Use of Random Forest Regression Model for Forecasting Food and Commercial Crops of India

Ali J. Ramadhan[1*], *S R Krishna* Priya[2], *N* Naranammal[2], *Suman*[3], *Priyanka* Lal[4], *Pradeep* Mishra[5], *Mostafa* Abotaleb[6] and *Hussein* Alkattan[6]

[1]University of Alkafeel, Najaf, Iraq
[2]PSG College of Arts & Science, Coimbatore, India
[3]SGT University, Gurugram, India
[4]Lovely Professional University, Phagwara, India
[5]JNKVV College of Agriculture, Rewa, India
[6]South Ural State University, Chelyabinsk, Russia

Abstract. Agriculture is the backbone of Indian Economy. Proper forecast of food crops and cash crops are necessary for the government in policy making decisions. The present paper aims to forecast Wheat and Sugarcane yield using Random Forest Regression. For the development of Random Forest models, Yield has been taken as dependent variable and variables like Gross Cropped Area, Maximum Temperature, Minimum Temperature, Rainfall, Nitrogen, Phosphorous Oxide, Potassium Oxide, Minimum Support Price and Area under Irrigation are taken as independent variables for both Wheat and Sugarcane crop. Values of R^2 for Wheat and Sugarcane is 0.995 and 0.981which indicates that the model is a good fit and other performance measures are calculated and results are satisfactory.

1 Introduction

In developing countries such as India, Agriculture is major source of employment and it contributes to the growth of Indian economy. In India crops are divided into four majorcategories - Food crop, Commercial crop, Plantation crop and Horticulture crop. Food crops such as Wheat provides food for human consumption. The most important food crop in India after Rice is Wheat. It is a rabi crop which is sowed in winter season (October – February) and harvested in spring season (March – April). India stands second in production of Wheat after China. Wheat crop is more flexible in terms of climatic and other conditions of growth. Wheat is grown in soil type like clay loam and sandy loam. Among the commercial crops, India stands in the second position in Sugarcane production after Brazil. It is grown in tropical and subtropical areas. Sugarcane is mainlygrown for sale in market and for industrial raw materials rather than for family consumption or to feed livestock.

Considerable work has been done in crop yield forecasting using Machine Learning Algorithm. Adil et al., [1], Sellam and Poovammal[9], Shivani and Preeti [10] Ujjainia et al., [12], have used Regression Analysis model for crop yield forecasting.

* Corresponding author: ali.j.r@alkafeel.edu.iq

Devika and Ananthi [3], Jyothi and Bhargavi [5], Ramesh and Vijay [8] and Veenadhari et al., [13] have used data mining algorithms for forecasting yield of different crops. Random Forest algorithm has been used by Berima [2], Everingham et al., [4] and Rahman et al., [7]for crop yield forecasting.

The present study is an attempt to use Random Forest Regression model for forecasting Wheat and Sugarcane yield.

2 MATERIAL AND METHODS

2.1 Data Description

All India annualdata on Wheat yield (in '000 kg/hectare) from year 1975 to 2017 and all India annualdata on Sugarcane yield (in '000 tonnes/hectare) from year 1951 to 2017 are taken as dependent variable for the development of the model. Gross cropped area (in '000 hectare), Maximum temperature (◦c), Minimum temperature (◦c), Rainfall (mm), Fertilizer consumption data(Nitrogen, Phosphorous Oxide, Potassium Oxide), Minimum support price, Area under irrigation (in '000 hectare) are taken as independent variables. Last 5 years (2013 – 2017) is used for validation of the model. Data is collected from different sources such as Indiastat.com, Ministry of Statistics and Programme Implementation [6] and The Fertilizer Association of India [11].

2.2 Machine Learning (ML) Algorithm

Figure 1 show the steps followed in Machine Learning algorithm.

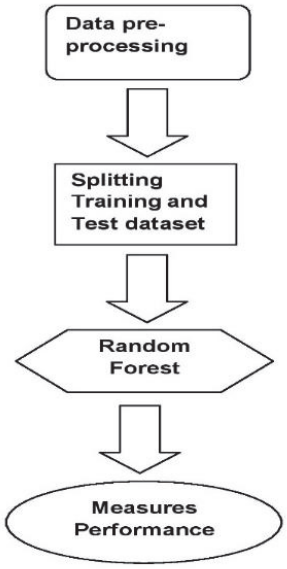

Fig. 1. Flowchart of Machine learning algorithm

2.2.1 Data pre-processing

Data pre-processing is a first and important step in ML. First, it is needed tocheck whether there is any null value(s) in the dataset, then it is needed to check whether the dataset is linear or nonlinear and the presence of outlier, multicollinearity. Following that,dependent and independent variables are defined.

2.2.2 Training and Testing data set

Training dataset is used for model building and testing dataset is used for validating the model. After Model building,the prediction is made andcompared with the testing data set and percentage of deviation is calculated.

In the present study,wheat yield data for 38 years(1975 - 2012)has been used as training set and remaining5years data (2013- 2017)has been used as testing set. Similarly,sugarcane yield data for 62 years (1951 - 2012) has been used as training set and data for 5 years (2013 -2017) has been used as testing set respectively.

2.2.3 Random Forest Algorithm

Random forest is a robust ML algorithm which can be used forregression and classification. Random Forest is an ensemble method, that the model is made up of a large number of small decision trees known as estimators, each producing its own prediction. Final prediction values are taken from average of each prediction which improves the accuracy of prediction.

2.2.4 Performance Measures

The execute random forest model has been appraised by using performancemeasures which include R-squared value, Relative importance, Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, Mean Absolute Percentage Error.

2.2.4.1 R – Square

$$R^2 = 1 - \frac{RSS}{TSS} \quad (1)$$

Where, RSS= Sum of Square of residuals and TSS = Total Sum of Square

2.2.4.2 Relative importance

Itcalculates the importance rank of variables. It shows which variable is important and rank the variable based on their contribution to R^2. It is calculated by the sum of the error reduction when divided by a variable.

2.2.4.3 Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

2.2.4.4 Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (3)$$

2.2.4.5 Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (4)$$

2.2.4.6 Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (5)$$

Where,

n- number of observations y_i – actual value and $\hat{y}_i$ – predicted value

3 RESULTS AND DISCUSSION

From table 1, R^2 values for wheat and sugarcane is 0.995 and 0.981 respectively which is high and the error values such as MSE, MAE, RMSE and MAPE are low. It indicates that the model is good fit.

Table 1. Performance Measures of Random Forest Regression for Wheat and Sugarcane Yield

Performance Measures	Wheat	Sugarcane
R^2	0.995	0.981
MSE	0.059	31.308
MAE	0.219	3.978

Performance Measures	Wheat	Sugarcane
RMSE	0.242	5.595
MAPE	6.951	5.234

Table 2 shows the relative importance of the independent variables for wheat yield forecasting. From the table, it is clear that Minimum temperature, Maximum temperature and Nitrogen are very important variables for wheat yield followed by other variables.

Table 2. Relative importance of variables for wheat yield forecasting

Rank	Variables	Relative Importance
5	AUI	0.247
7	MSP	0.226
2	N	0.203
3	P2O5	0.128
4	K2O	0.091
6	GCA	0.082
0	MINTMP	0.010
1	MAXTMP	0.006
8	Rainfall	0.005

Figure 2 shows the graph of relative importance of independent variables in yield forecasting of wheat crop.

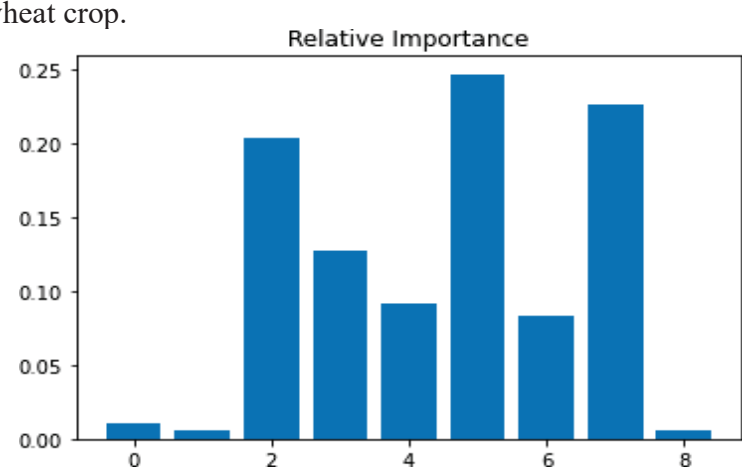

Fig 2. Graphical representation of Relative importance of variables in wheat yield

Table 3 shows the relative importance of the independent variables for sugarcane yield forecasting. From the table, it is clear that Rainfall, Minimum temperature and Maximum temperature are very important variables for sugarcane yield followed by other variables.

Table 3. Relative importance of variables for Sugarcane yield forecasting

Rank	Variables	Relative Importance
4	N	0.518
5	P2O5	0.214
6	K2O	0.111
3	GCA	0.105
1	MINTMP	0.029
2	MAXTMP	0.016
0	Rainfall	0.008

Figure 3 shows the graph of relative importance of variables in yield forecasting of sugarcane crop.

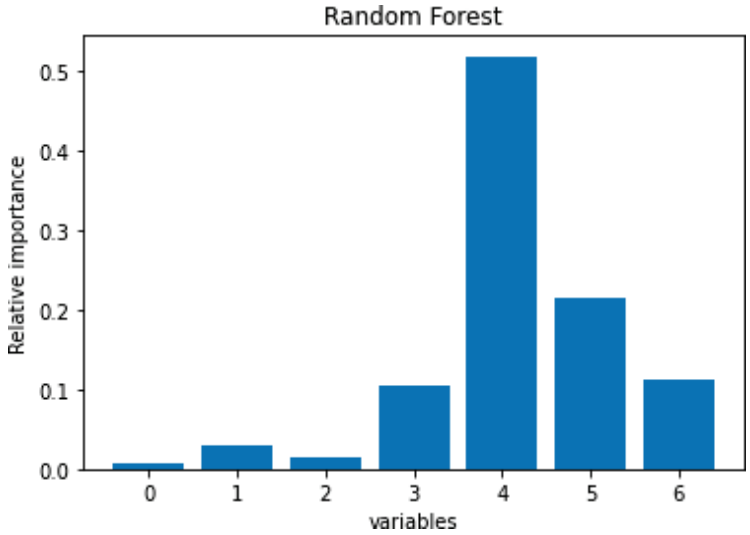

Fig. 3. Graphical representation of Relative importance of variables in sugarcane yield

From tables 4 and 5, there is upward trend in yield of both crops. The percentage of deviation of both yield forecastsof testing data set is low which indicate that the model is good fit.

Table 4. Actual and Forecasted value of Wheat yield

Year	Actual Value (in '000kg/ hectare)	Predicted value (in '000kg/ hectare)	% Of deviation
2013	3.146	2.933	0.21
2014	2.75	2.934	-0.18
2015	3.034	2.96	0.07
2016	3.200	2.97	0.22
2017	3.371	2.98	0.396

Table 5. Actual and Forecasted value of Sugarcane yield

Year	Actual value (in '000 tonnes/hectare)	Predicted value (in '000 tonnes/hectare)	% Of deviation
2013	70.522	67.245	3.28
2014	71.51	68.42	3.09
2015	70.72	68.92	1.81
2016	69.001	68.81	0.19
2017	80.198	68.67	11.53

Figures 4 and 5 show the actual and predicted value of wheat and sugarcane yield forecast obtained from the Random Forest Regression Model respectively.

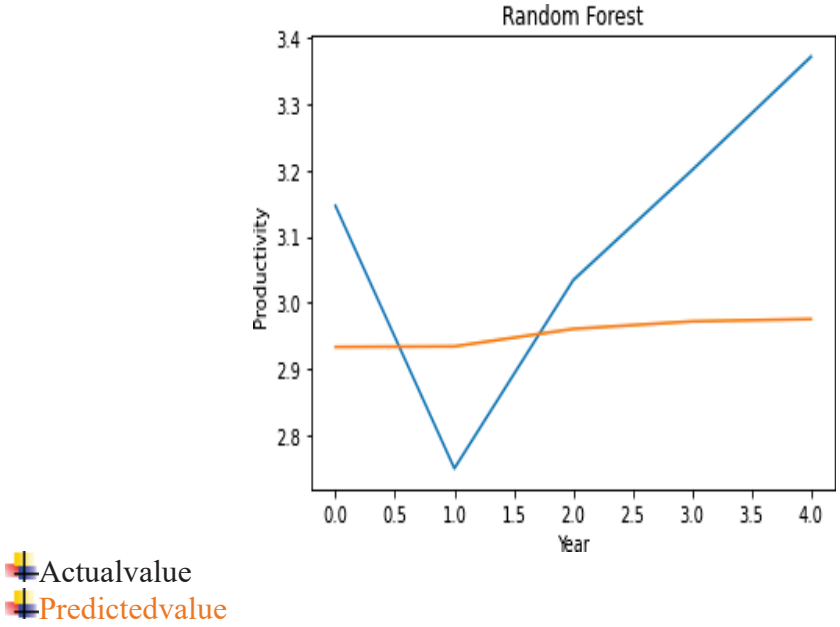

Fig. 4. Actual Vs Predicted graph of wheat yield forecast

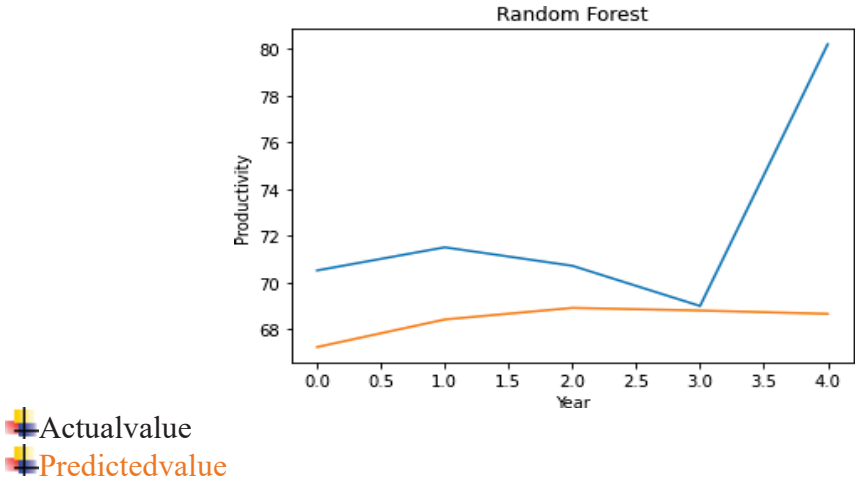

Fig. 5. Actual Vs Predicted graph of sugarcane yield forecast

4 CONCLUSIONS

The present study has been undertaken to inspect the forecasting ability of Random Forest Regression model and to predict the Yield of the Wheat andSugarcane cropsof India. From the results, Random Forest Regression model is good fit model and gives good accuracy for Wheat and Sugarcane crop yield because it is average of many possible predictions. Future researchers can use this model and add the upcoming years and can predict the yield and researchers can also try other machine learning algorithms and compare it with Random Forest regression.

REFERENCES

1. AdilN,DewanganS, Sharma K. Efficient Classification and Regression Techniques to Predict Crop Yield.International Journal of Scientific and Technology Research. 2019;08(11):378-382.
2. Al-Mahdawi, H. K., Albadran, Z., Alkattan, H., Abotaleb, M., Alakkari, K., & Ramadhan, A. J. (2023, December). Using the inverse Cauchy problem of the Laplace

equation for wave propagation to implement a numerical regularization homotopy method. AIP Conference Proceedings (Vol. **2977**, No. 1). AIP Publishing.

3. DevikaB, AnanthiB. Analysis of Crop Yield Prediction Using Data Mining Technique to Predict Annual Yield of Major Crops.International Research Journal of Engineering and Technology.2018;**05**(12): 1460 - 1465.
4. EveringhamY, Sexton J, SkocaiD, Bamber GI. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development.2016;**36**(2):27.
5. Al-Nuaimi, B. T., Al-Mahdawi, H. K., Albadran, Z., Alkattan, H., Abotaleb, M., & El-kenawy, E. S. M. (2023). Solving of the inverse boundary value problem for the heat conduction equation in two intervals of time. Algorithms, **16**(1), 33.
6. Ministry of Statistics and Programme implementation. Government of India (ON1962) &(ON1964).
7. RahmanAEM, AhmedFB, Ismail R.Random Forest Regression and Spectral Band Selection for Estimating Sugarcane Leaf Nitrogen Concentration Osinz EO-1 Hyperion Hyperspected Data. Int J Remate Sens.2013; Doi:10.1080/01431161.2012.713142.
8. RameshAM, VijaySR. A Survey of Data Mining Techniques for Crop Yield Prediction.International Journal of Advance Research in Computer Science and Management Studies.2014;**02**(09): 59 – 64.
9. Akbari, E., Mollajafari, M., Al-Khafaji, H. M. R., Alkattan, H., Abotaleb, M., Eslami, M., & Palani, S. (2022). Improved salp swarm optimization algorithm for damping controller design for multimachine power system. IEEE Access, **10**, 82910-82922.
10. ShivaniSK, PreetiSP. Prediction of Sugarcane Yield from Field Records Using Regression Model.International Journal of Recent Technology and Engineering.2019;**08**(04):1603-1606.
11. Ehsan khodadadi, S. K. Towfek, Hussein Alkattan. (2023). Brain Tumor Classification Using Convolutional Neural Network and Feature Extraction. Fusion: Practice and Applications, **13**(2), 34-41.
12. UjjainiaS, Gautham P, Veenadhari S. Crop Yield Prediction Using Regression Model. International Journal of Innovative Technology and Exploring Engineering.2020;**07**(05):269-273.
13. VeendhariS, MisraB, Singh D. Data Mining Technology for Predicting Crop Productivity –A Review Article, International Journal of Science and Technology.2011.