Crop Yield and Crop Production Predictions using Machine Learning

Divya Goel[1], Payal Gulati[2]
gulatipayal@yahoo.co.in
[1]Student, [2]Assistant Professor, Computer Engineering Department
J.C. Bose University of Science and Technology, YMCA, Faridabad

Abstract

Today Agriculture segment is a significant supporter of Indian economy as it represents 18% of India's Gross Domestic Product (GDP) and it gives work to half of the nation's work power. Farming segment are required to satisfy the expanding need of food because of increasing populace. Therefore, to cater the ever-increasing needs of people of nation yield prediction is done at prior. The farmers are also benefited from yield prediction as it will assist the farmers to predict the yield of crop prior to cultivating. There are various parameters that affect the yield of crop like rainfall, temperature, fertilizers, ph level and other atmospheric conditions. Thus, considering these factors the yield of crop is thus hard to predict and becomes a challenging task. Thus, motivated this work as in this work dataset of different states producing different crops in different seasons is prepared; which was further pre-processed and there after machine learning techniques Gradient Boosting Regressor, Random Forest Regressor, Decision Tree Regressor, Ridge Regression, Polynomial Regression, Linear Regression are applied and their results are compared using python programming.

Keywords

Agriculture, Machine Learning, Random Forest, Decision Tree, Linear Regression, Gradient Boosting Regression, Polynomial Regression, Ridge Regression.

1. IINTRODUCTION

India is a horticulturally based nation, where the greater part of the populace is reliant on this. It is one of the main occupations practised in our country and assumes a significant job in overall advancement of nation. As it is the backbone of Indian Economy. Around 60% of the land is utilized for agribusiness so as to satisfy the necessities of 1.2 billion individuals [21]. Thus, modernization in horticulture is significant and, in this way, will lead the ranchers of our nation towards benefit. Due to certain components such as environmental changes, uneven rainfall and water management, over the top utilization of pesticides and so on, the level of production is decreasing in India. For such variety of reasons most of the farmers don't accomplish expected yield of crop. To understand production levels, prediction of crop yield is applied which incorporates the prediction of yield of the crops on the data. Earlier, the estimation of crop yield was dependent on specific crops and the experience of cultivation of farmers.

There are various ways to deal with upgrade and improve the yield of harvest and their quality. Techniques related to Data Mining are likewise useful for anticipating yields of crop. Generally, Data Mining is utilized to examinations information from different frameworks and sums up it as a beneficial data. Data Mining programming [2] is a legitimate device that grants clients to describe and summarize apparent relationships just as examine information at different estimations or edges. In fact, Data Mining is discovering associations or examples of fields. These data can give information between connections, models or relationships. Then this knowledge can be changed into the recorded models and data on future examples or models. For example, a review of agrarian things urges ranchers to suggest and forestall future harvest incidents.

Manuscript received September 5, 2023
Manuscript revised September 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.9.3

Various analysts have been directed to build up a powerful strategy for yield forecast however centre have been reliably around measurable strategies and very little has been done in Machine Learning approach. The production of crop relies upon various different parameters [3] which may change with every meter square and depends on climatic conditions such as temperature, humidity, rainfall and pH value of soil, Region geography and also fertilizers. Different subsets of these parameters are utilized in various prediction models for different crops. Thus, prediction models are basically two fundamental sorts. There are Statistical models, that utilize single prediction work that incorporates all example spaces. Other is the Machine Learning Technology, another information for data search that able to connects the input variable with the output variable models.

Machine Learning can gain proficiency with the machine without characterized computer programming, so it improves machine execution by distinguishing and portraying the consistency and pattern of drive information. In this work, supervised machine learning techniques are applied for the prediction of crop yield. This type of technique helps to build generally precise and effective model since the learning data accompanies desired outputs and the goal is to find a general standard of mapping input to output. It includes building ML model that depends on labelled samples.

Till now, prediction is done on single state or single crop. But this proposed system explores the supervised machine learning approaches application in determining yield production of various crops and various different states and their districts. In this, various states data and different crop data is collected, analysed and initial data processing is completed. Techniques such as Linear Regression, Gradient Boosting Regressor, Random Forest Regressor, Decision Tree Regressor, Polynomial Regression, Ridge Regression have been used for yield prediction.

The paper is organized as follows: Section 2 covers the related work in the area of data mining and Machine learning and section 3 discusses the machine learning approaches applied in this paper for crop yield as well as production predictions. Section 4 covers the detail discussion on proposed system. The discussion and results are included in section 5. Finally, Section 6 concludes the research work.

2. RELATED WORK

In agriculture, Machine Learning is considered as a novel field, as variety of work has been done with the help of machine learning in the field of agriculture. There are different philosophies made and evaluated by the researchers all through the world in the field of agriculture and related sciences.

CH. Vishnu Vardhan Chowdary, Dr. K.Venkataramana [3], developed id3 algorithm for getting improved and great quality of crop yield of Tomato and is executed in Php platform and datasets are used as csv. Temperature, area, humidity and the production of tomato crop are the different parameters used in this study. R. Sujatha and P. Isakki [5], utilizes data mining techniques for prediction. This model worked on different parameters such as crop name, land area, soil type, pH value, seed type, water and also foreseen the boom and diseases of plants and in this way empowered to choose the descent crop dependent on climatic data and required parameters. N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy [8], proposed the SVM for crop yield prediction of rice. In this method, dataset used consists of different parameters such as place, temperature, precipitation and manufacturing. On this dataset, the implemented classifier

is sequential minimal optimization. They prepared the dataset through Weka tool to manufacture the set of rules on current dataset. In python, by using SVM algorithm outcomes were produced. S. Veenadhari, B. Misra and C. Singh [15], have built up an interactive site for finding the influence of climate and production of crop by utilizing c4.5 algorithm called Crop Advisor. Dependent on c4.5 algorithm, decision tree and ruled have been developed. It gives the idea how crop growth is affected by different climatic parameters. The data with respect to the related years environmental parameters like rainfall, temperature where gathered. The choices were dependent on the zone under the picked crop. Jun Wu, Anastasiya Olesnikova, Chi- Hwa Song, Won Don Lee [17], proposed selection tree which is fit for grouping all styles of farming records. A decision tree classifier turned into proposed for information of agriculture. It utilises new facts and can address each and in whole record. 10-fold cross validation method is utilised to check dataset, horse-colic and soybean dataset. Kiran Mai, C., Murali Krishna, I.V, A. Venu gopal Reddy [19], explained in their study that how data mining is incorporated with the other farming data such as meteorological data, usage of pesticides is useful for soothing out of use of pesticides. Topical information related to the business of agriculture which has contiguous properties was represented. Verheyen, K., Adrianens, M. Hermy and S.Deckers [20], explained statistical mining techniques in their study as they are regularly used to view the characteristics of soil. As K-mean is utilized for sectioning soils in blend with GPS based innovation.

3. MACHINE LEARNING APPROACHES

3.1 Random Forest

Random Forest is a supervised machine learning algorithm and is generally well known and powerful technique which is capable of accomplishing both classification and regression tasks, that works building large number of decision trees at the training time and outputting class that is mode of classes as classification or mean predictions as regression of each tree. The prediction is more powerful when there are more trees in forest. In this each tree is trained on subset of data as there are multiple trees. This algorithm can be developed in parallel.

Random Forest pseudocode:

1. Select randomly “i” features from total “n” features, where i << n
2. Calculate the node “m” among “i” features using the best split point.
3. Split the node into daughter nodes using best split.
4. Repeat steps 1 to 3 until “z” number of nodes has been reached.
5. Forest is built by repeating steps 1 to 4 for “y” number of times to create “y” number of trees.

3.2 Decision Tree

It is a predictive model which works by checking condition at each degree of tree and continues towards base of the tree where different choices are recorded. The condition relies upon the application and the result may be in terms of decisions. The different types of decision tree algorithm are c4.5, CART and ID3 algorithm.

3.3 Gradient Boosting

Gradient Boosting for regression builds an additive model in a forward stage-wise style, it takes into consideration the advancement of subjective differentiable loss functions. A regression tree is fit on negative gradient of given loss function in each stage. Boosting came out of whether a feeble learner

can be adjusted to turn out to be better. A frail speculation or feeble learner is characterized as one whose exhibition is in any event somewhat superior to random possibility.

3.4 Linear Regression

It is one of the most widely used data analysis and predictive modelling technique. It is a linear approach used to model a relationship between dependent and independent variables. It is expressed as y=mx+c ; where x is independent variable, y is dependent variable, m is slope of line and c is the intercept.

3.5 Polynomial Regression

It is a form of regression analysis in which relationship between independent variable i and dependent variable d is modelled as n^{th} degree polynomial in i. It fits non-linear relation between value of i and corresponding conditional mean of d, as E(d|i). As statistical estimation problem is linear, the regression function E(d|i) is linear in unknown parameters that are estimated from data.

3.6 Ridge Regression

It is a technique which is specialized to analyse multiple regression data which is multicollinearity in nature. It is fundamental regularization technique which may not be used by many due to complex things behind it.

4. PROPOSED SYSTEM

A Machine Learning (ML) manages issues where the connection among input and output factors isn't known or difficult to acquire. The "learning" term here signifies the programmed procurement of auxiliary depictions from instances of what is being described. Unlike traditional factual techniques, ML doesn't make presumptions about the right structure of the data model, which depicts the information. This characteristic is exceptionally helpful to model complex non-direct practices, for example, a capacity for crop yield expectation. ML strategies most effectively applied to Crop Yield Prediction (CYP). Supervised Learning techniques comprise of a dependent/outcome variable which is to be anticipated from a given arrangement of independent variables. Utilizing these arrangements of factors, function that guide inputs to outputs is found. The preparation process proceeds until the model accomplishes an ideal degree of exactness on the preparation data.

In the proposed framework, different states data are to be collected and different crops are taken into consideration from various different sources and thus dataset is obtained. Supervised learning is utilized for modelling, which gives the predicted yield and their order of production. The proposed framework is portrayed in following stages such as dataset collection, pre-processing the data, feature selection and apply different machine learning algorithms for getting the outcome. The work flow of the same is discussed below:

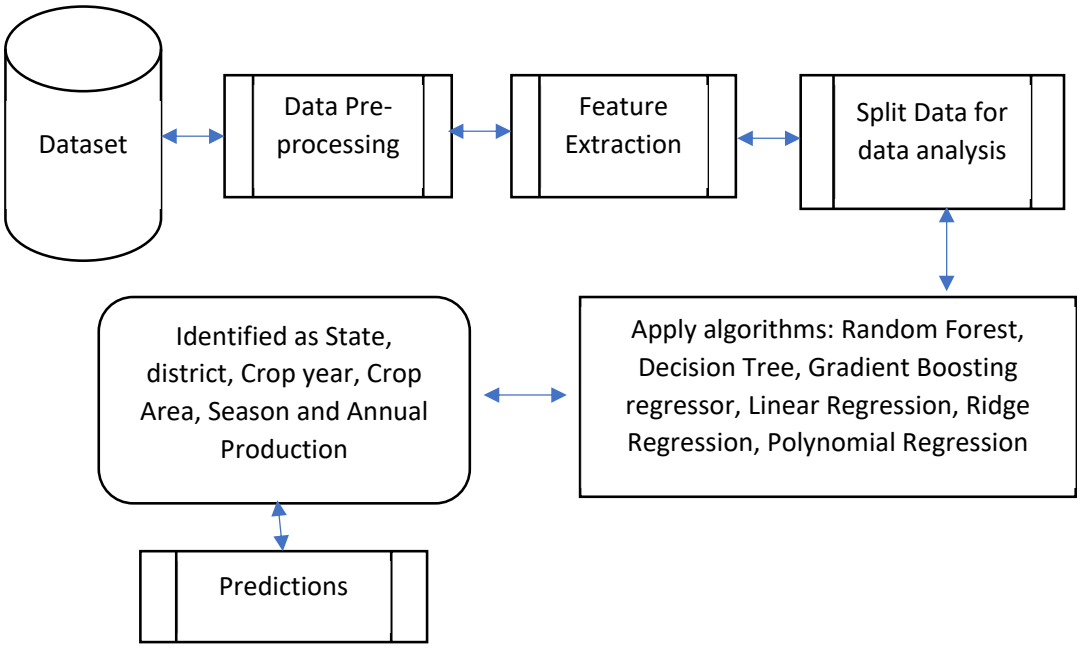

Figure 1. Architecture Crop Yield and Crop Production Prediction

4.1 Dataset Collection

Data is gathered from various sources. Previously, either a state or a single crop is used. In this paper, various datasets were merged together so predictions were made on novel dataset. This data is utilized for descriptive analysis. The dataset used in this paper consists of various states (Maharashtra, UP, West Bengal, Gujarat etc), different types of crops (sugarcane, coconut, wheat, gram etc), different seasons (Kharif, Rabi, Whole, Summer etc), different crop years and other parameters such as Rainfall, Temperature, pH, Humidity

4.2 Pre-processing the data

As it is known that pre-processing is considered as a key step before any algorithm is applied on the dataset. Here in this step, missing values in the dataset are treated.

4.3 Feature Extraction

Feature extraction ought to streamline the amount of data required to represent a huge dataset. Its goal is to extract useful characteristics from data. The characteristics include high, low and mean temperature, air humidity, soil pH, rainfall. In this phase, useful features were considered for the analysis purpose.

4.4 Split dataset into Train and Test set

This step includes training and testing of the input data. The stacked information is isolated into two sets, such as preparing and testing the data. Training set is mapped with the training set and during the training phase data is to be testing after learning from previous observations. The final data is formed and is processed by machine learning module.

4.5 Apply Machine Learning Techniques:

After the data is divided into training and testing set various machine learning algorithms are applied such as Random Forest, Decision Tree, Gradient Boosting regressor, Linear Regression, Ridge Regression, Polynomial Regression and based on these, crop yield as well crop production is predicted.

5. RESULT ANALYSIS

This section describes the outputs obtained after implementation of ML algorithms on the dataset obtained. Six different algorithms are applied on dataset. In this paper, Jupyter notebook, which is a suite of software products used in interactive computing is used for making crop yield and production predictions. The predictions made after applying machine learning techniques are shown in Figure 2-7 below.

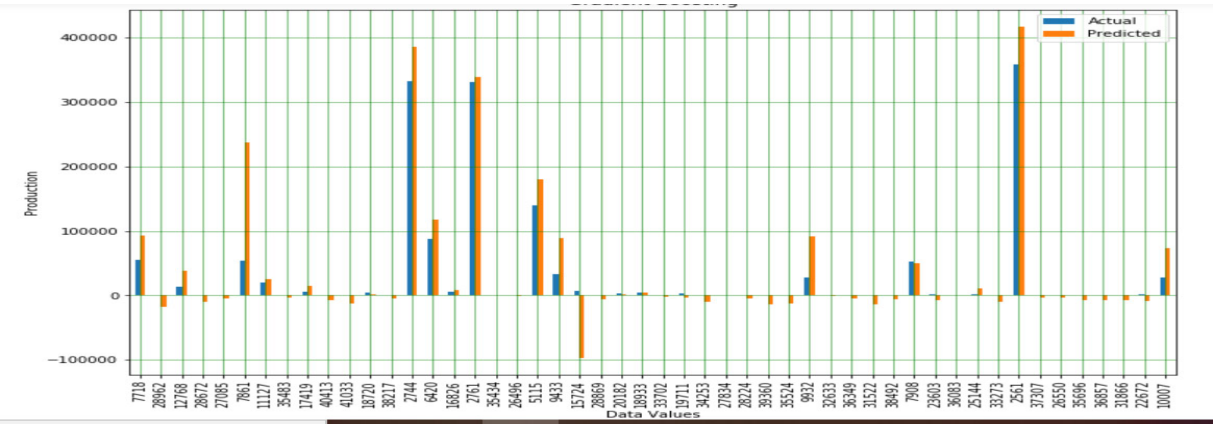

Figure 2 Predictions by Gradient Boosting

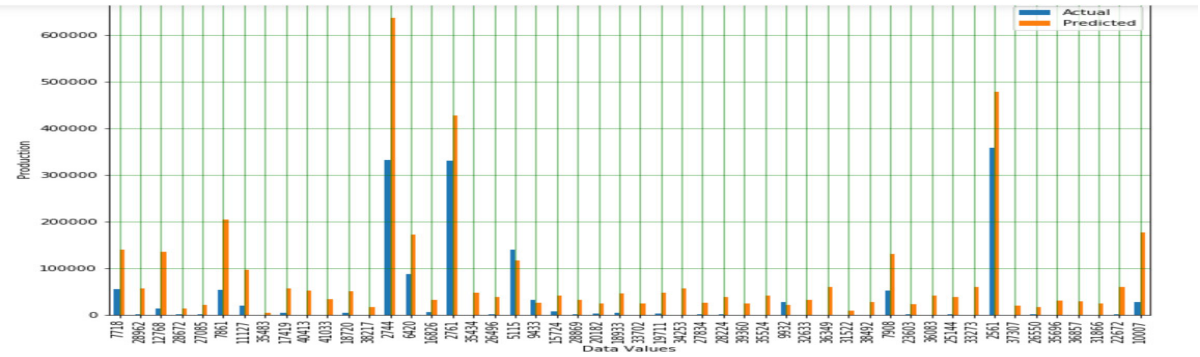

Figure 3 Predictions by Ridge Regression

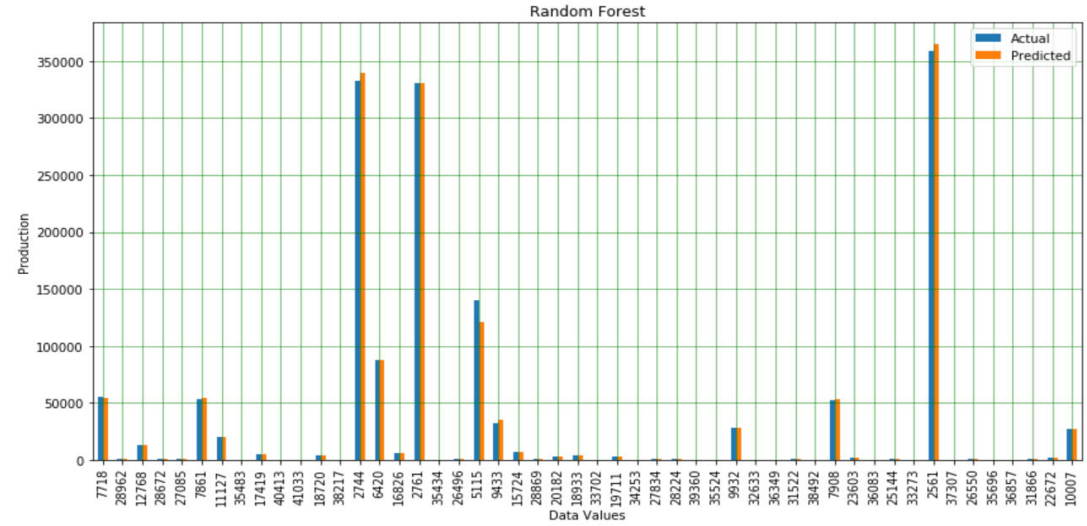

Figure 4 Predictions by Random Forest

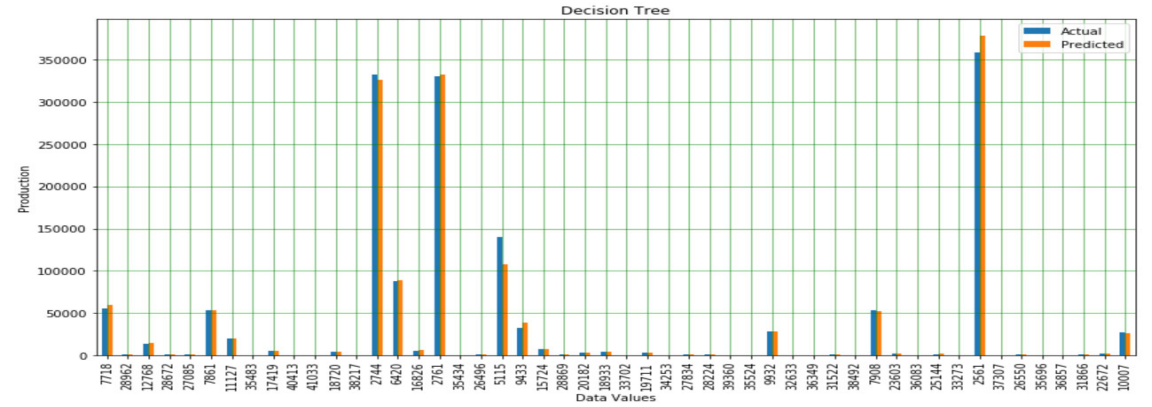

Figure 5 Predictions by Decision Tree

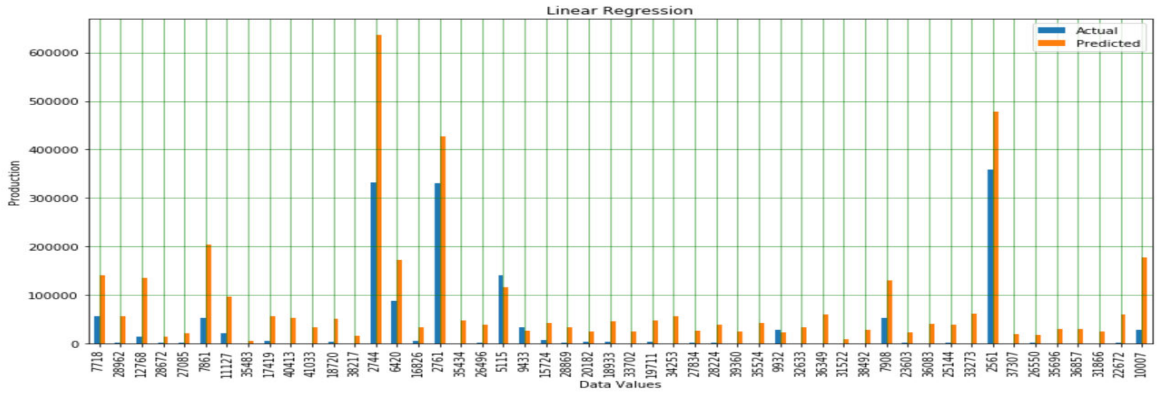

Figure 6 Predictions by Linear Regression

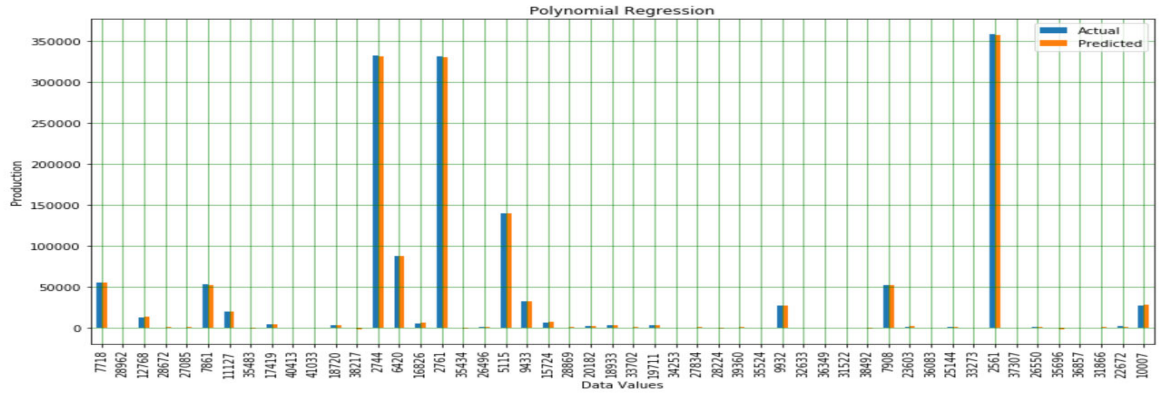

Figure 7 Predictions by Polynomial Regression

For analysing the accuracy of predictions made by machine learning algorithms mean absolute error, mean squared error, root mean square error, R-square and cross validation is considered in this work. The formulae for calculating parameters are given in eq(1), (2) and (3).

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \quad \ldots\text{eq(1)}$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \quad \ldots\text{eq(3)}$$

$$r^2 = 1 - \frac{\sum(y_i - \tilde{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad \ldots\text{eq(4)}$$

In this paper, two cases of target variable are considered. In first case, where target variable "yield" is considered, the Gradient Boosting Regressor is observed to be performing better with higher accuracy and cross validation of 87.9% as shown in Table 1. Figure 8 shows the comparison for cross validation accuracy when target variable is "yield".

Table 1 Different Techniques accuracy for target variable as "Yield"

	Test MAE	Test MSE	Test RMSE	Test R-Square	CrossValidation
Gradient Boosting Regressor	23.744808	4.676468e+05	683.847059	0.796290	0.879121
Ridge Regression	83.371910	2.289880e+06	1513.235102	0.002513	0.001220
Random Forest Regressor	19.437881	7.771160e+05	881.541815	0.661483	0.767536
Decision Tree Regressor	22.049008	1.137754e+06	1066.655599	0.504387	0.730118
Linear Regression	83.371914	2.289880e+06	1513.235103	0.002513	0.001220
Polynomial Regression	146.157727	2.236869e+06	1495.616746	0.025605	0.001220

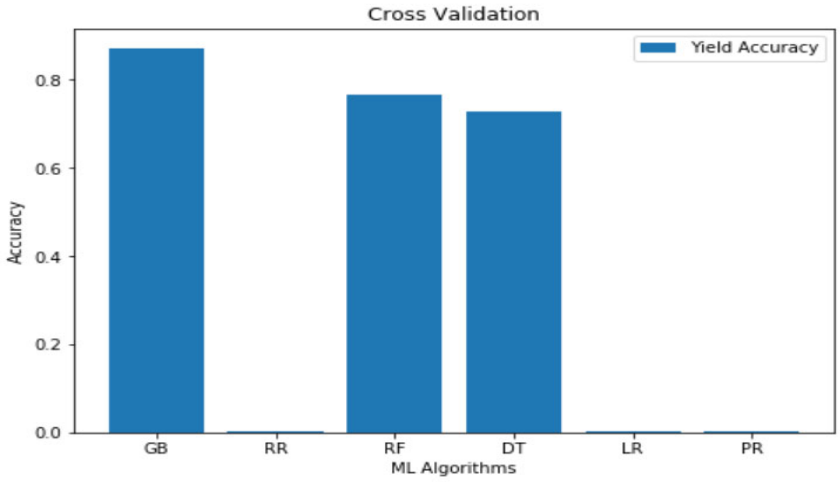

Figure 8 Comparison Graph for Cross validation Accuracy for Target Variable "Yield"

In another case, when target variable as "Production", then Random Forest Regressor is giving more accuracy with cross validation runs as 98.9% as shown in Table 2. Figure 9 shows the comparison for cross validation accuracy when target variable is "production".

Table 2 Different Techniques accuracy for target variable as "Production"

	Unnamed: 0	Test MAE	Test MSE	Test RMSE	Test R-Square	CrossValidation
0	Gradient Boosting Regressor	33085.368684	1.378599e+10	117413.755279	0.937892	0.940246
1	Ridge Regression	110728.210855	1.961227e+11	442857.411142	0.116443	0.109717
2	Random Forest Regressor	3560.027445	1.076494e+09	32809.969701	0.995150	0.989678
3	Decision Tree Regressor	5857.325847	1.929709e+09	43928.455944	0.991306	0.983545
4	Linear Regression	110728.211444	1.961227e+11	442857.411267	0.116443	0.109717
5	Polynomial Regression	1182.259271	9.462253e+08	30760.775980	0.995737	0.109717

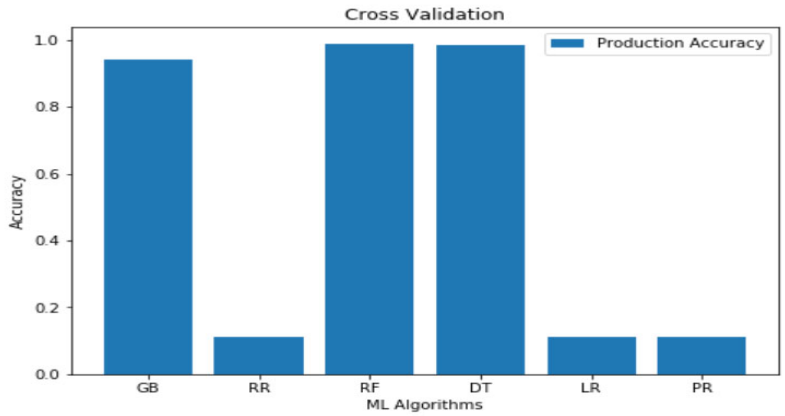

Figure 9 Comparison Graph for Cross validation Accuracy for Target Variable "Production"

6. CONCLUSION

In this paper, various techniques of machine learning have been applied on the agricultural data to evaluate the best performing technique. Here in this work six different supervised learning algorithms are considered. The dataset comprises of a variety of parameters that are useful for identifying status of crop. The results of these techniques were compared based on different errors and cross validation for obtaining accuracy. As per the observations, Gradient Boosting Regressor is giving more accuracy with cross validation runs as 87.9% when target variable is "Yield" but when target variable is "Production", the Random Forest Regressor is providing more cross validation accuracy i.e. 98.9%. This framework will assist to reduce the issues faced by farmers

and will serve as delegate to provide farmers with the information they need to gain high and maximize the profits.

REFERENCES

[1] Arun Kumar, Naveen Kumar, Vishal Vats, "Efficient crop yield prediction using machine learning algorithms", IRJET Volume: 05 Issue: 06, June-2018, pp 3151-3159

[2] P.Priya, U.Muthaiah, M.Balamurugan, "Predicting yield of the crop using machine learning Algorithm", IJESRT et al., 7(4): April-2018, pp 2277-2284

[3] CH. Vishnu Vardhan chowdary, Dr.K.Venkataramana, "Tomato Crop Yield Prediction using ID3", March 2018,IJIRT Volume 4 Issue 10 pp,663-62.

[4] Vaneesbeer Singh, Abid Sarwar, Vinod Sharma. "Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach", IJARCS 8 (5), May-June 2017, pp 1254-1259.

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

[6] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm", Agronomy for Sustainable Development, vol. 36, no. 2, 2016.

[7] Ankalaki, S., Chandra, N., Majumdar, J, "Applying Data Mining Approach and Regression Model to Forecast Annual Yield of Major Crops in Different District of Karnataka", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Special Issue 2, pp.25-29, 2016.

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

[9] Aditya Shastry, H.A. Sanjay, E. Bhanusree, "Prediction of Crop Yield using Regression Analysis", Indian Journal of Science and Technology, Vol. 9(38), pp.1-5, 2016.

[10] Kushwaha, A.K., Sweta Bhattacharya, "Crop yield prediction using Agro Algorithm in Hadoop", International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 5- No2, pp.271-274, 2015.

[11] D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473, 2015.

[12] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach", 2015 International Conference on Computational Intelligence and Communication Networks.

[13] Awanit Kumar, Shiv Kumar, "Prediction of production of crops using K-Means and Fuzzy Logic", IJCSMC, 2015.

[14] Fathima, G.N., Geetha, R., "Agriculture Crop Pattern Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Engineering, Vol. 4, Issue 5, pp.781-786, 2014

[15] S. Veenadhari, B. Misra and C. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters" 2014 International Conference on Computer Communication and Informatics, Coimbatore, 2014, pp. 1-5.

[16] Kaur, M., Gulati, H., Kundra, H., "Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications", International Journal of Computer Applications, Vol. 99– No.12, pp.1-3, 2014

[17] Jun Wu, Anastasiya Olesnikova, Chi- Hwa Song, Won Don Lee (2009), "The Development and Application of Decision Tree for Agriculture Data" IITSI, pp 16-20.

[18] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. "Estimation of neural network parameters for wheat yield prediction" In Max Bramer, editor, Artificial

Intelligence in Theory and Practice II, volume 276 of IFIP International Federation for Information Processing, pages 109–118. Springer, July 2008.

[19] Kiran Mai,C., Murali Krishna, I.V, an A.Venugopal Reddy, "Data Mining o f Geospatial Database for Agriculture Related Application", Proceedings of Map India,New Delhi, 2006,pp 83-96.

[20] Verheyen, K., Adrianens, M. Hermy and S.Deckers (2001), "High resolution continuous soil classification using morphological soil profile descriptions" Geoderma, 101:31-48.

[21]https://www.hilarispublisher.com/open-access/agriculture-role-on-indian-economy-2151-6219-1000176.pdf.