E-ISSN: 2618-0618
P-ISSN: 2618-060X

www.agronomyjournals.com
2024; SP-7(8): 817-820
Received: 08-05-2024
Accepted: 13-06-2024

Manoj Varma
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

Praveen Kumar A
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

Kaushal Kumar Yadav
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

Satyam Verma
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

Prabhat Kumar
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

Ankit Kumar Singh
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

Corresponding Author:
Praveen Kumar A
ICAR - Indian Agricultural Statistics Research Institute, New Delhi, India

A comparison of random forest based models for agricultural crop yield prediction

Manoj Varma, Praveen Kumar A, Kaushal Kumar Yadav, Satyam Verma, Prabhat Kumar and Ankit Kumar Singh

DOI: https://doi.org/10.33545/2618060X.2024.v7.i8Sk.1386

Abstract
Accurate crop yield forecasting is important for various stakeholders in the agri-food chain, including farmers, agronomists, commodity traders, and policymakers. Accurate and timely forecasts enable informed decision-making and better planning in the face of agriculture's inherent uncertainties. Machine learning can be leveraged to predict crop yields, guide crop selection, and inform actions during the growing season. Since crop yield depends on numerous factors, identifying the most significant variables is essential to enhance prediction accuracy. Feature selection algorithms help focus on relevant features, improving model performance while reducing computational time. This method is especially useful in agriculture, where yield is affected by factors like land use, water management, fertilizer application, and weather conditions. In this study, the Random Forest algorithm was applied both to develop a regression model and to perform feature selection across three different datasets. A multiple linear regression model was then built using the selected features.
The models' forecasting performance was assessed using statistical metrics, including Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Deviation (MAD). A comparison between Random Forest-based feature selection and Random Forest regression showed that the RF regression model delivered the most accurate forecasts across the datasets.

Keywords: Weather indices, feature selection, random forest and regression

1. Introduction
The impact of weather on crop growth varies depending on the crop's developmental stage. The influence on yield depends not only on the intensity of weather variables but also on how these variables are distributed across different growth stages, as different stages have varying sensitivities to weather fluctuations. To achieve precise forecasts, it is essential to divide the crop growth phase into narrow intervals and consider interactions between weather parameters in crop-weather models (Agrawal *et al*., 1983) [1]. Machine learning (ML) is well-suited for enhancing crop yield estimates as it uses data-driven techniques to identify patterns and relationships in input data. ML algorithms can establish functions that link predictors to outcomes like crop yield, simulating non-linear correlations between various data sources. By integrating the strengths of earlier technologies, such as crop growth models and remote sensing, with data-driven modeling, ML provides a flexible and effective approach for modeling and forecasting crop yield.
Random forests are highly effective for forecasting, as they do not overfit due to the Law of Large Numbers (Breiman, 2001) [6]. Combining machine learning with empirical domain knowledge enhances predictive accuracy (Droesch, 2018) [7]. Moreover, small sample sizes can lead to higher classification accuracy (Vabalas *et al*., 2019) [14], and accurate models have been developed using small datasets (Zhang *et al*., 2018). Machine learning (ML) techniques, with their minimal prior assumptions and data-driven nature, offer significant flexibility for crop yield modeling and forecasting. Various researchers have successfully applied ML techniques for crop yield forecasting, achieving satisfactory results (Droesch, 2018; Gopal and Bhargavi, 2019) [7-9]. Discriminant function analysis has been used to develop wheat yield forecast models, highlighting that rainfall and temperature during key growth

periods significantly impact wheat yield (Agrawal *et al*., 2012) [3]. This increases the number of variables in the model, requiring the evaluation of more parameters, which can be challenging with limited data. The solution is to find a model with a small, easily evaluated set of parameters that also accounts for the distribution of weather patterns throughout the crop growth period.

Given the numerous input parameters influencing crop yield, identifying key variables and excluding redundant ones is essential to improve predictive model accuracy (Springenberg *et al*., 2014) [12]. Feature selection algorithms help focus on relevant features, improving model performance and reducing computational time (Oreski *et al*., 2017; Gopal and Bhargavi, 2019) [11, 8, 9]. This is especially important in agriculture, where yield depends on factors such as land use, water management, fertilizer application, and weather. In Random Forest models, high-importance variables greatly influence outcomes, while low-importance variables can be eliminated to simplify and speed up the model (Breiman, 2001) [6]. Random Forest has proven to be highly effective for sugarcane yield modeling using data from a sugarcane mill (Bocca *et al*., 2016) [5].

2. Materials and Methods

In this study, the process started with gathering data on weekly weather parameters, such as minimum and maximum temperatures, relative humidity, total precipitation, mean temperature, and pressure, along with crop yield data for various districts. During the data preparation phase, weather indices were derived from these weekly parameters. Different weather indices were created for each dataset using specific functions.

Weather indices will be generated using the following expression:

$$Y = A_0 + \sum_{u=1}^{p}\sum_{v=0}^{1} a_{vw} Z_{vw} + \sum_{v'u=1}^{p}\sum_{v=0}^{1} a_{uv'v} Z_{uu'v} + cT + \varepsilon$$

$$Z_{uv} = \sum_{i=1}^{m} u_{ui}^{v} X_{ui}$$

$$Z_{uu'v} = \sum_{i=1}^{m} r_{uu'i}^{v} X_{ui} X_{u'i}$$

Where,

Where, r_{ui} / $r_{uu'v}$ represents the correlation coefficient between the yield and the u[th] weather variable or the product of the u[th] and u'[th] variables in the ith week, m is week of forecast p is number of weather variables used. C is constant, T denotes the year number, included to account for any long-term upward or downward trends in yield. (Agrawal *et al*., 2007) [2]

Weather indices were used in place of raw weather variables and Random Forest regression model have been developed. Also Random Forest-based feature selection algorithm was applied to identify the most important variables for further analysis. Crop yield forecast models were developed using multiple linear regression with the selected features. Forecasts for a testing dataset were generated with both models, and their performance was compared using measures such as Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

3. Data Description

The study utilized three datasets, each covering wheat crop yield and weather parameters (rainfall, maximum and minimum temperature, precipitation, and relative humidity) for different districts in Punjab. The first dataset is for Amritsar, the second for Jalandhar, and the third for Patiala. All weather data for these districts were sourced from NASA's POWER website (https://power.larc.nasa.gov/).

Each dataset was divided, with 80% used for training the model and 20% reserved for validation.

4. Results and Discussion

Accuracy measures such as MAD, RMSE, and MAPE were used to evaluate the prediction models. Lower values indicate better accuracy. For Amritsar, the RF feature selection model had the lowest MAD (683.70), RMSE (749.15), and MAPE (14.38). For Ludhiana, the RF regression model achieved the lowest MAD (343.50), RMSE (433.90), and MAPE (7.16). For Patiala, the RF regression model also had the lowest MAD (508.40), RMSE (588.71), and MAPE (10.62). Thus, the RF feature selection model was best for Amritsar, while the RF regression model excelled for Ludhiana and Patiala.

Table 1: Predicted Wheat Yield Values (Kg/ha) from Regression Models for Amritsar Data

Years	Actual values (Kg/ha)	Predicted values (Kg/ha)	
		RF FS	RF Reg.
2011	4283	3886.70	3804.58
2012	4975	3861.71	3928.26
2013	4654	3920.52	3857.90
2014	4869	3912.86	3718.17
2015	3914	3706.00	3675.70
2016	4478	4122.84	4023.18
2017	4948	4009.24	4213.64
2018	4866	4097.59	4076.73

Table 2: Predicted Wheat Yield Values (Kg/ha) from Regression Models for Ludhiana Data

Years	Actual values (Kg/ha)	Predicted values (Kg/ha)	
		RF FS	RF Reg.
2011	4964	4293.71	4286.50
2012	5375	4332.56	4248.73
2013	4853	4302.89	4305.46
2014	5226	4306.23	4125.14
2015	4462	4261.76	4097.34
2016	4670	4351.89	4489.69
2017	5093	4362.79	4543.26
2018	5144	4366.89	4539.97

Table 3: Predicted Wheat Yield Values (Kg/ha) from Regression Models for Patiala Data

Years	Actual values (Kg/ha)	Predicted values (Kg/ha)	
		RF FS	RF Reg.
2011	4836	4274.01	4102.93
2012	5472	4226.55	4010.75
2013	4798	4241.70	4061.30
2014	4968	4256.43	4046.94
2015	4496	4266.21	4045.80
2016	4585	4174.89	4316.97
2017	5165	4169.70	4368.99
2018	5272	4188.08	4229.75

Table 4: Comparison of regression models for wheat yield (Kg/ha) in Amritsar District

Algorithm	MAD		RMSE		MAPE	
	Training	Testing	Training	Testing	Training	Testing
RF FS	383.80	683.70	500.11	749.15	10.58	14.38
RF Regression	150.27	711.11	208.64	767.19	4.18	15.03

Table 5: Comparison of regression models for wheat yield (Kg/ha) in Ludhiana District

Algorithm	MAD		RMSE		MAPE	
	Training	Testing	Training	Testing	Training	Testing
RF FS	349.52	651.04	447.83	703.83	8.36	12.82
RF Regression	420.38	343.50	498.26	433.90	11.86	7.16

Table 6: Comparison of regression models for wheat yield (Kg/ha) in Patiala District

Algorithm	MAD		RMSE		MAPE	
	Training	Testing	Training	Testing	Training	Testing
RF FS	380.94	724.30	462.99	796.03	9.56	14.27
RF Regression	273.75	508.40	355.48	588.71	7.78	10.62

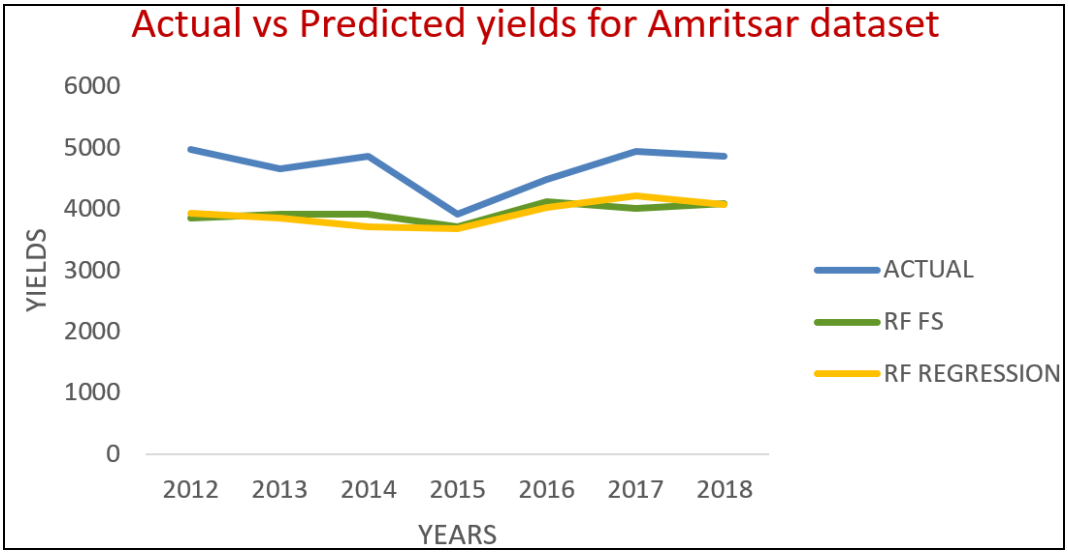

Fig 1: Fitting of RF regression model and RF feature selection based multiple linear regression model for Amritsar district wheat yield (Kg/ha) data

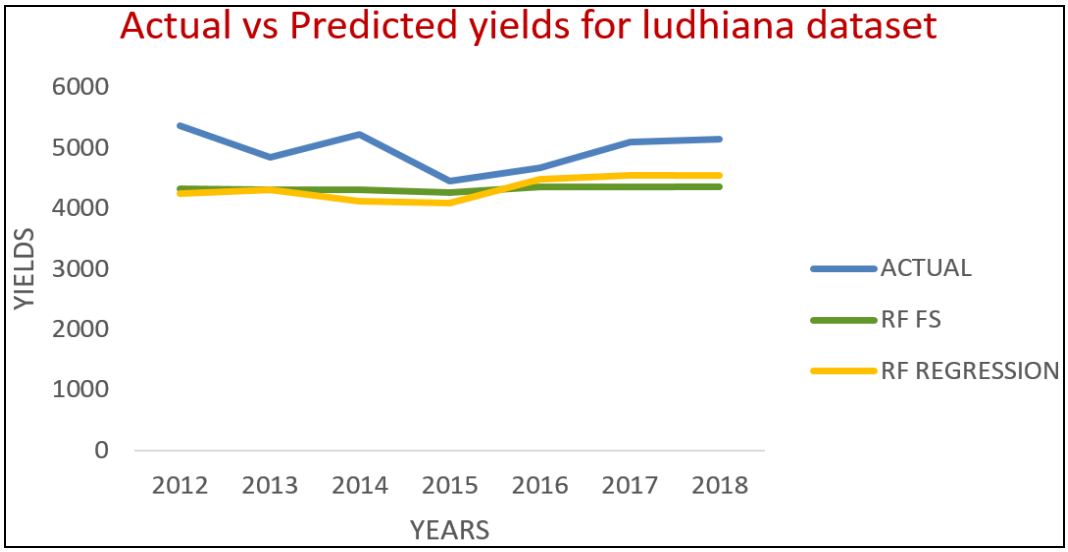

Fig 2: Fitting of RF regression model and RF feature selection based multiple linear regression model for Ludhiana district wheat yield (Kg/ha) data

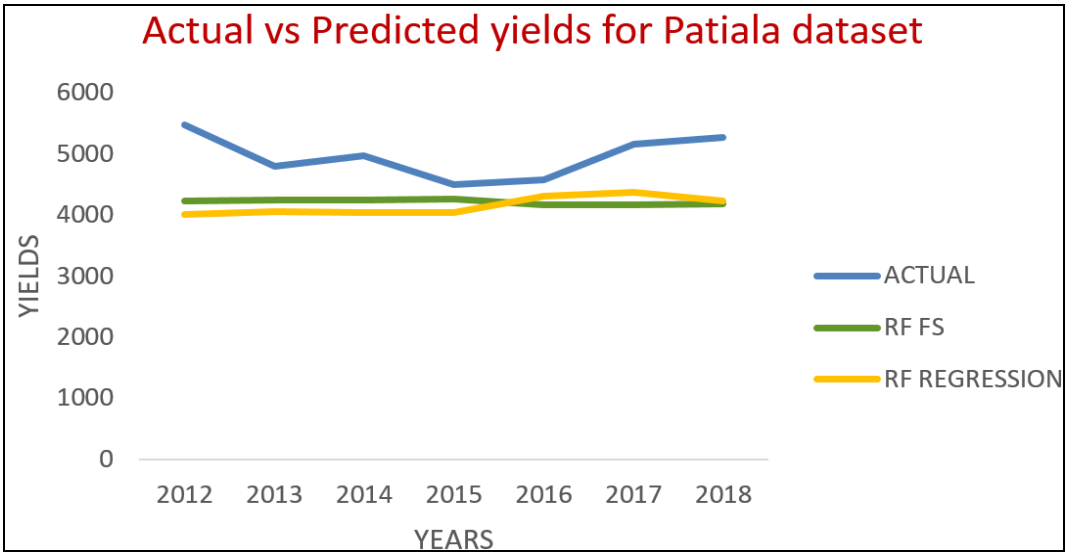

Fig 3: Fitting of RF regression model and RF feature selection based multiple linear regression mode for Patiala district wheat yield (Kg/ha) data

5. Conclusion

This study analyzed three different datasets, each yielding 30 weather indices. The Random Forest algorithm was employed for both regression and feature selection. Among the models examined, the regression model with feature selection achieved the highest prediction accuracy for wheat yield in Amritsar. In contrast, the RF regression model proved to be the most effective for wheat yield predictions in Ludhiana and Patiala districts.

6. Acknowledgement

We would like to extend our sincere appreciation to the India Meteorological Department for supplying the dataset and to the Director of ICAR – IASRI for offering the facilities at the institute.

7. Disclosure statement

The authors declare that there are no known conflicts of interest that could have influenced the work presented in this paper.

8. References

1. Agrawal R, Jain RC, Jha MP. Joint effects of weather variables on wheat yields. Mausam. 1983;34:189-194.
2. Agrawal R, Mehta S. Weather based forecasting of crop yields, pests and diseases - IASRI models. J Indian Soc. Agric Stat. 2007;61:255-263.
3. Agrawal R, HAS C, Aditya K. Use of discriminant function analysis for forecasting crop yield. Mausam. 2012;63(3):455-458.
4. Balogun AO, Basri S, Abdulkadir SJ, Hashim AS. Performance analysis of feature selection methods in software defect prediction: A search method approach. Appl. Sci. 2019;9(13):2764.
5. Bocca FF, Rodrigues LHA. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Comput. Electron Agric. 2016;128:67-76.
6. Breiman L. Random forests. Mach Learn. 2001;45:25-32.
7. Droesch AC. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ Res. Lett. 2018;13:1-12.
8. Gopal PM, Bhargavi R. Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches. Appl. Eng. Agric. 2019;35:9-14.
9. Gopal PM, Bhargavi R. Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. Appl. Artif. Intell. 2019;33:621-642.
10. Huang JZ. Introduction to Statistical Learning: With Applications in R. Springer; c2014.
11. Oreski D, Oreski S, Klicek B. Effects of dataset characteristics on the performance of feature selection techniques. Appl Soft Comput. 2017;52:109-119.
12. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. In: ICLR; c2014. p. 1-14. Available from: https://arxiv.org/abs/1412.6806.
13. Suruliandi A, Mariammal G, Raja SP. Crop prediction based on soil and environmental characteristics using feature selection techniques. Math Comput Modell Dyn Syst. 2021;27(1):117-140.
14. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. c2019, 14(11).
15. Whitmire CD, Vance JM, Rasheed HK, Missaoui A, Rasheed KM, Maier FW. Using machine learning and feature selection for alfalfa yield prediction. AI. 2021;2:71-88.