2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)
Galgotias University, Greater Noida, UP, India. Oct 30-31, 2020

Crop Yield Estimation in India Using Machine Learning

Ms Kavita
Information Technology
Manipal University Jaipur
Jaipur, India
Kavita.chaudhary@outlook.com

Pratistha Mathur
Information Technology
Manipal University Jaipur
Jaipur, India
Pratistha.mathur@jaipur.manipal.edu

***Abstract*— Agriculture in India is significant economic support. The population growth is the major challenge for food security. The population growth makes a rise in demand which requires farmers to produce more from the same agriculture land in respect to increase the supply. Technology can help farmers to produce more with the help of crop yield prediction. The main aim of this paper is to predict crop yield using area, yield, production, and area under irrigation. Four machine learning techniques Decision Tree, Linear Regression, Lasso regression, and Ridge Regression have been applied to estimate the crop yield. Cross validations methods, for validation, mean absolute error, mean squared error, and root mean squared error, were used to validate. The Decision tree outperforms other machine learning techniques.**

Keywords— crop yield, decision tree, LASSO regression, Linear regression, machine learning.

I. INTRODUCTION

In India, agriculture is the primary source of food for the large population as well as significant economic support. The rapid increase in the population of India and vital climate variations food supply and demand chain is required to be maintained. Many scientific techniques have been immerged with agriculture to maintain the balance between demand and supply of food. The notable variation in the climate creates a problematic situation for Farmers to decide how to be more dynamic and sustainable [2]. Agriculture requires more production from fewer inputs with the support of new technology, new farming methods, and time-saving products [3]. Thus, to determine food security problems, Crop yields Estimation plays a significant role [4].

Crop yield estimation can be used to help farmers to reduce the loss of production under unsuitable conditions and increase the production under suitable and favourable condition. Crop Yield positive prediction is affected by many factors, including farmer's practices, decisions, pesticides, fertilizers, weather conditions, and market values [5]. Crop yield estimation can be done using statistical data of previous year's yields along with the Rainfall, weather, and area-wise production [6].

Machine learning has recently evolved in many fields, including Agricultural domain. Many Machine learning techniques have been applied to predict the crop yield including support vector machine [2] [7], Decision tree [8], Artificial Neural Network [9] [6] [10], and Deep Learning [11] [12]. This research estimates the crop yield for India using data from 1950 to 2018. The prediction is made for five crops which are Rice, Wheat, Jowar, Bajra, Tobacco, and Maize using parameters including the area used for the crop sowing, production, Yield, and Area under irrigation. The prediction is attained using decision tree and random forest.

II. RELATED WORK

Crop yield estimation can be attained by implementing machine learning techniques. Multiple Linear Regression model was used to predict the crop yield using the Dataset which includes total cultivation area, water resources for irrigation (tanks and wells), canals length, and average maximum temperature [3]. Another study claimed that the computational model developed was producing better results than Lasso, Shallow neural network, regression tree, Deep Neural network approach was used to design the model. Root mean square error is 12% of the average yield and 50% of the standard deviation for the dataset validation using predicted weather data [11]. [13] Conducted research on four objectives as listed, First, study the artificial neural network model to predict the corn and soybean yields under unfavourable climatic conditions, Second, check the estimation capabilities of the model at state, regional and local levels, third, Artificial neural network performance is evaluated with reference to parameter variation, and fourth, compare the developed Artificial neural network model to other multiple linear regression models. [10] The study intended to use artificial neural networks to estimate rice production in various district of Maharashtra, India. The data was collected for 27 districts of Maharashtra from the Indian Government's public records. The obtained accuracy was 97.5% with the observed parameters were rainfall, minimum, maximum and average temperature, area, production, yield, and reference crop evapotranspiration for the years 1998 to 2002, Kharif season [6]. The research focuses on crop yield estimation of Kharif crops of Andhra Pradesh's district Vishakhapatnam. Rainfall plays a considerable role in Kharif crop production, so the authors first predicted Rainfall using modular artificial neural networks and further predicted crop yield by using the Rainfall and area data using support vector regression. These two methodologies were applied to increase the crop yield. The

978-1-7281-6324-6/20/$31.00 ©2020 IEEE

research work uses Machine Learning algorithms named, Artificial Neural network, Support Vector Regression, K-nearest Neighbor and Random Forest to estimated crop yield with better accuracy. The data used in research is consist of 745 instances where 70% of data are randomly assigned for training the model, and the remaining 30% is used for testing and evaluating the final model performance. The final result indicates that the Random Forest algorithm gains the highest accuracy [14]. [15] This Research proposes a novel model to predict the yield of soybean using Long Short Term Memory in southern Brazil, and Neural Network on satellite and weather data. The primary goal of this research is to, i) conducts a comparison study among multivariate OLS linear regression, random forest and LSTM neural networks on the basis on their performance. The forecasting is done on soybean data using Vegetation Indices, Land surface temperature and Rainfall as independent variables and ii) Estimate how early this model can predict the yield with reasonable accuracy. Among all the Algorithms Long Short Term Memory performers better for all the forecasts except DOY 16, For DOY 16 multivariate OLS linear Regression performs better. [2] The paper analyses the results attained by implementing Sequential minimal optimization classifier. WEKA tool was used to perform the experiment on the data of 27 districts data of Maharashtra, India. The results obtained from the experiment on the same Dataset indicated that other techniques performer better than Sequential minimal optimization. The validation was done using Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, and Root Relative Squared Error. In respect to accuracy and quality, BayesNet and Multilayer Perceptron has shown the highest accuracy and better quality, whereas Sequential minimal optimization has indicated the lowest accuracy and poor quality.

III. Material and Methods

A. Study Region

In this study, authors focus on India due to its climate variation from humid to dry in the southern and temperate alpine in northern India. India has a land area of 297319 ha and agriculture area of 179721 ha. Six major crops grown in India are rice, jowar, maize, bajra, tobacco, and wheat have been selected to perform this study. India is the top exporter of Rice in the world with the export quality of 12,060,844 tones [16].

B. Data sources

The Dataset used for the experiment in this research is originally collected from www.mospi.gov.in and https://data.gov.in, which is made public by government authorities. The obtained dataset has the following features: rainfall, area, area under irrigation, crop names, seasons, production, and yield for the year 1950 to 2018. Figure 1 represents the crop's production in the last 68 years.

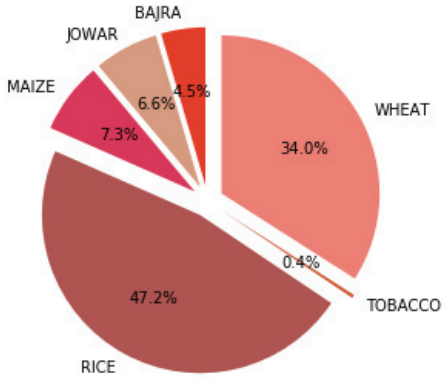


Fig. 1. Crop production mean for 1950-2018

This study applies Decision Tree, Linear Regression, Lasso regression, and Ridge Regression methodologies to predict the crop yield for India. And for validation, mean absolute error [1], mean squared error [17] and root mean squared error [18].

C. Methods

1) Decision Tree

Decision tree is a non-parametric machine learning algorithm used for regression and classification problems. The decision tree algorithm fulfils two major tasks, first, classifying the features which are appropriate for every decision, and second, concluding which choice to make based on chosen features. Decision Tree algorithm assigns a probability distribution to the plausible choice [19]. In decision tree, every node symbolizes a feature, and every branch leads to a decision, and the leaf node indicates the final result. For the construction of a decision tree, one feature should be selected as the root node to start the tree production, and further to complete the tree splitting of data is required.

2) Linear Regression

Linear regression is machine learning as well as the statistical algorithm. The main aim of linear regression is to generate mathematical models to describe the relationship between two variable. The model assumes a statistical relationship between input / dependent variable (x) and output / independent variable (y). The independent or response variable is calculated from the dependent variable. Another aim of regression is to examine the hypothesis by prediction and mathematical explanation [20].

3) Lasso regression

Lasso regression is least Absolute Shrinkage Selection Operator. In lasso regression value of the parameter controls both size and number of the coefficients, with higher values of leading to the greater number of covariates to be included in the linear model. In other words, the model shrinks some coefficients and sets others to 0, and hence tries to retain the useful features [21].

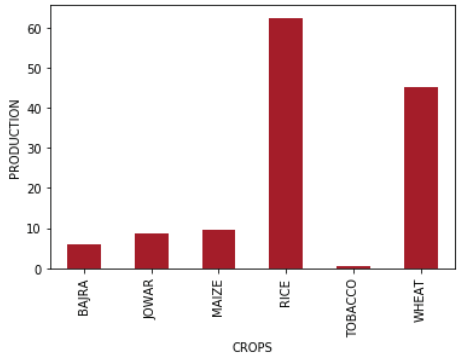
4) Ridge Regression

Ridge regression is applied to advantage when the predictor variables are highly collinear. This method is used to analyse multicollinearity in multiple regression data. This is most

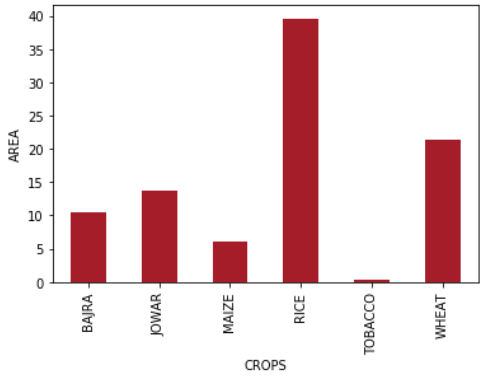
suitable when a dataset contains a higher number of predictor variables than the number of observation [22].

IV. Results

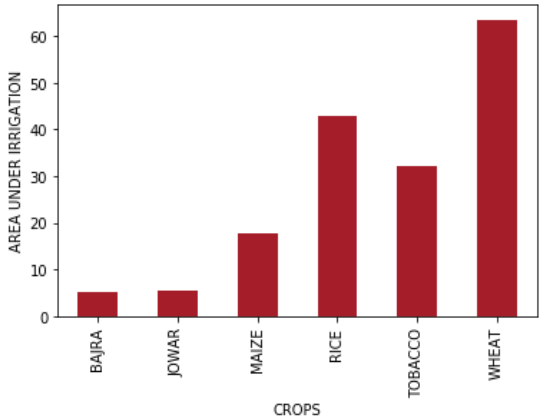
The Dataset features have a relationship among them, authors considered crop as a major feature and plotted the relationship represented in figure 2. The representation indicates the production and the area allotted to rice is highest. This satisfies the fact that India is the top exporter of rice to the world. Area under cultivation of rice and wheat took the maximum proportions and accounted for 75% of the food grains production in the country.



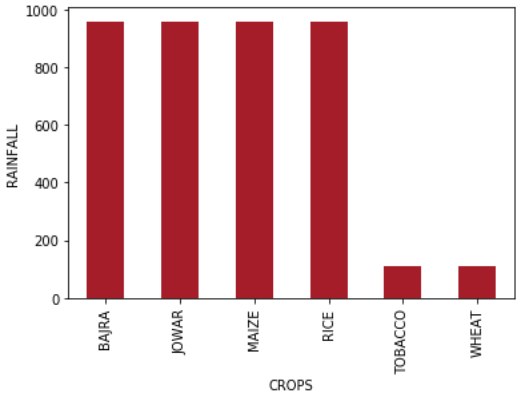
(a)



(b)



(c)



(d)

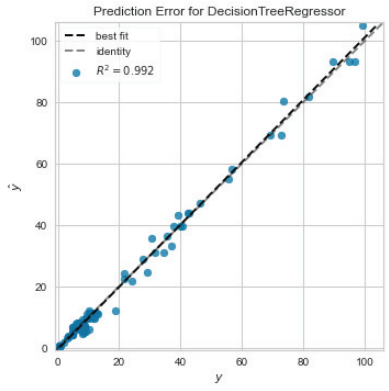
Fig. 2. Relationship between crop and features, (a) Represents relationship between crop and production (b) Represents Area allotted to crops (c) shows the area under irrigation and crop relationship (d) rainfall and crop mapping over the years

The prediction results are shown in Table 1. The research directs to the conclusion that decision tree produces better accuracy when compared to other machine learning algorithms.

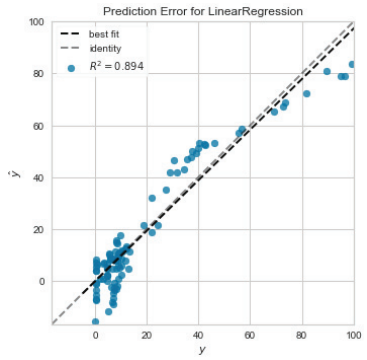
TABLE I Models comparison among decision tree, linear regression, lasso regression and ridge regression.

	Accuracy	Errors	
Models		***Mean Absolute Error (MAE)***	***Root Mean Squared Error (RMSE)***
Decision tree	98.62	1.45	2.11
Linear Regression	89.38	5.42	6.27
Lasso Regression	86.33	6.25	8.85
Ridge Regression	89.53	5.49	6.53

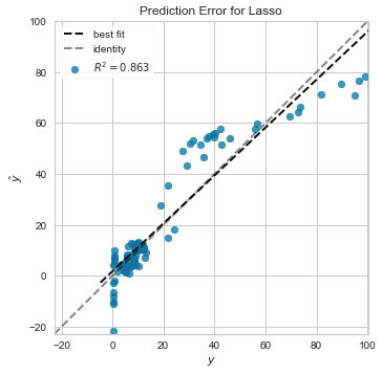
The decision tree illustrates the performance at country level with MAE = 1.45, and RMSE = 2.11 (table 1). The scatter plot of predicted errors of Decision tree, linear regression, lasso and ridge regression is demonstrated in figure 3.



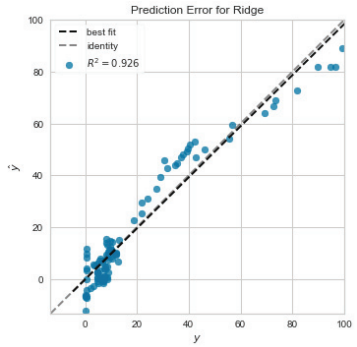
(a)



(b)



(c)



(d)

Fig. 3. Scatter plot of predicted errors for (a) Decision tree (b) linear regression (c) Lasso Regression (d) Ridge Regression

Linear regression and Ridge regression shows better accuracy than Lasso, Linear regression with accuracy 89.38 and Ridge regression with accuracy 89.53. Decision tree provides the most accurate estimation of crop yield for India country using statistical data. Machine learning does not interpret much, which makes it a black box technique. In this study, the machine learnings techniques are applied to predict the crop yield estimation for India, where Decision tree performed better than the other three regression methods.

V. Conclusion

As the population is proliferating, food demand and supply chain have become challenging to maintain. In the last many years, researchers have placed many efforts to predict crop yield production to help farmers. India is the country of villages and farmers. Technology can help farmers by estimating crop yield. This research implements machine learning techniques to predict the crop yield for India. The prediction so far has revealed that the decision tree performs better for country-level data. The study highlights the benefits of evolving techniques, as these techniques are associated with the agricultural dataset. This is beneficial for the small landholder farmers, from the prediction farmer can estimate the crop yield for the upcoming year and can grow the crop in obedience to prediction. The study can be taken forward by integrating remote sensing data with the statistical data.

References

[1] E. J. Coyle, "Rank order operators and the mean absolute error criterion," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 63–76, Jan. 1988.

[2] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Khon Kaen, Jul. 2016, pp. 1–5.

[3] P. S. Maya Gopal and R. Bhargavi, "Optimum Feature Subset for Optimizing Crop Yield Prediction Using Filter and Wrapper Approaches," *Applied Engineering in Agriculture*, vol. 35, no. 1, pp. 9–14, 2019.

[4] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction," *Span J Agric Res*, vol. 12, no. 2, p. 313, Apr. 2014.

[5] R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, Avadi,Chennai, India, May 2015, pp. 138–145.

[6] E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," *Environ Dev Sustain*, Aug. 2019.

[7] A. Mathur and G. M. Foody, "Crop classification by support vector machine with intelligently selected training data for an operational application," *International Journal of Remote Sensing*, vol. 29, no. 8, pp. 2227–2240, Apr. 2008.

[8] Kim, Nari and Lee, Yang-Won, "Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State," vol. 34, no. 4, pp. 383–390, Aug. 2016.

[9] I. Nitze, U. Schulthess, and H. Asche, "Comparison of Machine Learning Algorithms Random forest, Artificial neural network and support vector machine to maximum likelihood for supervised crop type classification," *Proc. Of The 4th GEOBIA* 35 May 2012.

[10] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," in *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Chennai, India, Jul. 2016, pp. 105–110.

[11] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Front. Plant Sci.*, vol. 10, p. 621, May 2019.

[12] A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," *Environ. Res. Lett.*, vol. 13, no. 11, p. 114003, Oct. 2018.

[13] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, no. 1, pp. 1–18, Jul. 2005.

[14] M. G. P. S. and B. R., "Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms," *Applied Artificial Intelligence*, vol. 33, no. 7, pp. 621–642, Jun. 2019.

[15] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti, "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil," *Agricultural and Forest Meteorology*, vol. 284, p. 107886, Apr. 2020.

[16] "FAOSTAT," 2017. http://www.fao.org/faostat/en/#rankings/countries_by_commodity_exports.

[17] Zhou Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[18] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014.

[19] D. M. Magerman, "Statistical decision-tree models for parsing," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* -, Cambridge, Massachusetts, 1995, pp. 276–283.

[20] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012.

[21] M. J. A. Chan-Lau, *Lasso Regressions and Forecasting Models in Applied Stress Testing*. International Monetary Fund, 2017.

[22] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*. John Wiley & Sons, 2006.