COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)

Rice and Wheat Yield Prediction in India Using Decision Tree and Random Forest

Dr. B M Sagar [1], Dr.N K Cauvery [2], Dr.Padmashree T [3], Dr.R. Rajkumar [4]

[1] Information Science & Engineering, RVCE, India.
[2] Deputy Director, RSST, India.
[3] Information Science & Engineering, RVCE, India.
[4] Information Science & Engineering, RNSIT, India.

*Corresponding Author: [1] sagarbm@rvce.edu.in, [2] cauverynk.rvic@rvei.edu.in, [3] padmashreet@rvce.edu.in, [4] rajkumar.manju@gmail.com

Abstract **- One of the main sources of revenue and growth in Indian economy is from agriculture. It is often a gamble for the farmers to obtain a decent yield, considering the unpredictable environmental conditions. This paper deals with the prediction of the yield of rice and wheat using machine learning algorithms using the annual crop yield production and the annual rainfall in the different districts of India. In this paper, a popular prediction model is developed using algorithms such as decision tree and random forest to predict the yield of most widely grown crops in India like rice and wheat. The features used were the area of production, rainfall, season and state. The season and the state were one hot encoded features. Mean square error was used to measure the loss. The dataset was prepared by combining the crop production in the various states and the rainfall dataset in the respective states.**

Index Terms—Machine Learning, XGBoost, Decision Tree, Random Forest, Data Preprocessing, Data Visualization, Prediction.

1. Introduction

In agriculture the yield is the amount of crops grown or products from the crops such as milk, wool or meat produced per unit area of land. As an alternative technique, the agricultural productivity can also be calculated using the seed ratio.

Rice is a type of seed originating from the grass species. It is considered as the staple food for most of the countries especially in Asia. It is a cereal crop and it is one of the most important commodities from agriculture point of view.

Wheat is also a grass species grown for its seed.It is a cereal grain which is widely consumed as staple food. Wheat requires different conditions to grow as compared to rice and is also an important commodity for agricultural trade.

In earlier days the farmer's experience was considered for the effective production but was of false hope. Due to the daily change in the various conditions, farmers feel the compelling need to produce more crops. Due to this present situation, many of the farmers are not having enough information and knowledge regarding the new crops and their farming. Also, they are not completely educated regarding the benefits they get while farming the new crops. In this regard, understanding the use of technology and forecasting crop performance in different environmental conditions, the productivity of various crops can be increased and is of beneficial importance to the farmers.

In the proposed system, Machine learning techniques are used and prediction algorithms like decision trees and random forest to predict rather accurately which crop cultivated yield better production at the present environmental conditions.

Rice and wheat is a primary staple diet throught out India and the requirement for it's large production is a must. There must be a method to increase the production or to grow it in ways that can maximize the yield over time and meet the demands of the people. This model helps to find out the yield what it can be for the crops rice and wheat and what the effective conditions are in terms of rainfall, to optimize the yield of the crops.

This model implements the use of decision trees and and random forest classifier, machine learning algorithms to predict the yield of various crops. The dataset used consists of the the following crop yield production on a yearly basis in the different districts of Punjab along with the rainfall on a yearly basis in these districts respectively.

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)

The decision tree algorithm uses only the essential features from the set of existing features in giving results for the desired output. However the drawback of this algorithm is that there is a probability for overfitting as it iterates through the dataset. Therefore the results are tightly bound to the dataset and if a network trained on this algorithm is used to predict future crop yield there will most probably be an error in the result. In order to overcome this problem the random forest classifier is used yielding higher accuracy results.

This model helps the farmers to increase their productivity of their respective fields and the wastage is minimized. Being the current day situation where the population is increasing implying the increase in demand this model will be useful for the farmers to maximize the yield and also to maximize their profits

2. Literature survey

In paper[1], the process of prediction of crop yield involves two stages. First stage is to analyse and then second is to categorize them using data mining algorithms. The authors have implemented the classification methods of Naive Bayes and K-Nearest Neighbour for crop yield prediction using the soil dataset. The results of the experiment classifies the soil into categories of low, medium and high. The soil with high confidence value can be used for larger crop yields, medium is for average crop yield and low value infers that the crop yield may be less than average. The dataset size is limited and as a future work, authors claim to use a larger dataset for higher accuracy.

The authors in [2] describe the crop prediction process with respect to the dataset of rainfall, crop type and soil type for a given land. The main algorithm used here is K-means and fuzzy logic. The dataset is used to generate the rule based fuzzy inference system. Both the algorithms are applied to predict the type of crop suitable for the anticipated rainfall and soil type to generate high production in crop yield. About 20 fuzzy rules are formulated and the type of crop is given as input for the system that predicts the production of the crop.

The system implemented is a useful tool for farmers to predict the productivity of the upcoming season and increase their productivity and thus provides by minimizing the losses. The authors do not consider the current weather conditions to predict the crop yield but instead use the data from previous few years. Also, there is no standard quantifiable metric used in the system for rainfall or crop yield.

As per the discussions in the paper [3], the authors have employed different data mining techniques in the field of agriculture. Various techniques include K-means, support vector machine and artificial neural networks to predict the crop yield with different climatic parameters like temperature, rainfall, evapotranspiration, wet day, etc. The authors have implemented C4.5 algorithm to develop decision tree and rules for prediction and an accuracy score of about 90% was achieved for selected crops and over 75% for all the crops. The dataset used in this system was for more than 20 years of data.

The paper [4] gives an analysis of yield prediction by using the environmental parameters such as area used for cultivation, rainfall recorded annually and the food price index. The algorithms used are Regression Analysis and Linear Regression with above parameters for crop yield prediction. The dataset used for analysis is for a period of 10 years. The algorithms use the environmental parameters as exploratory and response variables that helps to make the decision. The results depicts that the parameters influence an average of 70% for the crop yield. The paper lacks usage of weather conditions, minimum support price and soil parameters for yield prediction.

Weather forecasting and crop prediction both are integral part of farming [5]. The application built here uses collaborative filtering technique with weather factors such as temperature, wind speed, humidity, etc., for predicting the crop yield. The application also predicts the air quality index that predicts the pollution in the air. Multiple open source APIs are used to extract the data and analysis is carried out on it for prediction. The application suggests the best crop for the conditions that prevails for a better yield. The paper does not discuss on the results and accuracy of the prediction technique used.

3. The Proposed Model

A. METHODOLOGY

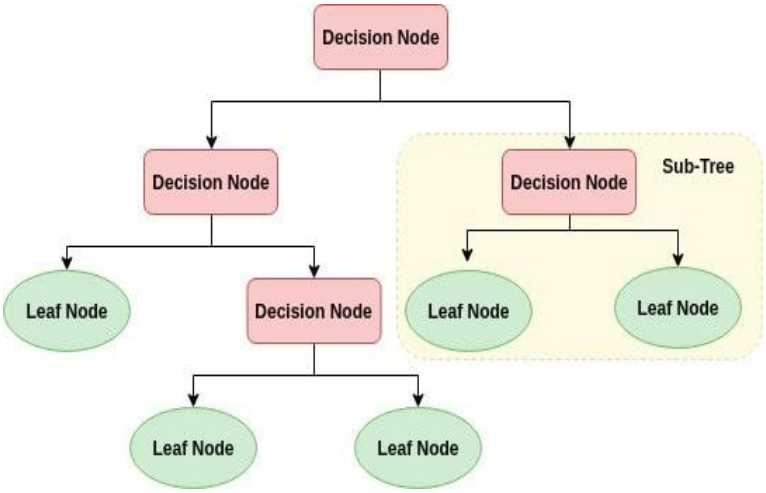
The proposed model makes use of three well-known machine learning algorithms applied to the customized input datasets for rice and wheat separately. They are:

1. Decision Tree Algorithm

It is a supervised learning algorithm, based on condition based decision making, which is applicable for both classification and regression. In Decision Trees, in order to predict a class label of an input record, the decision making starts from the **root** of the tree and classifies the input records by passing them down from the root to some terminal node of the tree, with the terminal node giving the classification of the input.

Each node in the tree is responsible for testing one feature, and every edge going down the node corresponds to plausible answers to the tested feature. This method is recursive and is repeated for every non leaf node.

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)



The following table shows the root mean square error (RMS) values using this algorithm:

Crop	RMS Value
Rice	0.02538
Wheat	0.02198

2. Random Forest Classifier

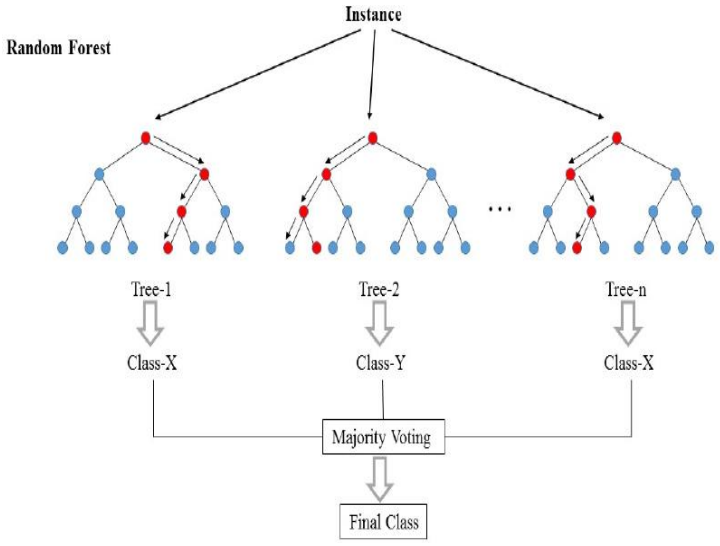
Random Forest is an ensemble learning technique, which uses a combination of a number of machine learning algorithms to obtain better performance in prediction.

Two key concepts of the classifier:

a) Training data is randomly sampled while building the trees.
b) Subsets of features are chosen randomly while splitting the nodes.

Bagging technique is used to create an ensemble of trees where many subsets of training data are produced with replacement.

In this technique, a dataset is divided into a number of subsets using randomization. Based on a single learning technique, a classifying model is built based on the samples of subsets. A classification with most votes, as a contribution of many trees in the forest is considered as the result for the corresponding input.



COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING

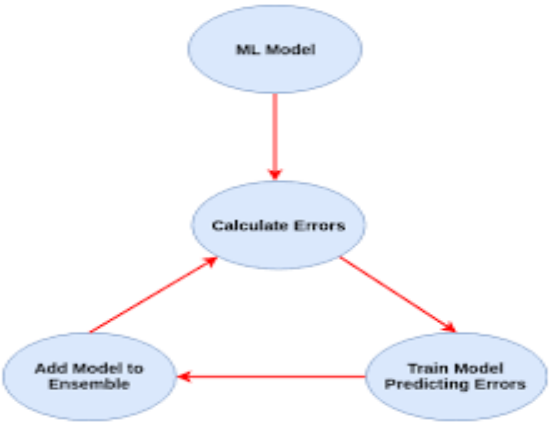
Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)

The following table shows the root mean square error (RMS) values using this algorithm:

Crop	RMS Value
Rice	0.01942
Wheat	0.01680

3. XGBoost Classifier (Extreme Gradient Boost Classifier)

XGBoost is an ensemble Machine Learning algorithm and uses a gradient boosting framework based on decision trees. It makes use of system optimization techniques such as parallelization, tree-pruning, to avoid overfitting of the model, and hardware resource optimization. It also uses algorithmic enhancements to yield results in the shortest possible time.



B. IMPLEMENTATION

1. Data Preprocessing

The crop yield dataset used in the proposed model is take from the website ‘data.gov.in’, titled ‘District-wise Season-wise Crop Production Statistics from 1997-2015’ and the rainfall dataset used is taken from Kaggle, titled ‘rainfall in India 1901-2015’.

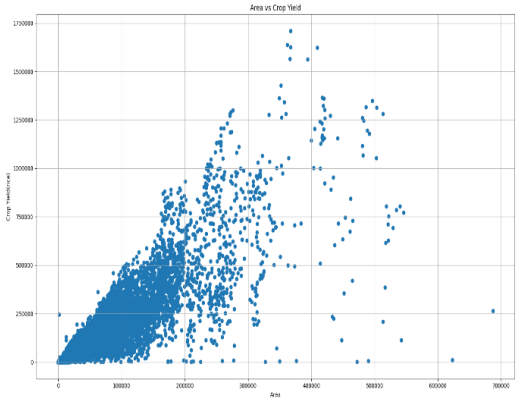
Out of many crops present in the dataset, it is split into two datasets, a rice dataset and a wheat dataset. For each of these datasets, the rainfall statistics is mapped according to the corresponding seasons, district and year. All the null values are filled with mean values. Thus, two processed datasets, containing crop yield and rainfall statistics are obtained.

2. Data Visualization

Several plots are plotted to understand the statistical data visually, for both rice and wheat. Units of measure for area of production, rainfall and crop yield are metres, millimetres and kilograms respectively.

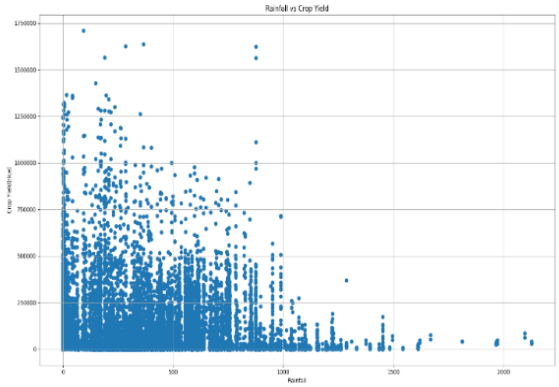
Rice

Area of Production v/s Crop Yield

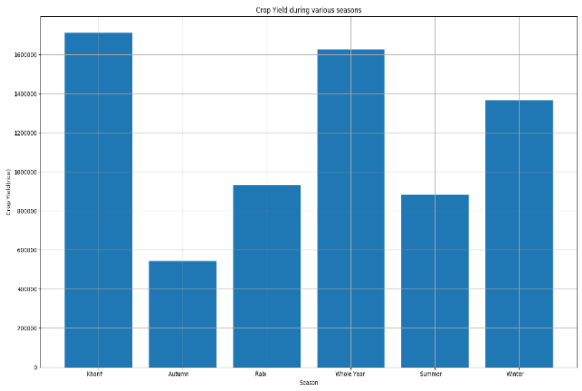


Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)

Rainfall v/s Crop Yield

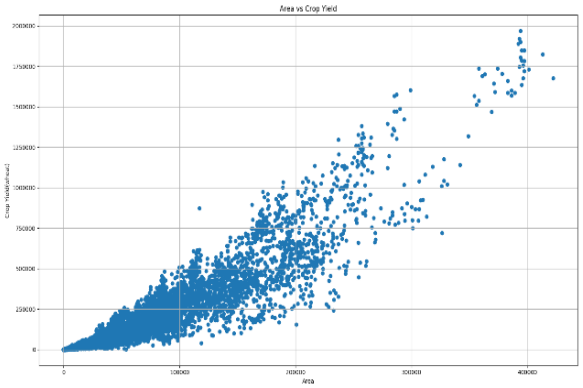


Seasons v/s Crop Yield

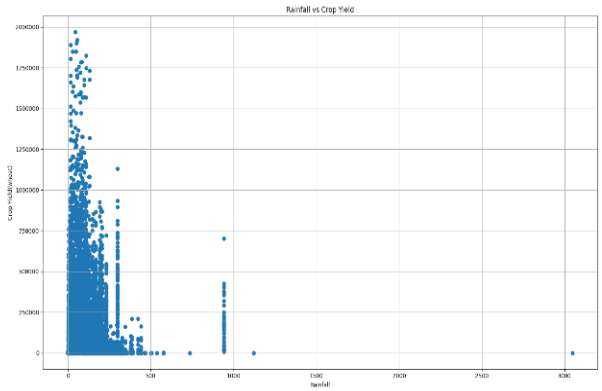


Wheat

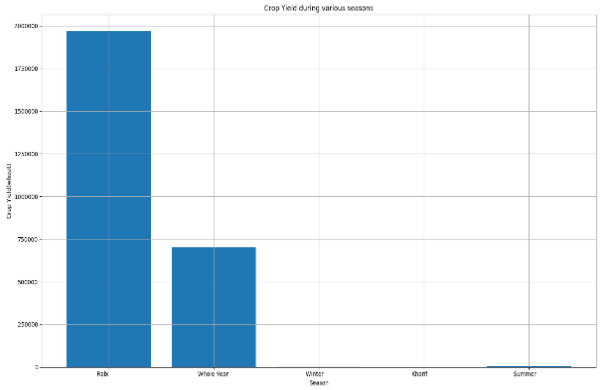
Area of Production v/s Crop Yield



Rainfall v/s Crop Yield



Seasons v/s Crop Yield



From the above plots, it can be interpreted that crop yield is linearly related to area. It is also observed that rice is mainly grown in Kharif season and wheat in Rabi.

Hence all these features are vital in predicting crop yield.

3. Feature Extraction

The features considered are state name, district name, area of production, yield, mapped rainfall data and season.

All the categorical features like state name, district name and season are one-hot encoded.

Area of production and yield are normalized, and the 'NaN' values if any are filled with mean values, for each of the features. Mean value method has been used because all the features considered are numeric and statistical.

COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)

All features except yield are considered to predict the yield. The dataset including these features is split into training data (75%) and testing data (25%) and fed into the above-mentioned Machine Learning models.

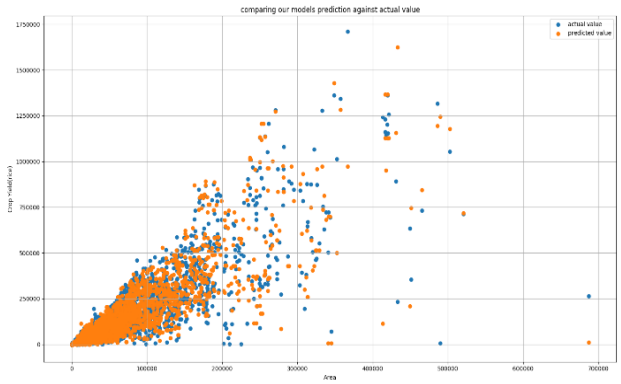
4. Training and observations

The training data is fit into models from 'sklearn library'. All parameters for the regressor are set to default. The results obtained are visualised in the graphs shown below:

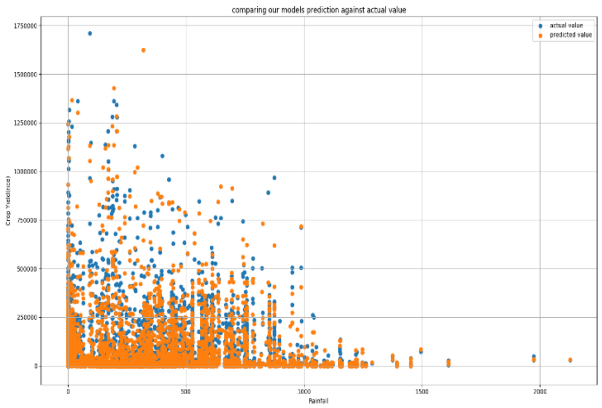
a) Decision Tree Regressor:

Rice

Area v/s crop yield

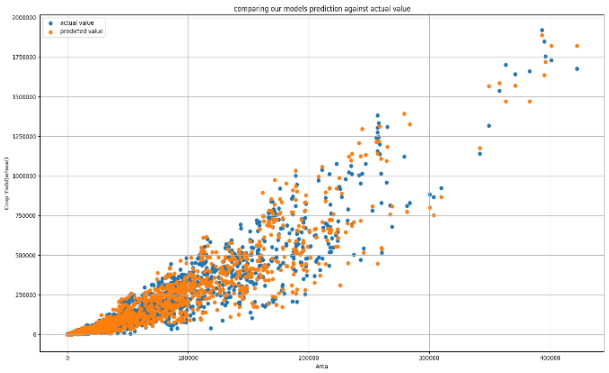


Rainfall v/s crop yield

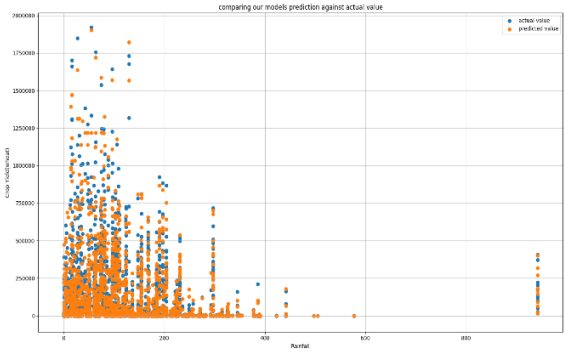


Wheat

Area v/s crop yield



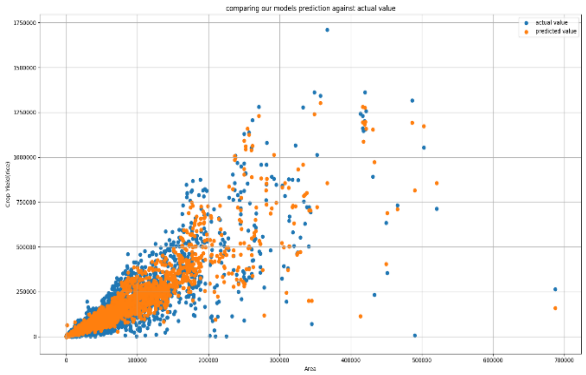
Rainfall v/s crop yield



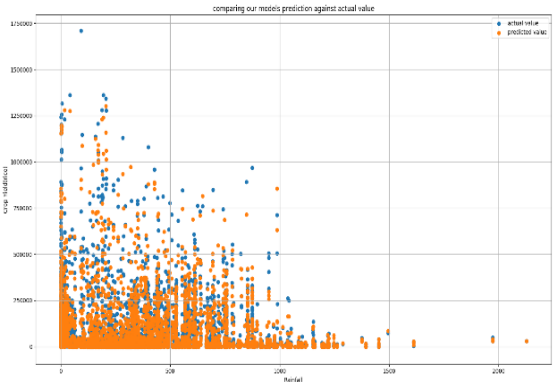
b) Random Forest Classifier

Rice

Area v/s Crop Yield



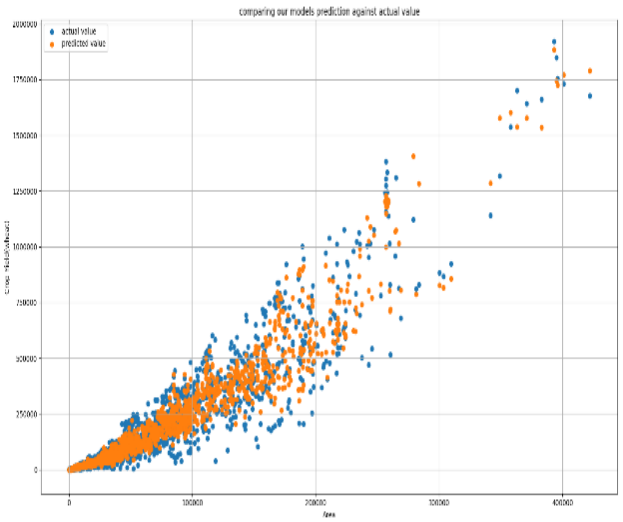
Rainfall v/s crop yield



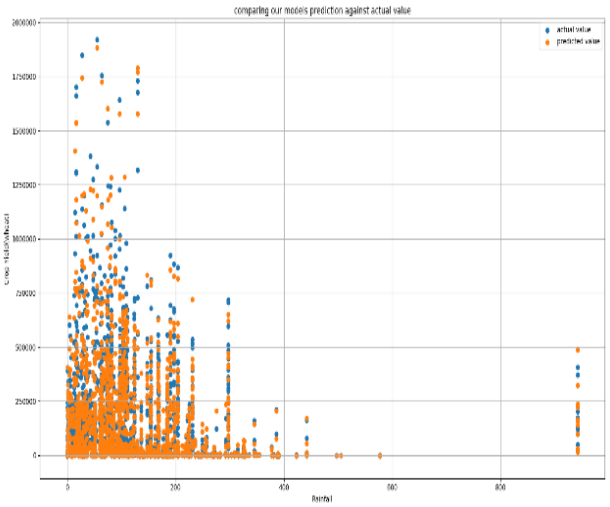
Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.1-8
ISSN : 2582-7464 (Online)

Wheat

Area v/s crop yield



Rainfall v/s crop yield



4. Results

In this paper, the algorithms such as decision tree and random forest are used to predict rice and wheat yield. The Mean square error has been used to measure the loss incurred by the model.

ALGORITHM	CROP	MSE
Decision Tree	Rice	0.0006418
	Wheat	0.0004834
Random Forest	Rice	0.0003772
	Wheat	0.0002825

The results clearly show that random forest algorithm fared much better than decision tree algorithm.This behaviour is expected as random forest algorithm is an ensemble model. Due to this, randomness in induced in the model which leads it to generalize better than stand alone models. This can also be seen in the graphs for decision tree and random forest (the plots for decision tree show that the model is slightly overfitting when compared to random forest model.)

5. Conclusion

From the results, it can be observed that decision tree algorithm overfits as mentioned earlier, and random forest does not

These experiments also showed that rice yield was dependent on the amount of rainfall experienced where as wheat yield was not affected much by it.

This paper can be further extended by adding more features such as temperature, soil conditions, etc., and using more advanced models such as XGBoost.

REFERENCES

[1] M. Paul, S. K. Vishwakarma and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 766-771, doi: 10.1109/CICN.2015.156.

[2] Awanit Kumar et al, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.8, August- 2015, pg. 44-56

[3] S. Veenadhari, B. Misra and C. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters," 2014 International Conference on Computer Communication and Informatics, 2014, pp. 1-5, doi: 10.1109/ICCCI.2014.6921718.

[4] V. Sellam and E. Poovammal, "Prediction of Crop Yield using Regression Analysis", Indian Journal of Science and Technology, 2016, Volume: 9, Issue: 38, Pages: 1-5, DOI: 10.17485/ijst/2016/v9i38/91714.

[5] Anil Hulsure , Yogesh Kale , Aditya Kalekar , Nandini Banswani, Vijay Ganesh, 2021, Weather Forecasting & Crop Recommendation, International Journal Of Engineering Research & Technology (IJERT) Volume 10, Issue 05 (May 2021),

DATASET REFERENCES:

[6] https://data.gov.in/catalog/district-wise-season-wise-crop-production-statistics?filters%5Bfield_catalog_reference%5D=87631&format = json&offset=0&limit=6&sort%5Bcreated%5D=desc

[7] https://www.kaggle.com/rajanand/rainfall-in-india