input with the rainfall in that area with soil's humidity as well as surrounding temperature as temperature also plays an important role in crop growth and predict the best crop to cultivate which can be selected by the farmers to grow that is the need of the today's generation. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. Which also affects the proportion of N-P-K in the soil in turns affects the rain. Thus, we mainly mainly focus on these parameters and use data mining techniques to solve this question. Data mining is also useful for predicting crop yield production [1], [2].

The rest of the paper is organised as follows. In section II some related literature are highlighted with their problem solving approaches. In section III, we have presented our simulation models and details about dataset including preprocessing stages. In section IV we have discussed the simulation results and it anslysis in two different subsections. Finally, we have concluded with some future directions in section V.

II. State of the art

This problem is identified before a couple of years. After that many attempts has been made by the researchers throughout the globe. However, there exist many limitations among the farmers and the technological support by the regions and states. Some of the directions for this issue are described here as follows.

In [3], the authors have predicted crop yield using boosting techniques, random forest, support vector machine, k-nearest neighbors and artificial neural network. This paper[3] proposed a method named Crop Selection Method (CSM) to achieve net yield rate of crops over season. CSM method,[3] may improve net yield rate of crops using limited land resource and also increases re-usability of the land. CSM algorithm works on prediction of crop yield rate based on favorable condition in advance and gives a sequence of crops with highest net yield rate.

In [4], the authors have predicted crop yield using decision tree classifier. They used rainfall, perception, production, temperature data to construct a random forest which is a collection

978-1-6654-1664-1/21/$31.00 ©2021 IEEE
DOI 10.1109/OCIT53463.2021.00068

of decision trees using $\frac{2}{3}$ rd of the data and they tested using $\frac{1}{3}$ rd of the data. Usually, decision tree classifiers uses greedy approach, where an attribute chooses at first step can't be used anymore which can give better classification if used in later steps. Also it overfits the training data which can give poor results for unseen data to overcome which they have combined results from different models to get a better result.

In [5], they have taken a dataset containing soil type, soil Ph, Humidity, Temperature, Rainfall, Wind, Production, Cost of Production and annual yield of that region for past 10-12 years and a decision tree classifier model has been implemented on the data for crop yield and K-Nearest neighbours has been applied for prediction of rainfall with 76.8% accuracy for crop yield prediction and 89.4% accuracy for rainfall prediction.

In [6], the authors used a deep neural network model to predict four crop yields namely: Aus-rice, Aman rice, Boro rice, Jute, Wheat and Potato using rainfall data, land types, chemical fertilizers, soil information. The DNN model is compared with RF, SVM and LR. DNN outperforms than other models with highest accuracy rate of 98% (Aus rice), 95% (Aman rice), 96% (Boro rice), 97% (Potato), 96% (Wheat) and 94% (Jute).

The authors in [7] investigated the crop suggestion model based on soil classification using machine learning techniques. The study proposed an SVM based model to suggest crops which are specific to soil conditions. The proposed SVM model outperforms KNN and bagged trees with 95% of accuracy.

In [8], the authors investigated the rice yield prediction performance of KNN, decision tree(DT) and Naive Based(NB) using 11 parameters of micronutrients and macronutrients. The prediction accuracy for Naive Based is 98%, DT is 94% and KNN is 97% is achieved. The study concluded that NB achieved better prediction rate and was suitable for rice yield prediction using soil parameters.

In [9] proposed a crop recommender model for farmers using machine learning models. The prediction model is prepared using ANN and the model performance is compared against DT, KNN, RF. ANN achieved a highest of 91% than other models. The crop suitability is predicted using rainfall, soil type, soil conditions, temperature and geographical location.

III. Our Contribution

In this paper, we have focused to predict the best crop to be grown in lands of farmers for the maximum yield. The crop is predicted by using various machine learning algorithms such as KNN and Random Forest. Further we have made a quantitative analysis of the accuracy. To predict the best crop we have used 7 parameters i.e. Nitrogen (N) , Phosphorous (P), Potassium (K), Temperature , humidity, PH value, rainfall from the Data-set. The above seven parameters are considered to recommend the crop name. Thus, the deciding parameters are known as X- input factor and recommended factor is taken as Y- output factor which is shown in Table I.

The dataset has 1547 records for training and 618 records for testing where the crop name is a categorical data and others

TABLE I
DATASET MODEL

N	P	K	Temperature	Humidity	ph	Rainfall	Crop
X							Y
—	—	—	—	—	—	—	—

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
...	...	...	...	...	...	...	...	...
1542	99	40	32	24.184712	69.948073	7.045543	163.270873	coffee
1543	89	28	33	26.444141	53.838762	6.993236	175.372331	coffee
1544	112	39	29	26.124922	63.374792	6.726529	147.803530	coffee
1545	111	28	26	27.773633	64.478587	6.937353	192.712124	coffee
1546	114	20	26	25.556567	62.670878	7.279057	193.586623	coffee

1547 rows × 8 columns

Fig. 1. Sample Dataset

are numerical data. The snapshot of the dataset is shown in Fig. 1. The training dataset and testing dataset are shown in Fig. 2 and Fig. 3.

The dataset considered here is a complete dataset that contains almost zero blank fields. Thus it is not required any onehot encoding or pre-processing for filling the bland fields. As the dataset contains only one categorical field, we have chosen to use random forest that may perform better. In the dataset 22 different crop types are taken. A snapshot of the dataset showing unique croptypes is shown in Fig. 4.

IV. Simulation and Analysis

Dataset contains 22 different variety of crops dataset manually separated into70% training and 30% testing datasets The dataset contains numerical and categorical attributes. Standard scalar preprocessing techniques has been applied on this dataset for numerical attributes for normalizing values and maintain equality. Label Encoder is used for categorical attributes to convert labels into a numeric form to convert

```
training_dataset.head()
```

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

```
[ ] len(training_dataset)
1547
```

Fig. 2. Details of Training dataset Dataset

```
testing_dataset.head()
```

	N	P	K	temperature	humidity	ph	rainfall	label
0	61	52	41	24.976695	83.891805	6.880431	204.800185	rice
1	67	45	38	22.727910	82.170688	7.300411	260.887506	rice
2	79	42	37	24.873007	82.840226	6.587919	295.609449	rice
3	78	43	42	21.323763	83.003205	7.283737	192.319754	rice
4	75	54	36	26.294655	84.569193	7.023936	257.491491	rice

```
[ ] len(testing_dataset)

618
```

Fig. 3. Structure of testing dataset.

```
print("Unique Categories ")
print(crops['label'].unique())
print("Total crops are: ",len(crops['label'].unique()))

Unique Categories
['rice' 'maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'mothbeans'
 'mungbean' 'blackgram' 'lentil' 'pomegranate' 'banana' 'mango' 'grapes'
 'watermelon' 'muskmelon' 'apple' 'orange' 'papaya' 'coconut' 'cotton'
 'jute' 'coffee']
Total crops are:  22
```

Fig. 4. Different crop types considered for recommendation.

them into the machine-readable form. After this much pre-processing, the dataset considered here is a complete dataset that contains almost zero blank fields.

We have simulated the models using KNN and Random Forest extensively on the given dataset 1. The simulation setup and models on this above dataset is as follows.

A. Simulation

KNN and Random Forest for multi classification: K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. Ex:Take few rows of rice and bajra. Further, calculate K-Nearest Distance i.e K is odd value with new row and plot new row to nearest category. 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 first phase. The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2 for each new rows.

1) K-Nearest Neighbour: KNN stands for K-Nearest Neighbour. In this supervised learning algorithm which is used for classification, we classify based on how it's neighbours are classified. It stores all it's previous cases and classifies new ones based on how similar it is to the previous cases. Here, K signifies the amount of neighbours we take for comparing the distance. We find the distance between new and previous cases

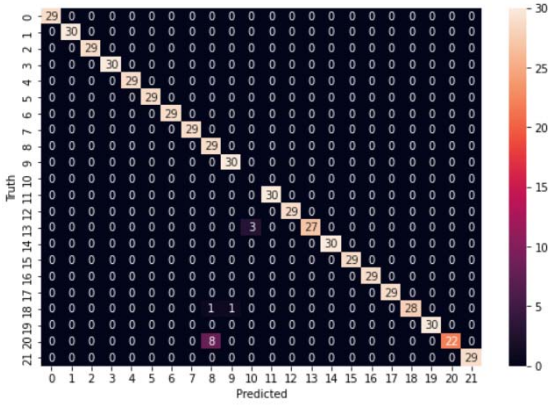

Fig. 5. Confusion Matrix of KNN.

based on the equation 1. The distance D using minkowski equation is:

$$D(X_1, X_2) = \left[\sum |X_1 - X_2|^{1/p}\right]^p \quad (1)$$

Here, value of p is taken as 2 for minkowski distance. Further, X_1 represents coordinates of the neighbour nodes and X_2 represents coordinates of the new node. By calculating all the k-nearest point distances we can get a decision based on those coordinates. Consider the Fig. 1 above where you can see seven X variable columns and one Y variable column. In those X columns we have no categorical variables and everything are numerical variables which are further scaled down using Standard Scalar and Y is a categorical variable. So before applying formula we don't need to convert the categorical variables into numerical before training but we need to encode the dependent variable Y as it is a categorical and it is easy to display the mathematical model and to train the model too. So in this model we have used label encoder to convert the categorical into the numerical and which is used to train to the model.

The categorical column is transformed into the numerical column by applying label encoding which is shown in the Fig. 7. Here we can see 8 columns, out of which 7 are numerical and those are also known to be input variables and 1 categorical variable that would be output of the model that we create and as it is a categorical variable. we need to encode it to the numerical which is useful while we train the model, as supervised models couldn't train on the categorical values directly. We need to convert them to the numerical and train them.

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix presented in figure 5.

2) Simulation of RANDOM FOREST CLASSIFIER: Random Forest classifier is simply a bagging technique. It is a classification algorithm where we come to a decision based on

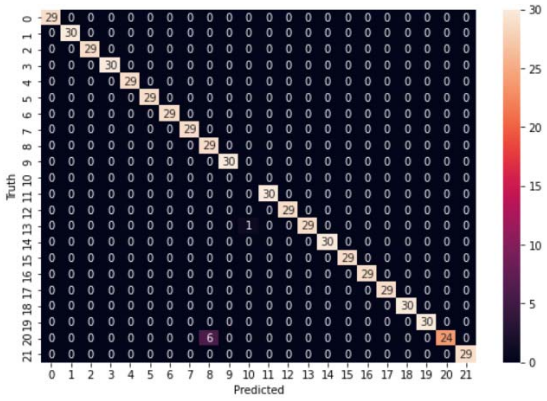

Fig. 6. Confusion Matrix of Random Forest.

```
y_sample_utility_testing_dataset
```

```
array([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
       20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 11, 11, 11, 11,
       11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
       11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,  3,  3,  3,  3,  3,  3,  3,
        3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,  3,
        3,  3,  3,  3,  3,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,
        9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,
        9, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
       18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 13, 13, 13,
       13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
       13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 14, 14, 14, 14, 14, 14, 14,
       14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14,
       14, 14, 14, 14, 14, 14,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,
        2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,
        2, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
       19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,  1,  1,  1,
        1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
```

Fig. 7. Dataset after applying label encoder.

several trees. As in our model we are classifying the crop based on the user input we can use the random forest classifier. This algorithm build each tree to design the uncorrelated forests which is going to predict accurate answer by using multiple decision trees. A decision tree generates only one scenario of a tree but where as random forest is a group of trees which checks every possible uncorrelated trees and generate the accurate average answer. This random forest algorithm produces good predictions either by using the classification as well as the regression tasks. It generates a good accuracy over the decision tree. And using this random forest classifier we can prevent over-fit. As we are going to have multiple trees with separate conditions and by having maximum vote technique we get the result. So as this is maximum vote result there is high probability to get a correct answer.The testing performance won't be affected as the number of trees increases. Random forest also produces lower bias. In this random Forest we do have multiple decision trees which are generated from the original dataset by multiple (row sampling + feature sampling) sampling techniques. So we are going to find multiple correlations with in the dataset and get an accurate answer. Here row sampling and feature sampling happen with replacement of the records. It mean the rows which appear in Decision tree 1 can also be appeared in Decision tree 2 but not every row matches. so it is called as sampling technique. In general decision trees have low bias

```
accuracies
```

```
array([0.96774194, 0.97741935, 0.95145631, 0.97734628, 0.97411003])
```

```
accuracies.mean()
```

```
0.9696147823363608
```

```
accuracies.std()
```

```
0.009737887893012488
```

Fig. 8. Results of cross validation.

and high variance it mean we are having low training error but high testing error. But when we combine all the decision trees and use the majority vote technique, then the high variance is converted to the low variance. Hyper parameter denotes number of Decision trees that we are going to use in our Random Forest machine learning model. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix presented in the figure 6.

B. Analysis of Models

This paper is related to classification of crops. Random Forest and KNN are mostly used for classification problems. Dataset contains Temperature,Humidity,Ph level,Rainfall,Nitrogen,Phosphorous and Pottassium these factors mostly follows trends and changes similarly So,KNN and Random Forest are better algorithms for this dataset.KNN-contains some instances if any new instance matches with these instances then it is a neighbhour. Random Forest-multiple instances average and majority voting gives decision of output but decision tree makes only single decision. Logistic Regression is used for predicting the categorical dependent variable using a given set of independent variables. The decision tree model gives high importance to a particular set of features. But the random forest chooses features randomly during the training process. Therefore, it does not depend highly on any specific set of features. This is a special characteristic of random forest over bagging trees.

Here, we have trained our dataset with KNN and Random Forest Classifier. It is observed that Random Forest Classifier shows better accuracy when it is stimulated with our dataset. KNN couldn't work as much as Random Forest because KNN and Random Forest both depends on the majority vote technique. But mainly KNN depends on the distance between already existing data and the new data. And it is mostly depends on classes of the neighbours. So it mean if value of K changes then the output also varies. hence accuracy is unstable. This model is Good when we are having abundant amount of data to track. When coming to the Random Forest the main benefit is sampling techniques. It finds all the correlations between the dataset and model multiple decision trees based on some sampled data of the dataset. Sampling includes both

TABLE II
ACCURACY ANALYSIS

Model	Accuracy in %
K-Neighbors Classifier	96.96147823363608
Random Forest Classifier	98.05825242718447

the row as well as feature sampling. So there is a high chance of having different correlations among the dataset. Hence we get a different outputs for different decision trees based on the co-reations of respective trees. The experimented model uses cross validation techniquex. It is a resampling procedure used to evaluate machine learning models and access how the model will perform for an independent test dataset and used KNN classifier i.e N-neighbors as 5 and applied classifier to each model finally, accuracy is computed which is shown in 8. Based on those results by using majority vote technique we get a maximal accurate correlated result. There is very high accuracy because we are developing multiple decision trees and in turn getting majority answer from that. So Random Forest is much preferable but it is highly computation compared to KNN. Random forest have low bias and low variance.

V. CONCLUSION

In this paper, we have shown that we can suggest the crop that should be cultivated in a particular region by the help of its soil quality, N-P-K values, humidity and expected rainfall. The accuracy of the prediction may improve if we will have a proper tested soil with additional features. We have used K-Neighbors Classifier and Random Forest Classifier for the crop recommendation. This system may help the farmers to decide the crop , to be chosen for the upcoming season that will not degrade its product outcome followed by loss. Using this system, the profit of the agricultural sector can be improved through maximizing crop harvest , which will lead to gain of interest among the youngsters for technology based farming. In future a hybrid model can be designed with a Data-set having large number of attributes that will make a strong and robust model.

REFERENCES

[1] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020.

[2] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and electronics in agriculture*, vol. 151, pp. 61–69, 2018.

[3] R. Kumar, M. Singh, P. Kumar, and J. Singh, "Crop selection method to maximize crop yield rate using machine learning technique," in *2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*, pp. 138–145, IEEE, 2015.

[4] P. Priya, U. Muthaiah, and M. Balamurugan, "Predicting yield of the crop using machine learning algorithm," *International Journal of Engineering Sciences & Research Technology*, vol. 7, no. 1, pp. 1–7, 2018.

[5] A. Patil, S. Kokate, P. Patil, V. Panpatil, and R. Sapkal, "Crop prediction using machine learning algorithms," *International Journal of Advancements in Engineering & Technology*, vol. 1, no. 1, pp. 1–8, 2020.

[6] T. Islam, T. A. Chisty, and A. Chakrabarty, "A deep neural network approach for crop selection and yield prediction in bangladesh," in *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 1–6, IEEE, 2018.

[7] S. A. Z. Rahman, K. C. Mitra, and S. M. Islam, "Soil classification using machine learning methods and crop suggestion based on soil series," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pp. 1–4, IEEE, 2018.

[8] V. Singh, A. Sarwar, and V. Sharma, "Analysis of soil and prediction of crop yield (rice) using machine learning approach.," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, 2017.

[9] Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "Agroconsultant: Intelligent crop recommendation system using machine learning algorithms," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–6, IEEE, 2018.