An Ensemble Approach for Crop Recommendation and Yield Estimation Using Machine Learning

[1]Ch.Lakshmi Sahithi, [2]G.Eshwar, [3]Dr. R. Jegedeesan, [4]Dr. M Sujatha, [5]N.Venkateswaran

IV Year CSE Students[1,2,3,4] Associate Professor[3,4,5]

Jyothishmathi Institute of Technology And Science, Karimnagar, 505 481, Telangana.

Abstract

In many developing countries like India agriculture is the most widely used habitations and provide a major part of economy of country. Agriculture is the main occupation of many rural populations in India. Many of the farmers now-a-days are facing hardships due to lack of technical knowledge to use precision farming and unpredictable weather patterns such as temperature, rainfall and many other parameters. Therefore choosing the right crop to grow and estimate the yield is a crucial part of improving real-life farming situations. This paper considers the attributes like soil type, area, state, pH value, potassium, phosphorous, nitrogen, crop for crop recommendation and yield prediction. The user can predict the most suitable crop and its estimated yield for a given area. Machine learning algorithms allow choosing the most profitable crop list or estimating the crop yield for a user selected crop. This model uses the machine learning algorithms like K-Nearest Neighbours(KNN), Random Forest Regressor, Random forest Classifier, Multivariate linear regression(MLR),Support Vector Machines(SVM) are used. Additionally the system also suggests the best time to use the fertilizers to boost up the yield. There is a web interface for the farmers to flexibly use as end users. Therefore, this proposed model assists the farmers in choosing the suitable crop that can be grown in a particular region during a specific season or specific period of time and estimate its yield and predict the most profitable crop. The algorithm that has highest score is used. Among all, the Random Forest showed the best results with 95% accuracy. Hence compared to the existing model this modelhelps the farmers in maintaining their time by assisting them in the decision making process.

Keywords : Precision Farming,K-Nearest Neighbours(KNN), Random Forest Regressor, Random forest Classifier, Multivariate linear regression(MLR),Support Vector Machines(SVM),Gini index.

I.Introduction

Agriculture has an extensive history in India. It holds around 56% of the total GDP in India[4]. Machine learning (ML) approaches are used in many fields, ranging from supermarkets to evaluate the behaviour of customers to the prediction of customers' buying patterns. Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, [8]soil, use of fertilizer, and seed variety. Nowadays, crop yield prediction models can estimate the actual yield reasonably, but a better performance in yield prediction is still desirable. ML studies consist of different challenges beyond data mining techniques[7] when aiming to build a high-performance predictive model[4]. It is crucial to select the right algorithms to solve the problem at hand, and in addition, the algorithms and the underlying platforms need to be capable of handling the volume of data. Machine Learning provides a practical approach that can provide better yield estimation based on several patterns and correlations[10] and discover knowledge from datasets[9]. The models need to be trained using datasets, where the outcomes are represented based on past experience[2,5]. The predictive model built using several features and as such, parameters of the models are determined using historical data during the training phase.

In addition to yield prediction we have also added a feature helping them in recommending the right crop to cultivate. The crop recommendation is also one of the majorly problems in precision agriculture[5]. Not all precision agriculture[16] systems offer best results. Precision agriculture[16] is a technology of site-specific farming. Selecting a crop for farming is one of the essential decisions[8] farmers have to make as their whole revenue and yield depends on this decision. But in agriculture it is important that the recommendation made are accurate and precise because in case of errors it may lead to huge material and capital loss. Recommendation of crop is one major domain in precision agriculture[7,16]. We are recommending the right crop[6] using parameters like state, rainfall, soil type, pH value.[2,16]

We have used the proposed method to minimize the errors[3] and give better predictions which in turn would help the farmers.

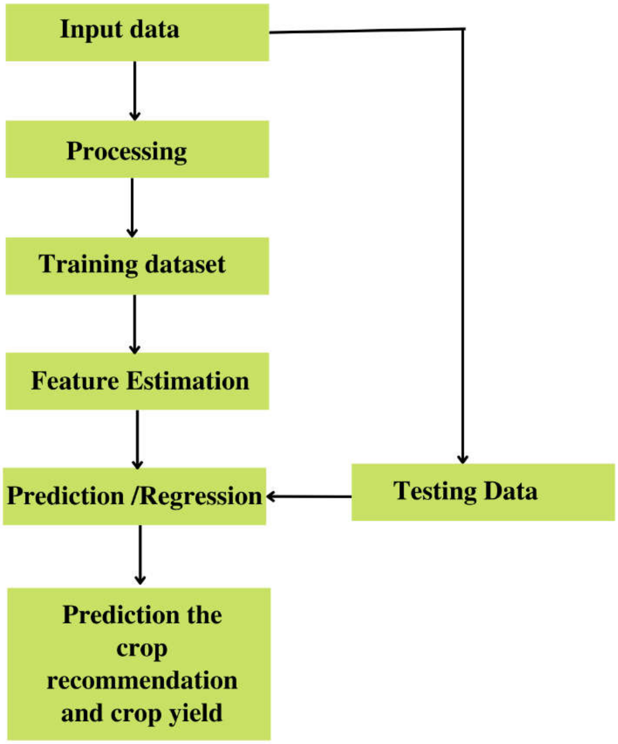

Fig 1 : Work Flow of Proposed System

Working of Proposed System

It works in four steps:

1.Select random samples from a given dataset.

2.Construct a decision tree for each sample and get a prediction result from each decision tree.

3.Perform a vote for each predicted result.

4.Select the prediction result with the most votes as the final prediction.

II. Related work

Authors: Thomas vanKlompenburg,Ayalew_Kassahun,Cagatay_Catal-2020[16]

In this paper a systematic literature review is done to get all the necessary features and incorporate the algorithms and features that have been used in crop yield prediction. This paper retrieved 567 relevant studies from six electronic databases of which 50 studies for further analysis using inclusion and exclusion criteria. These selected studies were investigated, analysed, and features were used for further research. The most used features are soil type, temperature, rainfall and the most used is Artificial Neural Networks in these models. The additional analysis is mostly used Convolutional Neural Networks(CNN) and other widely used deep learning are long-short Term memory(LSTM) and Deep Neural Networks.

Authors:S. Veenadhari,Bharat Misra,CD Singh-2014[1]

In this paper we learned thatpredicting the crop yield well ahead of its harvest would help the policy makers and farmers for taking appropriate measures for marketing and storage. Such predictions will help all the industries for planning the required materials. Several models of predicting and modelling crop yields have been developed in the early days with variable rate of accuracy, as these don't take into account characteristicsoftheweather, and aremost features. In the present day software tool named `Crop Advisor' has been developed as an flexible web page for estimating the influence of climatic parameters on the crop yields.C4.5 algorithm is used to find out the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh. This web page indicates the relative influence of different climatic parameters on the crop yield, other agro-input parameters responsible for crop yield are not considered in this tool, since, application of these input parameters varies with individual fields in space and time.

Authors:Y. Jeevan Nagendra Kumar,V. Spandana,V.S. Vaishnavi,K. Neha,V.G.R.R. Devi-2020[12]

In this paper,crop yield and prediction include forecasting the yield of the crop from the past data which contains the attributes like ph, humidity,rainfall,crop name. This attributes are required for predicting the crop for that wheather conditions. This is done through random forest machine learning algorithm. It will achieve the best accurate value for crop prediction. This algorithm uses least number of model to predict the yield of crop.

Authors: Potnuru Sai Nishant,Pinapa Sai Venkat**,**Bollu Lakshmi Avinash,B.Jabber-2020[11]

This paper consists of various parameters like state, district, season, area and then user predicts the yield of the crop in which year. This paper, uses advanced techniques like kernel Ridge, Lasso and ENet and other algorithms to predict the yield and it also used the stacking regression for enhancing the algorithms to give a better prediction than previous.

III. Methodology

Data Collection

The datasets include specific soil features that are collected for various states which constitutes the attributes like State name, season, crop, area, production, soil-type[2,16]. The values of crop attribute consists of various crops like are arecanut, banana, dry chilles, coconut, paddy, sugarcane, cotton, ginger, groundnut, maize, wheat, jowar etc. There are number of cases of each crop available in the training set. Attributes of depth consideration like soil type, nitrogen, potassium are also considered [2,16]. The dataset is splitted into 75% and 25% [9]for both training and testing. Here we imported train_test_split function of sklearn. Then use it to split the dataset. Also, *test_size = 0.2*, it makes the split with 80% as train dataset and 20% as test dataset. The *random_state* parameter seeds random number generator that helps to split the dataset. We used Random Forest Classifier*,* which fits multiple decision tree to the data. Finally we train the model by passing variables to the *fit* method. Once the model is trained, we need to Test the model. For that we will pass a variable to the predict method.

Working

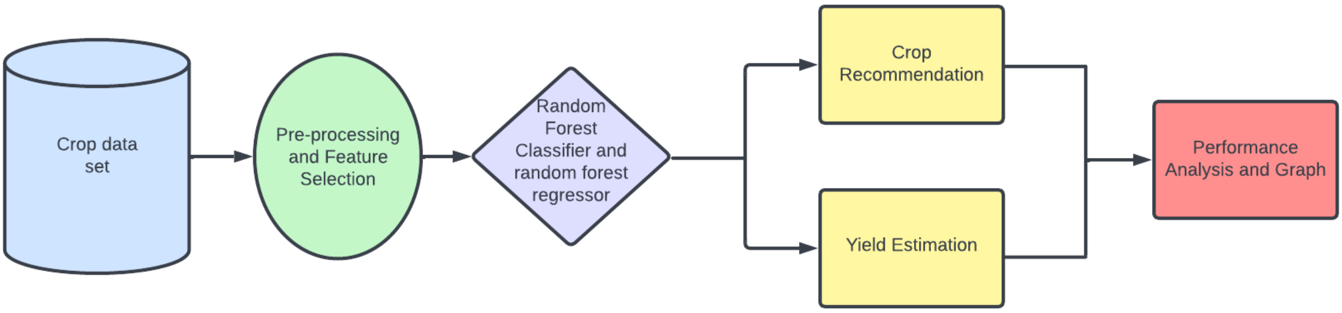
Crop data set
Pre-processing and Feature Selection
Random Forest Classifier and random forest regressor
Crop Recommendation
Yield Estimation
Performance Analysis and Graph

Fig 2 : Proposed Architecture For Crop Recommendation and Yield Estimation

Normalization : Normalization is a rescaling of the data from the original range so that all values are within the new range of 0 and 1.Normalization requires that you know or are able to accurately estimate the minimum and maximum observable values. You may be able to estimate these values from your available data.

A value is normalized as follows:

- y = (x – min) / (max – min)

A value is standardized as follows:

- y = (x – mean) / standard_deviation

Where the *mean* is calculated as:

- mean = sum(x) / count(x)

And the *standard_deviation* is calculated as:

- standard_deviation = sqrt(sum((x – mean)^2) / count(x))

Model Selection

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column.Here we imported train_test_split function of sklearn. Then use it to split the dataset.Also, *test_size = 0.2*, it makes the split with 80% as train dataset and 20% as test dataset.The *random_state* parameter seeds random number generator that helps to split the dataset.The function returns four datasets. Labelled them as *train_x, train_y, test_x, test_y.* If we see shape of this datasets we can seethe split of dataset.We usedRandomForestClassifier*,* which fits multiple decision tree to the data.

Finally we train the model by passing *train_x, train_y* to the *fit* method.Once the model is trained, we need to Test the model. For that we will pass *test_x* to the predict method.

The Random Forests Algorithm

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset[9,16]. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute[2,16]. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result[17]. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- The algorithm can be used in both classification and regression problems.
- Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

Feature Selection Method

Random forests also offers a good feature selection indicator[17]. Scikit-learn provides an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction. It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses gini importance or mean decrease in impurity (MDI) [17] to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.

Gini Index : The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favours mostly the larger partitions and are very simple to implement. In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.

The Gini Index varies between 0 and 1, where 0 represents purity of the classification and 1 denotes random distribution of elements among various classes. A Gini Index of 0.5 shows that there is equal distribution of elements across some classes.

Mathematically, The Gini Index is represented by

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Equation --- 1

a)Predict The Crop Recommendation

The crop dataset is given as input and the preprocessing of the input dataset and feature selection is done using gini index. The Random Forest Classifier is used for Crop recommendation. The input dataset is splitted into training and testing.

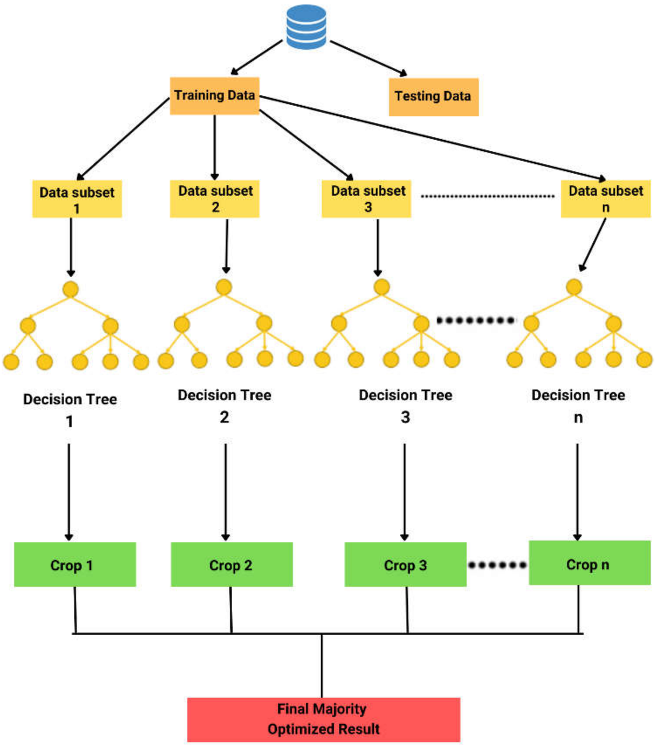

Fig 3 : Crop Recommendation using Random Forest Classifier

Experimentation Analysis

In the actual dataset, we chose only 4 features :

State_Name,Season,Soil type,area

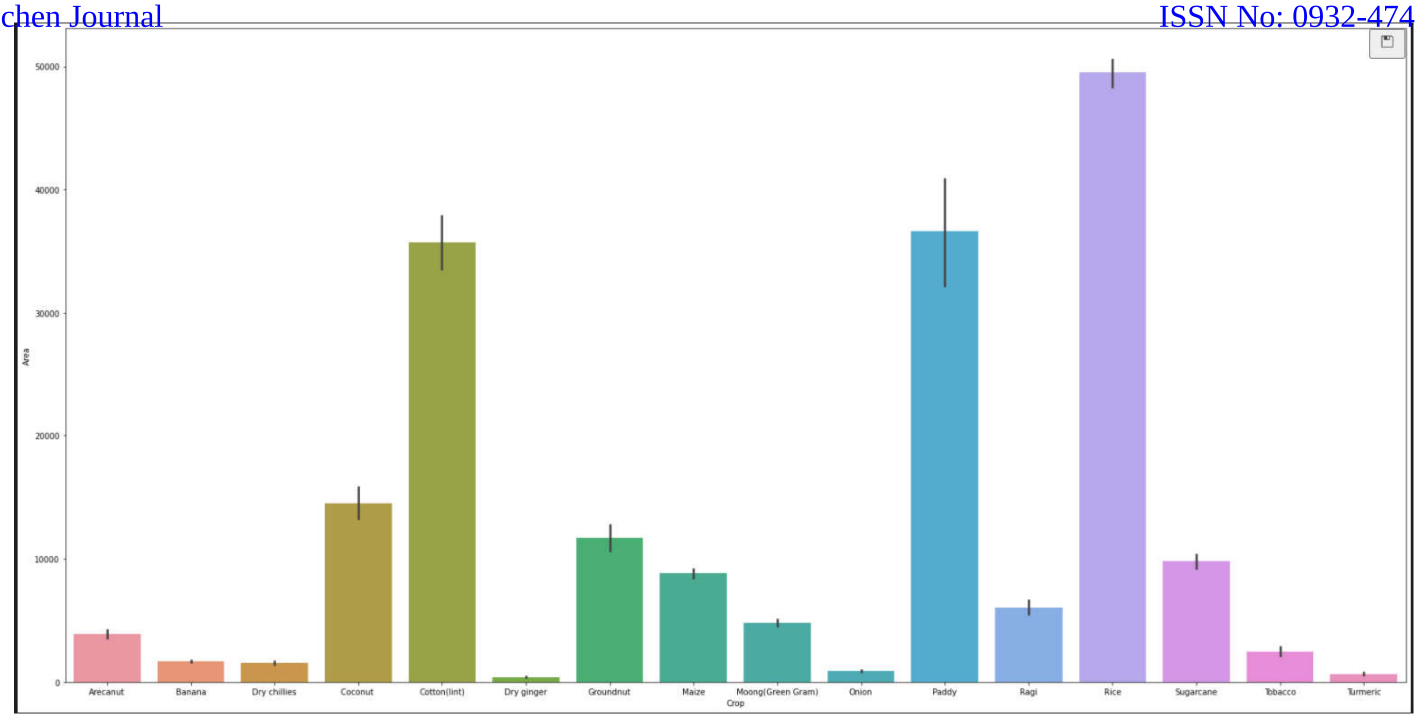

Fig 4 : Crop Recommendation

From the above mentioned(fig-4) we can conclude that rice has 1st highest area and Cotton is 2nd highest area. These two crops are mostly cultivated in upto 3500 to 5000 acres.

Accuracy on test set: We got a accuracy of 0.87% on test set.

b)Predict The Crop Yield

The crop dataset is given as input and the preprocessing of the input dataset and feature selection is done using gini index. The Random Forest Classifier is used for Crop recommendation. The input dataset is splitted into training and testing.

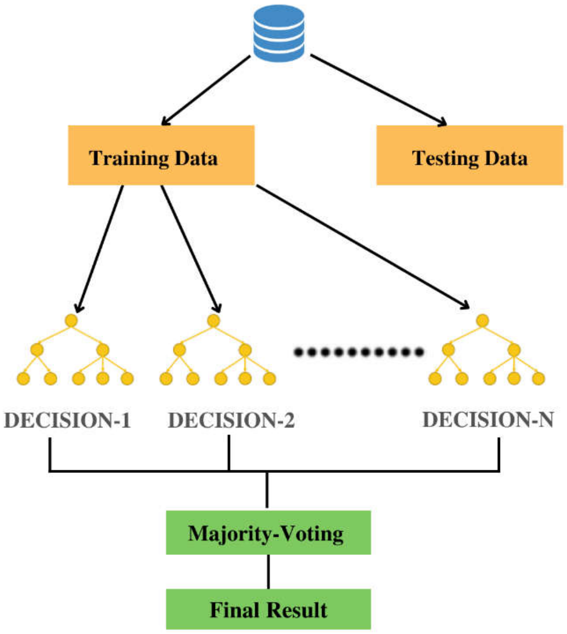

Fig 5: Crop Estimation using Random Forest Regressor

Experimentation Analysis

In the actual dataset, we chose only 4 features :

State_Name,Season,Soil type,area

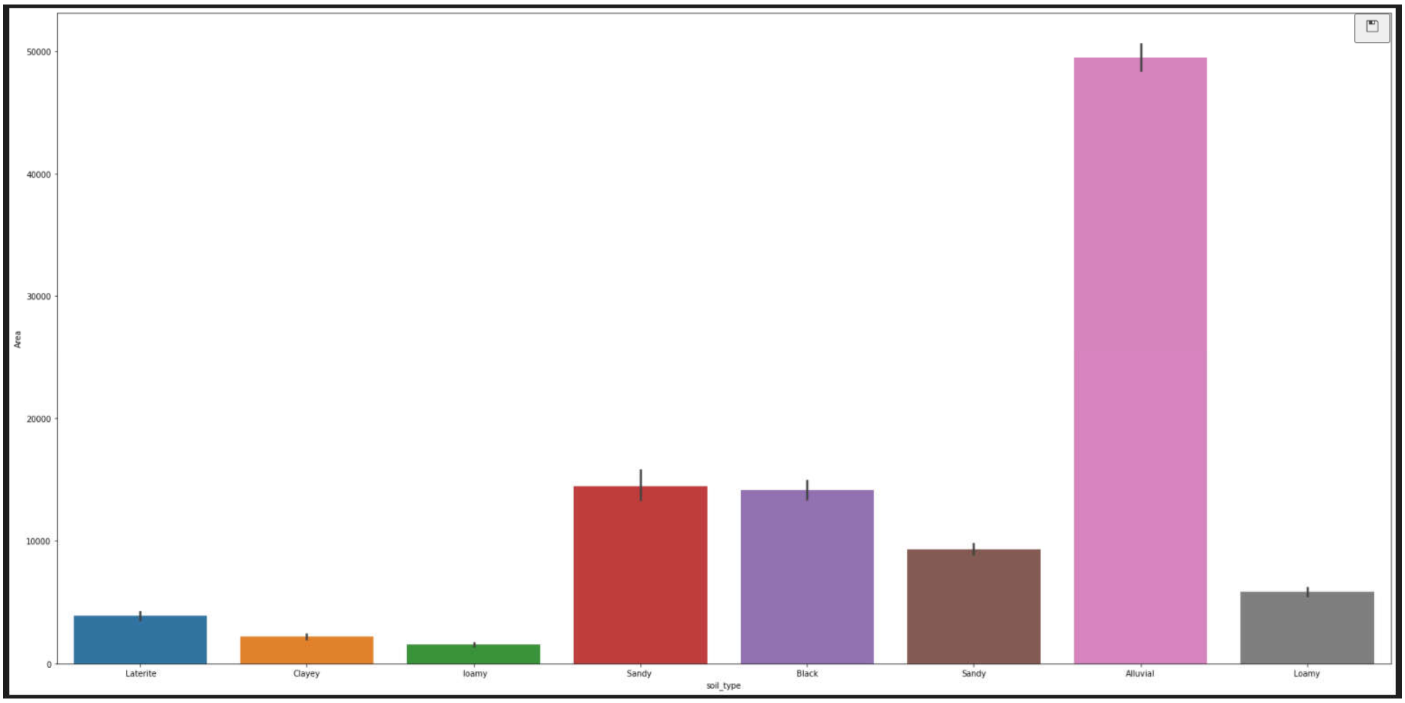

Fig 6 : Yield Prediction

From fig-6 the various amounts of areas with types of soils can be seen in which alluvial soil is present in most of the areas.

Accuracy on test set: We got a accuracy of 0.96% on test set.

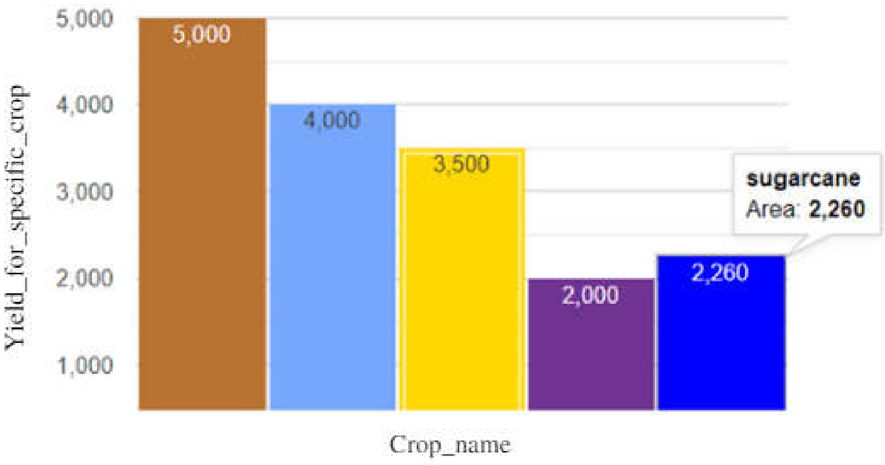

Fig 7 : Analysis of Yield Prediction

IV . Conclusion

India has a nation where agriculture plays a vital role. In the welfare of farmers, exploration of the nation.So our work will help farmers sowing theright seeds based on soil requirements to increase productivity and acquire profit out of such a technique.Thus farmers can plant right crop increasing his yield and also increasing overall productivity of the nation.The algorithm that has highest score is used. Among all, the

Random Forest showed the best results with 95% accuracy. Our future work is aimed at designing an IoT based tool that can be used to test soil by farmers by their own and later can be attached to any smart phone to connect our app and get the various parameters like nitrogen, potassium and phosphorous levels, soil type, pH values for crop recommendation and yield estimation. We are also planning to extend the dataset to foreign countries.

V.References

[1] Veenadhari, S., Bharat Misra, and C. D. Singh. “Machine learning approach for forecasting crop yield based on climatic parameters.” International Conference on Computer Communication and Informatics. IEEE, 2014.

[2]Kumar, Y. Jeevan Nagendra, et al. “Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector.” 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020.

[3] Bhosale, Shreya V., et al. “Crop yield prediction using data analytics and hybrid approach.” Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 2018.

[4] CH. Vishnu Vardhanchowdary, Dr. K. Venkataramana, “Tomato Crop Yield Prediction using ID3.” International Jounal of Engineering Reasearch and Technology (IJERT) (Volume 4 Issue 10 pp. 663–62), 2018.

[5] Girish L, Gangadhar S, Bharath T R, Balaji K S, Abhishek K T. “Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning.” National Conference on Technology for Rural Development (NCTFRD-18), 2018.

[6] Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. “Crop yield prediction using machine learning algorithms.” Fifth International Conference on Image Information Processing (ICIIP) (pp. 125–130). IEEE, 2019.

[7] Akshatha, Shailesh Shetty S, Anet P James, Athira M Saseendran, Chaitra M Poojary, “Crop Analysis and Profit Prediction using Data Mining Techniques” (Id:39), International

[8] Kamatchi, S. B., and Parvathi, R. (2019). Improvement of Crop Production Using Recommender System by Weather Forecasts. Procedia Computer Science, 165, 724–732.

[9] Nishiba Kabeer, Dr Loganathan. D and Cowsalya. T. “Prediction of Crop Yield Using Data Mining.” International Journal of Computer Science and Network (IJCSN) (2019).

[10]Palanivel, K., and Surianarayanan, C. “An approach for the prediction of crop yield using machine learning and big data techniques.” International Journal of Computer Engineering and Technology, 10(3), 110- 118.

[11].Potnuru Sai Nishant,Pinapa Sai Venkat,Bollu Lakshmi Avinash,B.Jabber-2020.

[12] Y. Jeevan Nagendra Kumar,V. Spandana,V.S. Vaishnavi,K. Neha,V.G.R.R. Devi-2020

[13] Kevin Tom Thomas, Varsha S, Merin Mary Saji, Lisha Varghese, and Er. Jinu Thomas. “Crop Prediction Using Machine Learning.” International Journal of Future Generation Communication and Networking (2020).

[14] Ms. Fathima, Ms Sowmya K, Ms Sunita Barker, Dr Sanjeev Kulkarni. “Analysis of crop yield prediction using data mining technique.” International Research Journal of Engineering and Technology (IRJET) – 2020.

[15] Pavan Patil, Virendra Panpatil, Prof. Shrikant Kokate. “Crop Prediction Using Machine Learning.” International Research Journal of Engineering and Technology (IRJET) (2022).

[16]ThomasvanKlompenburg,AyalewKassahun,CagatayCatal *InformationTechnologyGroup,WageningenUniversity&Research,Wageningen,theNetherlands,DepartmentofComputerEngineering,BahcesehirUniversity,Istanbul,Turkey.*

[17]van Klompenburg, Thomas ; Kassahun, Ayalew ; Catal, Cagatay. / Crop yield prediction using machine learning: A systematic literature review. In: Computers and Electronics in Agriculture. 2020 ; Vol. 177.

[18] Dr.A.Nirmal Kumar, Dr.R.Jegadeesan, Dr.D.Baswaraj, J.Greeda 2019 “Improved Migration Performance In Virtualized Cloud Datacenters” International Journal of Scientific & Technology Research. Volume 8, Issue 09,page no.1515-1518 September 2019. (Scopus indexed)

[19].Annadi Jahnavi, Hanumandla Bhavana, Dr. R. Jegadeesan “An Implementation of Detecting Password Pattern In Dictionary Attack” International Journal of Advanced Science and technology. ISSN 2005-42386, Page no: 84-92 July,2019(Indexed by Scopus, Elsevier).

[20].Dr.A.Nirmal Kumar, Dr.R.Jegadeesan, Dr.C.N.Ravi, J.Greeda2019 “A Secure Transaction Authentication Scheme using Blockchain based on IOT” International Journal of Scientific & Technology Research. VOLUME 8, ISSUE 10,Page no:2217-2221 OCTOBER 2019. ISSN 2277-8616 (Scopus indexed)

[21].Dr R Jegadeesan, Dr.C.N.Ravi, Dr.A.Nirmal Kumar 2020 “Automatic Rice Quality Detection Using Morphological and Edge Detection Techniques” ICCCE 2020 3rd International Conference on Communications and cyber Physical Engineering, Metadata of the chapter that will be visualized in Springer Link. Volume, issue, May.2020 Page No.233-242, Springer conference.