Crop Yield Prediction Using Improved Random Forest

Padma T. [*], *and Dipali Sinha*

Dept. of Master of Computer Applications, Sona College of Technology, Tamil Nadu

Abstract. Agriculture has an important role in India's economic development. Crop productivity is affected by the rising population and the country's ever-changing climate. Crop yield estimation is a challenge in the farming sector. Numerous studies have been conducted in the agricultural sector to better estimate crop yield through machine learning techniques. It is an efficient method for anticipating crop yields and determining which crops to cultivate. Random Forest has been widely utilized for this purpose. A set of parameters in the Random Forest classifier must be stay tuned. The machine learning algorithm will yield better results with correct hyper parameter adjustment. This work presents a hybrid approach to agricultural yield estimation using a Random Forest classifier and the Random Search method with a 0.99 R2 score, 0.045 MSE, and 0.022 MAE, the suggested method outperformed other existing approaches such as Decision Tree (DT), Multiple Linear Regression (MLR), Random Forest (RF), and Grid Search (GS) optimized RF. Validation methods such as R2, Mean Squared Error, and Mean Absolute Error to cross-validation have been used to confirm the authenticity of the outcomes. The purpose of this study is to apply the crop yield prediction approach into action to assist farmers in solving agricultural production concerns.

1 Introduction

Agriculture is the engine that drives the Indian economy. Agricultural yield is essential in ensuring global food security since it influences the amount of food that can be produced for consumption. The demand for food increases in direct proportion to the growth in world population. Improving agricultural production is thus essential for catering to this demand, alleviating hunger, and ensuring a sustainable food supply. Emerging technologies such as Artificial Intelligence (AI) are reshaping the agriculture economy, particularly crop prediction. Farmers can get significant insights into elements that affect agricultural yields, such as weather patterns, soil moisture levels, and pest infestations, by utilizing AI to anticipate crop yields. The technology can analyze large datasets and find patterns to predict crop yields with a high degree of accuracy. This information allows farmers to optimize

[*] Corresponding author: padmat@sonatech.ac.in

their crop management practices, including irrigation, fertilization, and pest control, to increase agricultural yield.

Farmers can minimize crop production risks by predicting crop yield using machine learning (ML), a subset of AI. For instance, a farmer can alter their planting schedule or switch to a more suited crop type if they anticipate a low yield to reduce potential losses. Furthermore, applying ML to agriculture may result in sustainable farming methods. Farmers can use less water, fertilizer, and pesticides by using precise crop production projections, which can contribute to environmental protection. Agricultural forecasting, commonly referred to as crop yield prediction using ML, is the practice of forecasting crop yields using machine learning algorithms. This technology can assist farmers and agricultural specialists in making more educated decisions about crop planting, harvesting, and management.

Machine learning is a practical approach that can deliver more accurate yield prediction based on a variety of parameters. It is capable of learning from datasets by discovering correlations and patterns. Datasets that reflect the outcomes of prior experiments must be used to train the models [1]. Predictive models make predictions based on historical information. Various machine-learning techniques have been explored to improve agricultural yield forecasting [2]. Crop yield prediction using ML techniques typically relies on a variety of data inputs, including weather, soil, previous yield, and other environmental factors. Machine learning approaches for predicting crops analyze these data sources to detect patterns and forecast future crop yields.

This study primarily concentrates on the Random Forest which is applied to forecast the yield of crops, which aids farmers in choosing and cultivating the most lucrative crop, so lowering the likelihood of loss and raising productivity and the value of farming land. Hyper parameter tuning is a key component of machine-learning algorithms. It improves a machine learning model's effectiveness and is fine-tuned before the training process begins. Therefore, in this study, the Random Search algorithm has been used to optimize the Random Forest method's hyper parameters to improve the performance of a predictive model.

The following is the paper's outline: The relevant studies are included in the second subsection. The proposed model is discussed in the third subsection. The results of the experiment are shown in Section 4. Section 5 concludes by summarising the results and offering suggestions for further enhancements.

2 Related Work

For efficient crop management, sustainable agriculture, and food security, Kaur et al. [3] discuss the significance of crop yield forecasting. The authors emphasize the significance of various data sources used in agricultural production prediction models. These consist of satellite images, soil information, and weather data. They point out that prominent models employed in crop yield prediction include support vector regression, random forest, and neural networks Jhajharia, K. et al. [4] explored diverse ML algorithms for forecasting yields for various crops in Rajasthan, India. The results demonstrate that the random forest technique prevailed over the other predictive models. In this research, the most effective model, according to Iniyan, S. et al. [5] is the feature engineering-based LSTM, which provides better accuracy in comparison to other regression models. A comprehensive review of crop yield prediction using numerous deep learning models has been presented by Oikonomidis, A. et al. [6]. It became apparent that the convolutional neural network (CNN) performs the best when evaluated using the measure of Root Mean Square Error (RMSE). To predict agricultural yield, Balakrishnan, N. et al. [7] introduced an AdaSVM and AdaNaive ensemble approach. They concluded that the suggested model outperforms the conventional SVM and Nave Bayes models.

3 Methodology

3.1 Pre-processing

The pre-processing stage analyses the initial input data as a first step and gets the raw data ready for use in further processing [8]. The dataset has been rescaled using equation (1) to generate precise estimations.

$$X' = {}^{X - X_{min}}/_{X_{max} - X_{min}} \tag{1}$$

3.2 Random Forest

The popular method known as Random Forest makes use of the idea of ensemble learning [9], that is the process of incorporating multiple models to deal with a challenging task and improve the effectiveness of the model.

Algorithm 1: Random Forest

```
Step 1: From a given dataset, choose bootstrap samples randomly
Step 2: Make a decision tree for each set, then use the result to
make predictions
Step 3: Each predicted result shall be averaged
Step 4: As the final forecast, pick the outcome with the average value.
```

3.3 Random Search Optimization

A Random Search (RS) is an optimization approach used to identify the best hyper parameter combination for a machine learning algorithm. Random search is comparable to grid search, but instead of searching over a pre-defined grid of hyper parameters, it samples hyper parameters at random from a predefined search space. Random search is a computationally efficient algorithm for tuning hyper parameters in machine learning models

Algorithm 2: Random Search optimization

```
Step 1: Define a hyperparameter search space.  A distribution or a range of
values can be usedto define the search space
Step 2: Draw hyperparameters at random from the search space
Step 3: With the sampled hyperparameters, train and evaluate the model: On a
training set, train the model with the sampled hyperparameters. Using Mean
Square Error, assess the model's performance on a validation set
Step 4: Continue steps 2-3 for a set number of repetitions
Step 5: Depending on the evaluation metric, determine the hyperparameters that
produced thebest performance
Step 6: Train the model using the best hyperparameters on the entire dataset.
```

3.4 Proposed Model

The proposed approach incorporates a random forest model along with a random search optimization technique to enhance the precision of crop yield estimates. The Random Search algorithm has been employed for tweaking the hyper parameters of the Random Forest algorithm that cannot be learned during training. It will define a range for the machine-learning model's hyper parameters and generate multiple models using randomly selected combinations from the specified hyper parameters range. The average mean square error (MSE) is utilized as the fitness value to steer the search process. Finally, the model with the least error and the optimized set of hyper parameters will then be returned. The architecture of the proposed system is given in Figure 1. Following is the process for the suggested hybrid Random Forest classifier and Random Search approach for crop yield prediction.

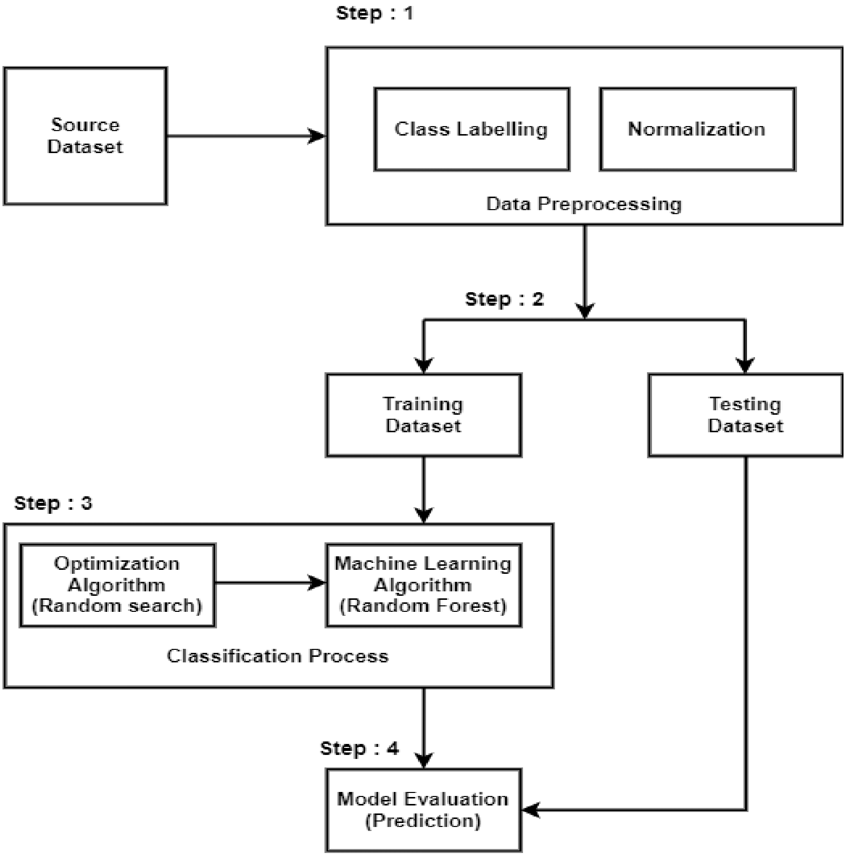

Fig. 1. Proposed Model

Algorithm 3: Crop yield prediction with hybridized Random Search optimized RandomForest Algorithm

```
Step 1: Take the dataset for predicting crop yields and divide it into a train set and a test set.
Step 2: Define a process with a random forest algorithm
Step 3: Make a grid with a maximum quantity of trees to be generated and the number of attributesthat must be chosen before splitting the node
Step 4: Create a function that executes the Random Search algorithm that takes input including aconfigured process, parametric grid, and train and test data
Step 5: Establish an objective function that is given a list of hyper parameters and outputs anMSE score
        $MSE = f(hyperparameters)$
Step 6: Set the number of iterations
Step 7: Traverse through all randomly selected combinations of hyper parameters provided in thegrid and evaluate the objective function
Step 8: Traverse through all randomly selected combinations of hyper parameters provided in thegrid and evaluate the objective function
Step 9: Determine the optimum hyper-parameter that minimizes the MSE.
```

4 Results and Discussions

The proposed random search optimized random forest is implemented and discussed in this section.

4.1 Dataset

This study utilized a crop yield prediction dataset for India from Kaggle. It includes characteristics such as temperature, precipitation, rainfall, pesticides, crop type, and field area. Because some machine learning models cannot deal with string values as input, label encoding is used to convert string data to numerical values.

4.2 Performance Measure

This study employed the metrics listed below in Table 1 to evaluate the effectiveness of the suggested system. The measures taken into account for assessing system performance include R2, Mean Square Error, and Mean Absolute Error.

Table 1. Performance Measures

Measure	Formula
R^2	$1 - \frac{SSR}{SST}$
MSE	$MSE = 1/n * \Sigma(i = 1\ to\ n)\ (yi - y\hat{}i)\hat{}2$
MAE	$MAE = 1/n * \Sigma(i = 1\ to\ n)\ \lvert yi - y\hat{}i \rvert$

4.3 Experimental Results

The standard models, including DT, MLR, RF, and GS Optimized RF, have been compared to the recommended prediction system.

Table 2. Performance Analysis of the Proposed System

Prediction Model	R -Square	MSE	MAE
DT	0.81	0.069	0.076
MLR	0.79	0.071	0.081
RF	0.91	0.078	0.082
GS optimized RF	0.95	0.063	0.065
Proposed Model	0.98	0.049	0.057

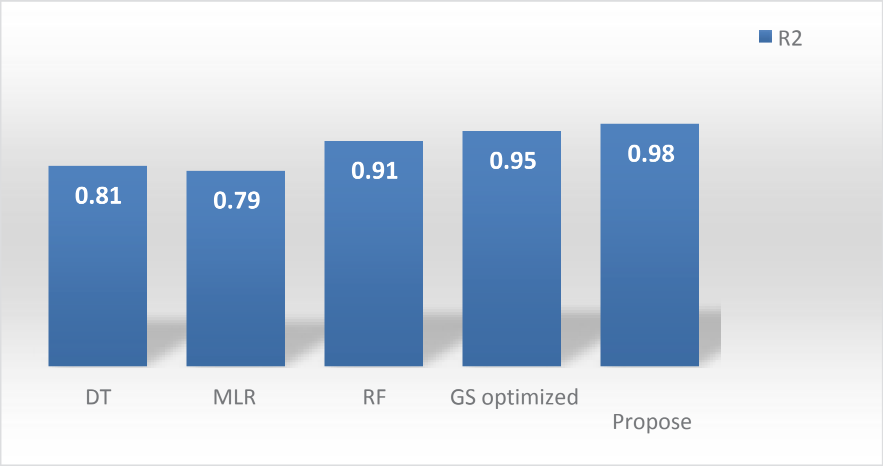

Fig. 2. Comparative Analysis of R^2

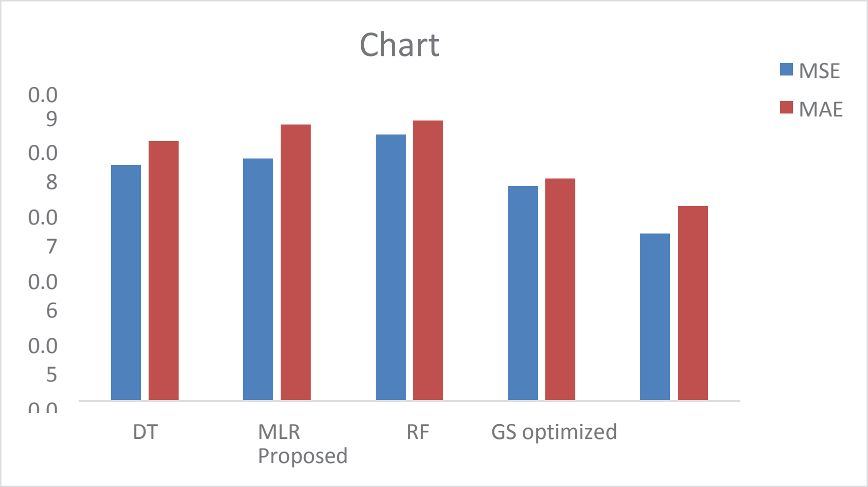

Fig. 3. Comparative Analysis of MSE and MAE

5 Conclusion

Crop forecasting analysis is crucial for the farmer to make decisions. Numerous studies have been conducted in the agricultural field to forecast crop production. In this investigation, the author employed various models of regression to forecast crop yield and observed that the random search optimized random forest technique performed better. According to results from experiments on crop yield data, the best outcome, with an R2 value of 0.98, is produced by an improved Random Forest. Despite the high forecast accuracy, this work still has certain shortcomings. Future research will examine more deep learning and machine learning models to create more accurate prediction systems.

References

1. D. S. Wigh, J. M. Goodman, & A. A. Lapkin, Comput. Mol. Sci **12**, 1603, (2022)
2. N. Bali, & A. Singla, Arch. Computat. Methods Eng. **29**, 95–112 (2022)
3. M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, & N. Khan, IEEE Access, **9**, 63406-63439 (2021).
4. K. Jhajharia, P. Mathur, S. Jain, & S. Nijhawan, Proc. Comput. Sci., **218**, 406-417 (2023)
5. S. Iniyan, V. A. Varma, & C. T. Naidu, Adv. Eng. Softw **175**, 103326 (2023)
6. A. Oikonomidis, C. Catal, & A. Kassahun, N. Z. J. Crop Hortic. Sci. 51 1- 26 (2023)
7. N. Balakrishnan, & G. Muthukumarasamy, Int. J. Adv. Res. Comput. Sci. Softw. Eng. 5, 148 (2016)
8. C. Fan, M. Chen, X. Wang, J. Wang, & B. Huang, Front. Energy Res., **9**, 652801 (2021)
9. T. G. Keerthan Kumar, C. A., Shubha, & S. A. Sushma, Int J. Innov Technol Explor Eng, 9(1), 1301-1304 (2019)
10. P. Liashchynskyi, & P. Liashchynskyi, arXiv preprint arXiv:1912.06059 (2019)