ISSN: 0067-2904

Machine Learning Based Crop Yield Prediction Model in Rajasthan Region of India

Kavita Jhajharia, Pratistha Mathur*
Information Technology, Manipal University Jaipur, Jaipur, India

Received: 24/2/2022 Accepted: 30/12/2022 Published: 30/1/2024

Abstract

The present study investigates the implementation of machine learning models on crop data to predict crop yield in Rajasthan state, India. The key objective of the study is to identify which machine learning model performs are better to provide the most accurate predictions. For this purpose, two machine learning models (decision tree and random forest regression) were implemented, and gradient boosting regression was used as an optimization algorithm. The result clarifies that using gradient boosting regression can reduce the yield prediction mean square error to 6%. Additionally, for the present data set, random forest regression performed better than other models. We reported the machine learning model's performance using Mean Squared Error, Mean Absolute Error and R-squared and identified that after the inclusion of gradient boosting regression, the accuracy increased to 92.77%. The MAE value decreased from 26.20 Mg/ha to 21.58 Mg/ha. The results indicate that machine learning models can improve the prediction of crop yield.

Keywords: Machine Learning, Crop yield Prediction, Decision Tree, Random Forest regression, Gradient boosting regression.

1. Introduction

Agriculture is an essential source of income in India [1] [2], as 59% of the country's population is employed in the agriculture domain, and 70% of the population is directly or indirectly reliant on agriculture for livelihood. The crop yield estimation is of most importance yet very difficult to predict due to factors such as peculiar climate, soil condition, water resources, and population growth. A gradual increase in population and global climate change are major concerns that require research to improve agriculture.

Agriculture is attaining new services, methods, and technologies to produce more food with the available inputs. Hence, crop yield prediction is a crucial step in agriculture. Present crop yield prediction methods involve crop simulation models, which are computerized descriptions of crop growth, continuous development, and crop yield estimation using mathematical equations on various variables such as soil, climate, seed quality, etc. [3] [4] [5]. The main objective of this study is to analyze data and use parameters such as soil, temperature, irrigation, fertilizer, and land use to develop innovative methods for crop yield prediction [6] [7] [8] [9].

In Rajasthan state, the total land is 342.7 lac ha with a net cropped area of 183 lac ha [10] (Rajasthan Agriculture Road map, 2016), and 22.5% of the state's economy is based on the agriculture sector. Figure 1 represents the location of Rajasthan State in India and the study's region. Rajasthan state is located in the northwestern side of the country and has the largest

*Email: kavita.jhajharia@jaipur.manipal.edu

geographical region 10.4% of the total area of India. As of 2020-2021, the state is the 9th largest contributor to India's overall GDP with $130 billion. Rajasthan state has a large area dedicated to crop production, but due to farm mechanization, which is important to increase production and quality, there a large gap in the experimental yield and the obtained value at the farm level. The population is growing in India, and with the rise in population, food demand is increasing. To feed the population, latest technologies and tools are required to be incorporated into the agriculture sector. Furthermore, timely advice to the farmers regarding the crop yield helps them plan the appropriate strategies to improve the produce.

With limited resources and environmental constraints, it is difficult for farmers to maintain crop productivity with good quality. The crop yield can be estimated using a machine learning algorithm to enhance the it without compromising the quality. A machine learning model process the factors that affect the crop yield and provide a more accurate prediction. However, past studies have focused more on having a suitable environment for agriculture [11]. In present circumstances, researchers focus primarily on analytical techniques, which provide limited information about the crops. The extracted data may not be enough to predict crop yield. Climate and weather changes have an impact on agricultural production. Suitable weather conditions lead to high production of the crop. High-quality seeds result in higher crop productivity (however, the issue with using high-quality seeds is that to predict the yield genotype and phenotype of the crop must be examined) [12]. Other factors such as water, a nutrient in the soil, and weeds can affect crop productivity [13].

The study by Ortiz-Bobea in [14] presents a model of weather effects on total productivity factors in agriculture at the global level, and the final model indicated that anthropogenic climate change has reduced total productivity factors by 21%. The research in [15] conducted a detailed study of climate, water, and crop yield models to identify the climate impact on the crops. The authors in [16] used meteorological data and introduced a weather forecasting model for crop yield in Europe. In addition, A study based on precision agriculture on a statistical model and incorporation of spatial dependence in the model for Canadian Prairies was introduced by Bornn and Zidek [17].

For crop growth and crop yield, a suitable condition was discussed on AVHRR for Poland [18]. To yield the maize crop, researchers considered remote sensing data of leaf area index and soil moisture and proposed a model which used sequential data integration [19]. Since the effect of extreme weather conditions of seasons on the Mediterranean crop is an issue, it should be included in the crop models for better predictions [20].

The researchers in [21] applied an ensemble Kalman filter to integrate the soil moisture estimation to reduce the errors encountered due to ambiguity in the temporal rainfall distribution-based crop model. Chlorophyll content present in the leaf area index also contributes to crop yield prediction [22]. To predict the crop, a study was conducted using four vegetation indices SAVI, PVI, NDVI, and GVI and a neural network-based crop yield prediction model [23].

Precision agriculture utilizes tools and technologies to fulfil the need for soil and crops for the best productivity. In precision agriculture, real-time data on farms and weather are gathered using sensors, deployed on the farm, and predictions of the crop yield are made to assist farmers in making the right farming decision [24]. The data gathered by sensors are enormous in size and therefore can be processed using big data analytics. The outcome of such data can provide benefits to farmers as well as to the nation's economic development [25].

Big data analytics and machine learning algorithms can increase crop yield by many folds. To implement the machine learning algorithms, the present study used guar seed (cluster seed), groundnut, bajra (pearl millet), moong (green gram), gram, rapeseed & mustard and wheat, and implemented decision tree, random forest and gradient boosting regression. The objective of the current study is 1) To implement machine learning techniques to predict the crop yield for the upcoming years and 2) Validate the results using MAE, MSE, and R2 validation metrics. Most past studies were concentrated on image processing techniques and statistical models for prediction. The proposed work uses machine learning, which will increase the computation and prediction efficiency compared to statistical models.

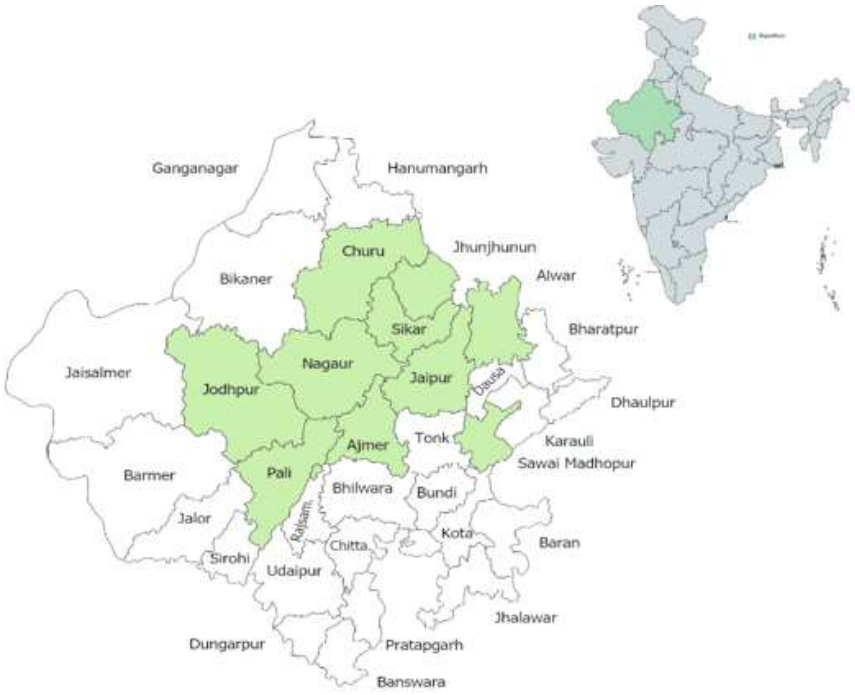


Figure 1: Rajasthan state location in India

The article is organized as follows: Section 2.1 describes data sources and data pre-processing, section 2.2 illustrates methods implemented in the study and validation matrices. Section 3 discusses the achieved results including a comparison of the model's performances, and finally, section 4 provides the conclusion of the study.

2. Materials and Methods

2.1 Data Acquisition

The data for this study was acquired from the from the agricultural department of the government of Rajasthan for the years 1997 to 2018 from ten major crop-producing districts, namely Ajmer, Alwar, Churu, Jaipur, Jodhpur, Nagaur, Pali, Sawai Madhopur, and Sikar, of state Rajasthan, (Figure 1). The final data focus on seven crops including, gaur seed (cluster seed), groundnut, bajra (pearl millet), moong (green gram), gram, rapeseed & mustard, and wheat along with the area (hectare), Production (Tonnes), yield (Tonnes/Hectare) and rainfall in the past 21 years. The data was pre-processed before applying the machine learning algorithms.

2.2 Methods

Initially, the data was obtained from multiple government agencies, and the raw data was pre-processed to eliminate irrelevant and unnecessary data. This step also included the conversion of categorical data to numeric data. Further, the missing values were identified and filled with the appropriate mean values required. The data was divided into features and labels that were further divided into training and testing datasets. Figure 2 depicts the framework of the study.

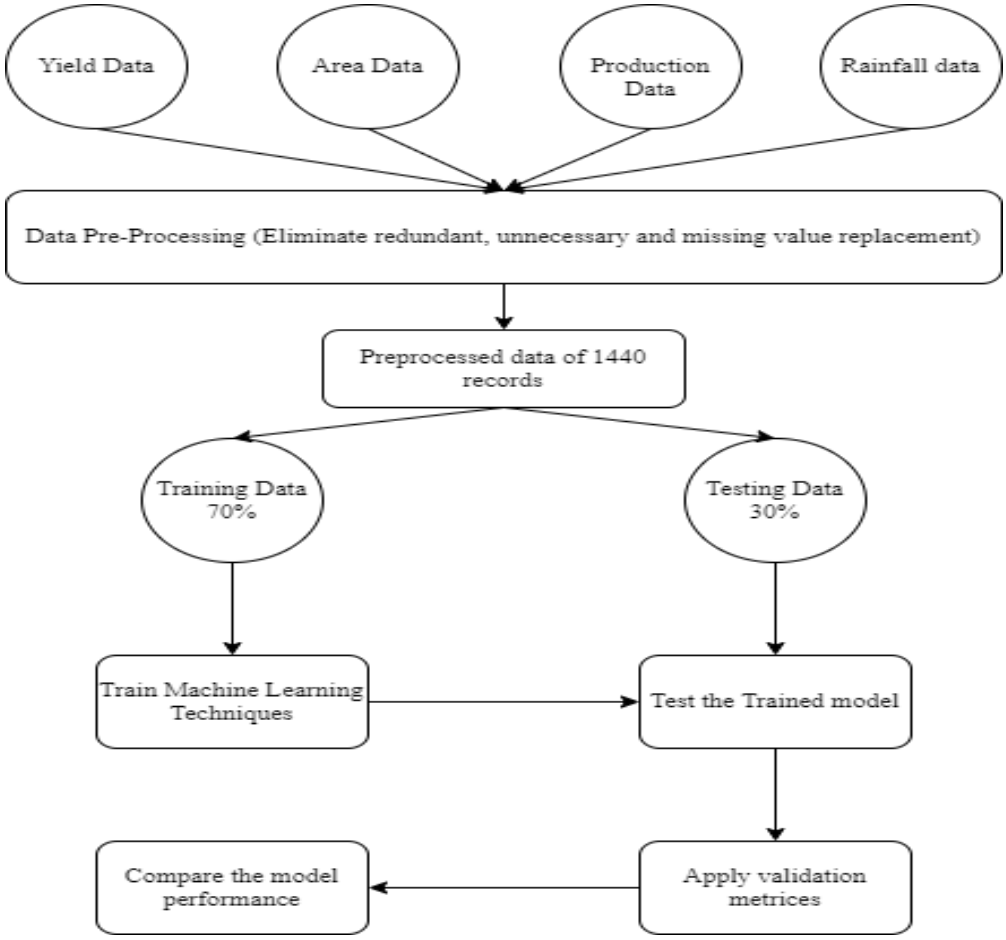


Figure 2: Framework of the study

2.2.1 Decision Tree

The decision tree Regressor algorithm is in the sub-domain of supervised machine learning that can also be used for regression/value prediction analysis and classification tasks. For the current study, the objective is to predict the value of the target variable i.e., crop production by the directives that train the model by providing training data. It uses a tree to represent the model or possibly solve the problem. In decision trees, attribute selection is the most important task, which can be done in two mechanisms, Information gain and Gini impurity. It uses various parameters of country/state such as area, the area under irrigation, crop year, and crop season (Kharif, Rabi or Whole Year).

2.2.2 Random Forest Regressor

Random Forest algorithm is also a supervised machine learning algorithm. In an unplanned manner, it creates a forest with many trees. Generally, in Random Forest Regressor, if the number of the trees is high then the accuracy will be high. Random forests handle the stumbles created by missing values and do not overfits the model when we have a higher number of trees present in the forest. Mostly, there exist two stages in this algorithm, the first one being the

creation of a random forest, and the other one is to pull out predictions from the regressor created in stage one. It randomly selects a few rows of the dataset to make a stump tree and tries to identify the maximum number of trees for a condition to decide the prediction. The model used η_estimator value ten and random states 101.

2.2.3 Gradient Boosting Regression

Gradient boosting combines weak prediction models and generates ensemble models. Gradient boosting algorithm is used for regression and classification and fits models that predict the continuous values. It uses multiple fixed-size decision trees, selected by the η_estimator parameter, to build an additive model. The model fitting process is initialized by a constant value, which can be the mean value of the target, and in the following stages, the negative gradients are predicted to fit the estimator. The model used η_estimator as 100, random_state as 42 and max_depth as 4. The learning rate is used to add the new trees sequentially to reduce residue errors in the predictions.

1.4 Tables and Figures

2.3.1 Analysis of production over the years for multiple categories

Figure 3 represents the production of crops (Tonnes) and area (Unit Hectare) for Rajasthan from 1997 to 2018. In Rajasthan, the maximum crop production is 941557 tonnes on 709268 Hectares whereas the average production is 102605.7 tonnes over the 97763.05 hectares. Figure 4 illustrates the analysis of production for different sample crops, including Bajra and wheat. Bajra sometimes has given unexpected production despite the large area allotted but wheat has always outgrown expectations. Figure 5 analyses the mean production of the crops for the selected ten major districts. The illustration indicates that the production of bajra and wheat is high in almost every district. Figure 6 shows the production for the same region (Ajmer) from 2017 to 2021. Changing patterns motivate us to find reasons behind these changes, and factors that affect crop production are not always independent.

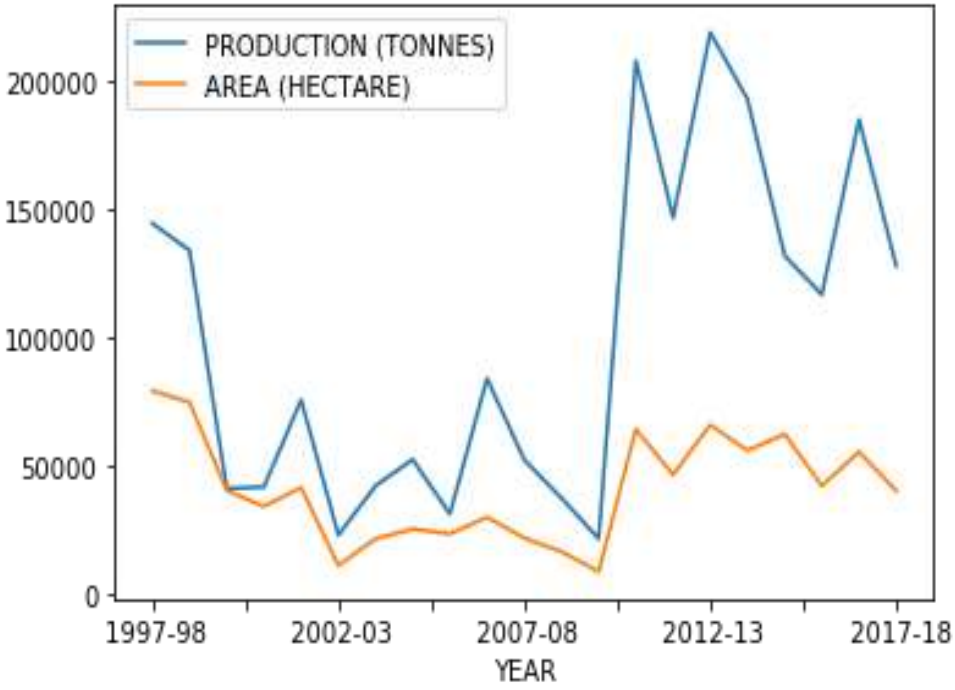
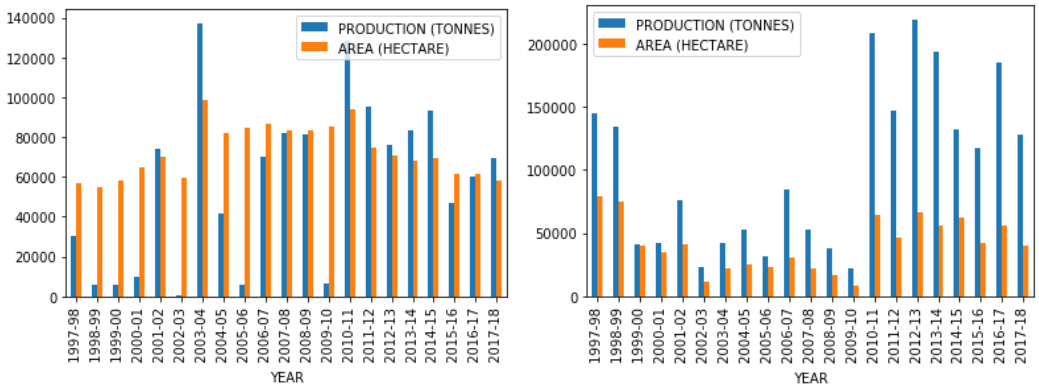


Figure 3: Analysis of crop production and area over the years



(a) Bajra (b) Wheat

Figure 4: Analysis of production and area for different crops over the years

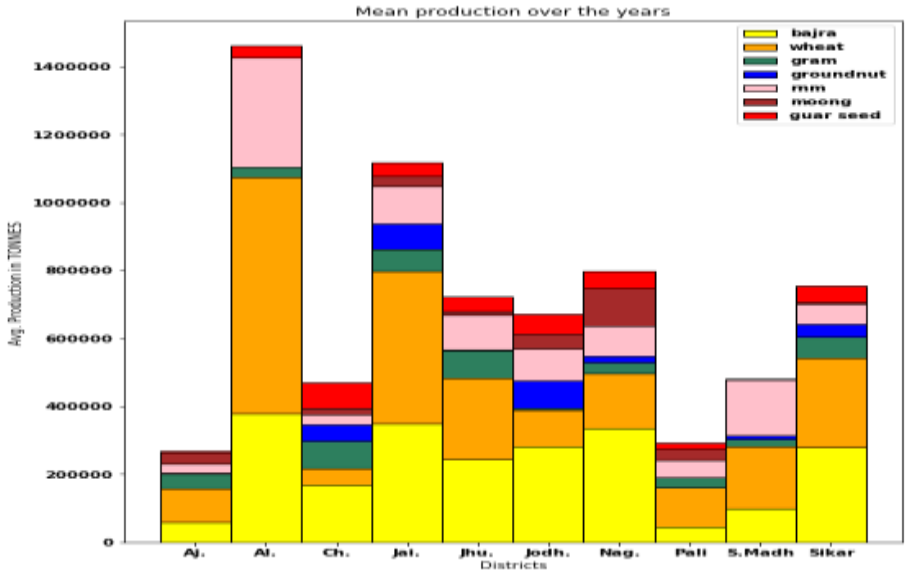


Figure 5: Crop-wise mean production over the years for the selected districts

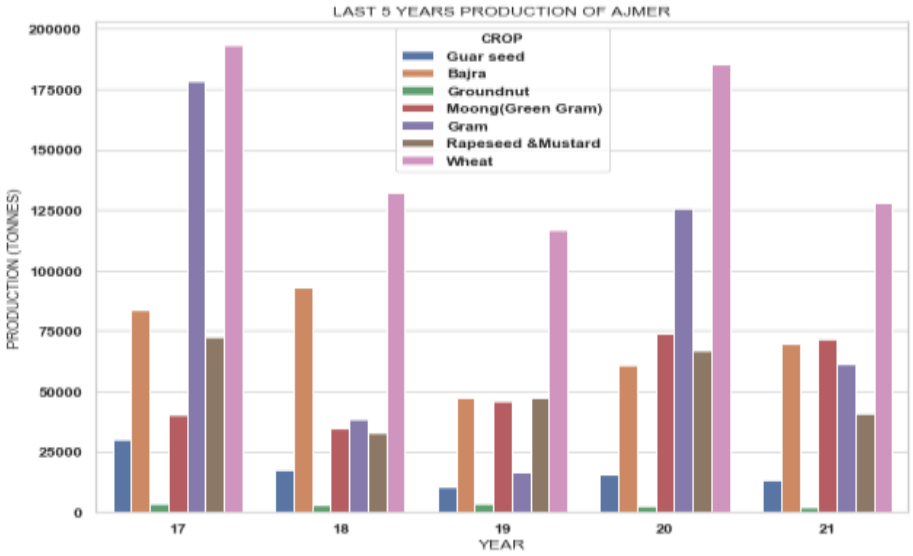


Figure 6 : Analysis of the last five years production for selected crops in Ajmer District

Figure 7 shows the district-wise rainfall in mm units. The maximum rainfall was 296310.50 mm and the average rainfall was 7267.52 mm, the distribution of the rainfall is uneven.

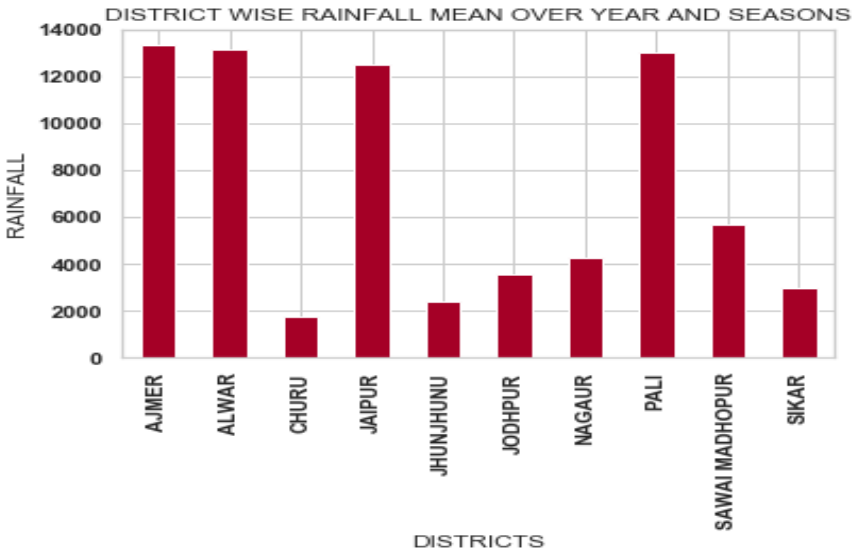


Figure 7: Analysis of rainfall in selected districts

3. Validation Metrix

Model accuracy is assessed with the help of validation metrics, such as mean absolute error (MAE), mean squared error (MSE), and Correlation coefficient (R2). They are demarcated in equation 1, 2 and 3 below:

$$\text{MAE} = \frac{1}{k} \sum_{j=1}^{k} |X_j^p - X_j| \tag{1}$$

$$\text{MSE} = \frac{1}{k} \sum_{j=1}^{k} (X_j - X_j^P)^2 \tag{2}$$

$$\text{R2} = 1 - \frac{\text{Sum of Squares of Residues}}{\text{Total sum of squares}} \tag{3}$$

MAE and MSE are used to measure the difference between the predicted value and the actual value. MAE represents the variation between the actual value and predicted value obtained by the average absolute difference over the data set. MSE signifies the distinction between the original and predicted value obtained by squaring the average difference over the data set. The coefficient of determination (R2) signifies the coefficient of values that fit in compared to the original value, the higher the value the better the model.

4. Results and Discussion

The current study implemented three methodologies, decision tree, random forest and gradient boosting regression for the crop yield prediction using the anaconda platform. The result of all three was compared with linear regression, lasso regression, and ridge regression. Table 1 summarizes the models' comparison using mean squared error and accuracy score.

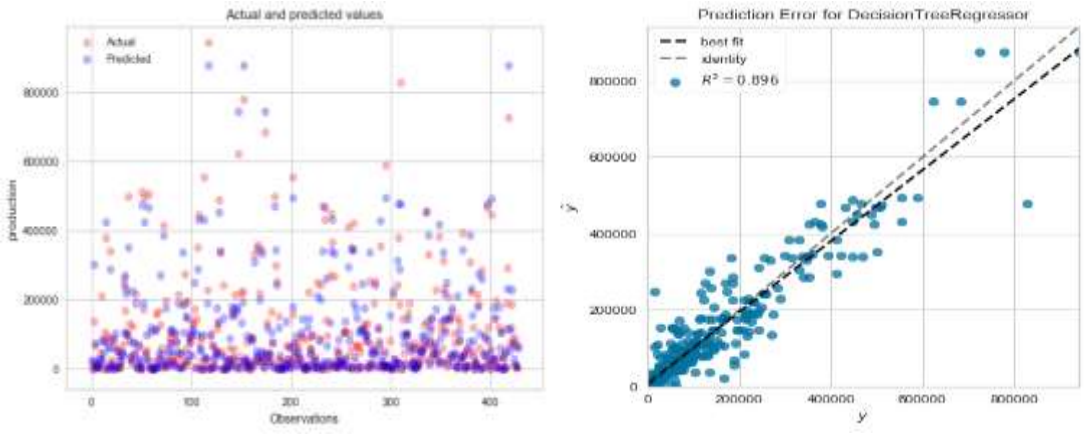
Table 1: Accuracy score and mean squared error of Rajasthan's Agricultural Data

Model	Accuracy score	Mean Squared Error
Decision Tree Regressor	0.9008	25331.7032
Gradient Boosting Regression	0.9277	21581.3307
Lasso Regression	0.7037	56344.8707
Linear Regression	0.7043	56347.2529
Random Forest	0.9271	22334.6126
Ridge Regression	0.7043	56195.1971

The above table shows that gradient-boosting regression outperforms all the other methodologies with 92.7% accuracy. The regression techniques are not able to perform better than decision tree and random forests, thus we have selected the best-performing models for prediction. As required for the machine learning model implementation, the data was divided into test and train data with a ratio of 3:7, which indicates that 30% of the entire data was used for testing the model and 70% of the data was used for training. Hence, all the models were trained on the data from the year 1997 to 2018, and the crop yield estimation was done.

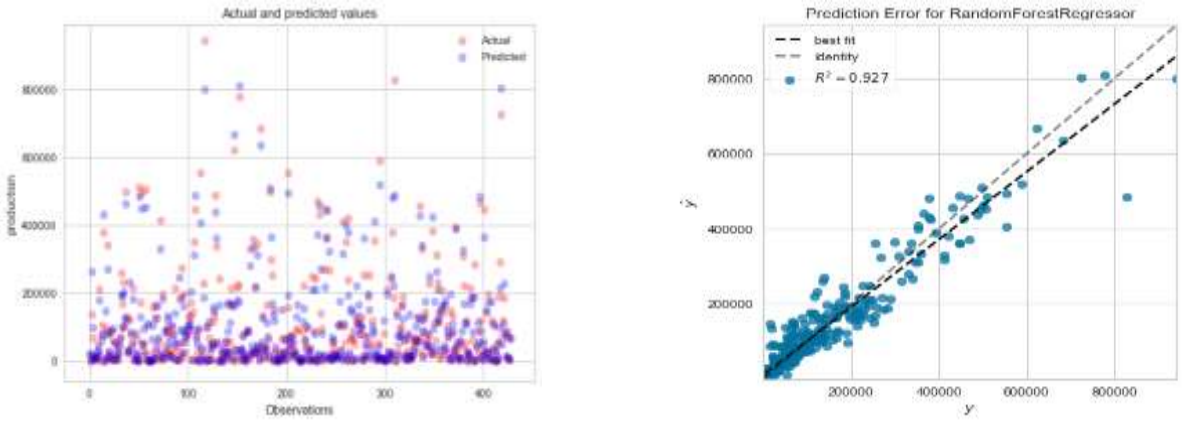
The decision tree provided considerable results for the present research with 89.64% accuracy, MAE is 26.20, and MSE is 21.53. Decision trees make the prediction more intuitive to understand, highlighting how each of the factors affects it. Figure 8 (a) shows the results of the decision tree with the actual value and the predicted value, which indicates that outliers are less, and (b) represents the R^2 value of 0.896 for error prediction.

Random Forest performed better compared to the decision tree with 92.71% accuracy and MAE is 22.33, and MSE is 15.13. Figure 9 (a) represents the results of the random forest with the actual value and predicted value, and (b) shows the representation for R^2 with a value of 0.92. For the current study, the value for η_estimator was 10 and the total random states were 101. Gradient Boosting Regression outperforms all the other models and provides 92.77 % accuracy with an MAE value of 21.58, and an MSE value of 15.01. The model used η_estimator as 100, random_state as 42 and max_depth as 4. Figure 10 (a) represents the comparison of actual and predicted values, and (b) represents the R^2 value with 0.928. Table 2 contains the consolidated accuracy, MAE, MSE, and R^2 values for all three approaches.



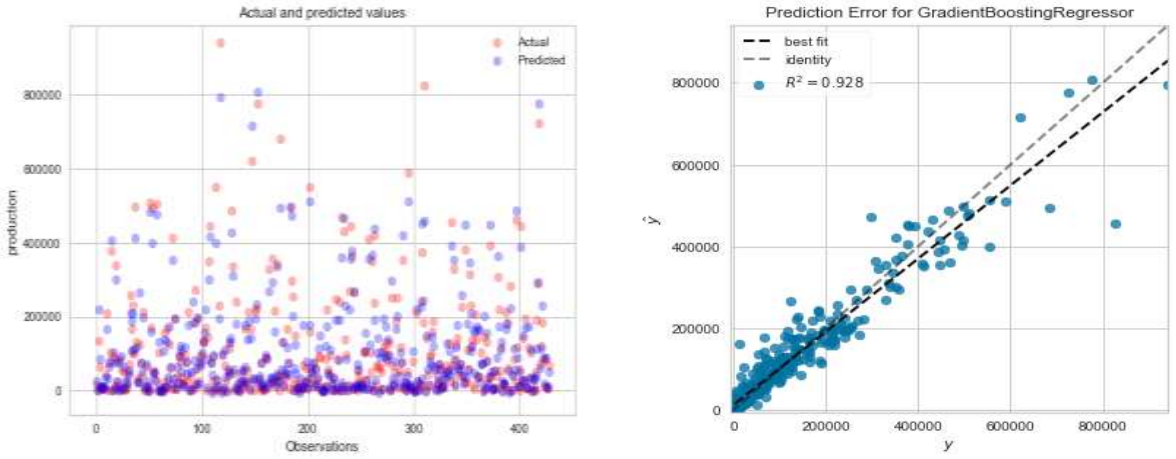
(a) (b)

Figure 8: (a) The results of the decision tree with the actual value and predicted value, and (b) shows the R2 value



(a) (b)

Figure 9: (a) The results of the random forest with the actual value and predicted value, and (b) shows the representation for R2 with the value



(a) (b)

Figure 10: (a) Comparison of the actual and predicted value for Gradient Boosting Regression, and (b) the R2 value representation

Table 2: Performance of Random Forest, Decision Tree, and Gradient Boosting Regression

Model	Accuracy	MAE	MSE	R^2
Decision Tree	89.64	26.20	21.53	89.64
Random Forest	92.71	22.33	15.13	92.71
Gradient Boosting Regression	92.77	21.58	15.01	92.77

In this study, gradient boosting regression attained a 92.77% accuracy score to predict the crop for the present dataset. From the results of the study, we identified that some crops have reduced production but there is still an increase in the price of the crop. This shows that the production for the crop has been decreasing but the demand for it has not, and this can be observed from the positive slope. The results also reveal that wheat and Bajra are the most produced crops in the selected ten districts of Rajasthan, but many other crops can be focused on to get more benefit. The increase in the production of such crops is less compared to their demand and these crops will be more profitable to produce. Table 3 shows the crop list along with the price variance.

Table 3 : Crops with a slow increase in production but a high increase in prices

Crop	Production variance	Price variance
Arhar/Tur	5237.572915	260.551948
Groundnut	1758.235163	199.339827
Jowar	3648.323679	234.95671
Jute	-43011.14643	125.248918
Moong	941.797058	308.993506
Sunflower	-4511.467222	223.906926
Sesamum	1494.291172	279.404762
Urad	2670.533797	284.469697
Safflower	-1031.251935	122.445887
Niger seed	-144.009443	249.534632

4. Conclusion

The crop yield depends on several factors and the research on the domain is immensely useful for farmers. The study was conducted with the objective to identify the most efficient machine learning techniques for crop yield prediction in Rajasthan state. The region of the study was ten selected districts of Rajasthan state based on the data from 1997 to 2018.

Among all the applied machine learning algorithms ridge regression, lasso regression and linear regression, could not produce good results but decision trees and random forests gave considerable results. Gradient boosting regressor produced the best results for the present dataset. The predicted results acquired from the various techniques were evaluated by validation metrics. The R2 for gradient booster was highest with 92.77 values and lowest for the decision tree with 89.64 values. The study emphasizes the benefits of machine learning algorithms in crop yield prediction.

In the future, the study can be extended to other regions of the country. Finding crops that have special changing patterns over the years, such as the decrease in production, can help understand the reasons behind that in a more specific way. The results can be used to help farmers decide on their crops for more monetary gain and with little risk. Additionally, the government can be better prepared for anomalies with better resource arrangements such as insurance, logistics and resources.

6. Disclosure and conflict of interest

The authors declare that they have no conflicts of interest.

References

[1] A. K. Kushwaha and M. Tech, "Crop yield prediction using Agro Algorithm in Hadoop," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, p. 5, 2015.

[2] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Jan. 2016, pp. 1–4.

[3] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods," *Agric. For. Meteorol.*, vol. 218–219, pp. 74–84, Mar. 2016.

[4] J. N. Brown, Z. Hochman, D. Holzworth, and H. Horan, "Seasonal climate forecasts provide more definitive and accurate crop yield predictions," *Agric. For. Meteorol.*, vol. 260–261, pp. 247–254, Oct. 2018.

[5] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Comput. Electron. Agric.*, vol. 151, pp. 61–69, Aug. 2018.

[6] P. K. Aggarwal, "Uncertainties in crop, soil and weather inputs used in growth models: Implications for simulated outputs and their applications," *Agric. Syst.*, vol. 48, no. 3, pp. 361–384, Jan. 1995.

[7] P. Grassini *et al.*, "How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis," *Field Crops Res.*, vol. 177, pp. 49–63, Jun. 2015.

[8] M. Safa, S. Samarasinghe, and M. Nejat, "Prediction of Wheat Production Using Artificial Neural Networks and Investigating Indirect Factors Affecting It: Case Study in Canterbury Province, New Zealand," J. Agr. Sci. Tech., Vol. 17, pp. 791-803. 2015.

[9] F. F. Bocca and L. H. A. Rodrigues, "The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling," *Comput. Electron. Agric.*, vol. 128, pp. 67–76, Oct. 2016.

[10] "| NITI Aayog." https://www.niti.gov.in/ (accessed Aug. 27, 2021).

[11] A. Morshed, R. Dutta, and J. Aryal, *Recommending Environmental Knowledge As Linked Open Data Cloud Using Semantic Machine Learning*. 2013.

[12] B. Parent and F. Tardieu, "Can current crop models be used in the phenotyping era for predicting the genetic variability of yield of plants subjected to drought or high temperature?," *J. Exp. Bot.*, vol. 65, no. 21, pp. 6179–6189, Nov. 2014.

[13] J. Han *et al.*, "Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China," *Remote Sens.*, vol. 12, no. 2, Art. no. 2, Jan. 2020.

[14] A. Ortiz-Bobea, T. R. Ault, C. M. Carrillo, R. G. Chambers, and D. B. Lobell, "Anthropogenic climate change has slowed global agricultural productivity growth," *Nat. Clim. Change*, vol. 11, no. 4, pp. 306–312, Apr. 2021.

[15] Y. Kang, S. Khan, and X. Ma, "Climate change impacts on crop yield, crop water productivity and food security – A review," *Prog. Nat. Sci.*, vol. 19, no. 12, pp. 1665–1674, Dec. 2009.

[16] P. Cantelaube and J.-M. Terres, "Seasonal weather forecasts for crop yield modelling in Europe," *Tellus Dyn. Meteorol. Oceanogr.*, vol. 57, no. 3, pp. 476–487, Jan. 2005.

[17] L. Bornn and J. Zidek, "Efficient stabilization of crop yield prediction in the Canadian Prairies," *Agric. For. Meteorol.*, vol. 152, pp. 223–232, Jan. 2012.

[18] K. Dabrowska-Zielinska, F. Kogan, A. Ciolkosz, M. Gruszczynska, and W. Kowalik, "Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices," 2002.

[19] A. V. M. Ines, N. N. Das, J. W. Hansen, and E. G. Njoku, "Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction," *Remote Sens. Environ.*, vol. 138, pp. 149–164, Nov. 2013.

[20] M. Moriondo, C. Giannakopoulos, and M. Bindi, "Climate change impact assessment: The role of climate extremes in crop yield simulation," *Clim. Change*, vol. 104, pp. 679–701, Feb. 2011.

[21] A. J. W. de Wit and C. A. van Diepen, "Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts," *Agric. For. Meteorol.*, vol. 146, no. 1, pp. 38–56, Sep. 2007.

[22] D. Haboudane, J. R. Miller, N. Tremblay, P. J. Zarco-Tejada, and L. Dextraze, "Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture," Remote Sens. Environ., vol. 81, no. 2–3, pp. 416–426, Aug. 2002.

[23] S. S. Panda, D. P. Ames, and S. Panigrahi, "Application of vegetation indices for agricultural crop yield prediction using neural network techniques.," Remote Sens., vol. 2, no. 3, pp. 673–696, 2010.

[24] A. McBratney, B. Whelan, T. Ancev, and J. Bouma, "Future Directions of Precision Agriculture," *Precis. Agric.*, vol. 6, no. 1, pp. 7–23, Feb. 2005.

[25] D. Howe *et al.*, "The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, Sep. 2008.