International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Crop Yield Prediction using XG Boost Algorithm

Rohit Ravi, B. Baranidharan

Abstract*: The main objective of this research is to predict crop yields based on cultivation area, Rainfall and maximum and minimum temperature data. It will help our Indian farmers to predict crop yielding according to the environment conditions. Nowadays, Machine learning based crop yield prediction is very popular than the traditional models because of its accuracy. In this paper, linear regression, Support Vector Regression, Decision Tree and Random forest is compared with XG Boost algorithm. The above mentioned algorithms are compared based on R^2, Minimum Square Error and Minimum Absolute Error. The dataset is prepared from the data.gov.in site for the year from 2000 to 2014. The data for 4 south Indian states Andhra Pradesh, Karnataka, Tamil Nadu and Kerala data alone is taken since all these states has same climatic conditions. The proposed model in this paper based on XG Boost is showing much better results than other models. In XG Boost R^2 is 0.9391 which is the best when compared with other models.

Keywords: Crop yield, precision agriculture, Random forest, Support vector machine, XG Boost Algorithm Artificial Intelligence, Machine learning.

I. INTRODUCTION

Modern day civilization started with the agriculture. Without agriculture, civilization won't exist and thrived up to now. According to the record, India is the second most agricultural product production country next to China. According to common estimates, more than 50% population of our country depends on agriculture and its related industry. Thus agriculture is the backbone of our country and still with the advancement of information technology sector, agriculture is providing 50% of the employment to our country [1]. Currently, it supplies 17.32% of GDP and its GVA (Gross Value Added) is nearby 23.82 Lakh crore in the financial year 2016-17 [2]. From the statistical data, it is observed that India is producing 7.39% of total global agricultural product. In 2017-18 it is estimated that around 275 million tons of grains is produced in India [3].

But still as mentioned previously, India is not on the top in production. The main reason behind this is that still in many part of our country, farmers are not exposed to using the appropriate technology in agriculture field. We must have to take the help of technologies to make our agriculture practices better and enhanced. In this section, we are going to learn these topics which will help us in enhancing the way of cultivation. In this article, we are going to discuss about the prediction of crop yield & proposal scheme that will guide us in cultivating better crop for the particular region to increase their production. The confidential crop field pictures, the latest climate data and harvest data of the last few years are processed to envisage the forecasted crop harvest and even prescribe convenient product for a specific farmland. The proposed model is based on machine learning techniques, image processing, image segmentation, etc. for crop yield forecasting.

Manuscript published on January 30, 2020.
* Correspondence Author
Rohit Ravi*, M.Tech, Department of Computer Science & Engineering, SRM Institute of Science & Technology, Chennai, India. Email: rohitravi29@gmail.com
Dr. B. Baranidharan, Associate Professor, Department of Computer Science & Engineering, SRM Institute of Science & Technology, Chennai, India. Email: baranidb@srmist.edu.in

II. MACHINE LEARNING TECHNIQUES

Machine learning is an element of Artificial Intelligence which provides ability to learn without being specifically programmed for its learning activity. It helps in automating the evaluation and processing the huge amount of data previously collected. Therefore, it is going to remove some of the burden on the man power with improved accuracy.

Machine learning targets on the development of computer programming which is going to advance the applied domain area using previous data. Finding out the crops yield based on the soil's appearance becomes annoying for the farmers. There is also a need to counter the agricultural pollution, and improve its productivity.

Efficient usage of agricultural land is urgent for establishing food preservation of the country. In this paper, we propose a crop yield prediction system based on XG Boost algorithm. Machine Learning [13] is generally classified in three parts: -

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Some of the machine learning algorithm which may help us in making our research successful are described below:-

A. Linear Regression

In Machine learning, we are having input variables which are used to find out output variable. And the relation between these input and output is being computed by machine learning. And In linear regression, [14] this relationship is described in below form of equation,

$$y=a+bx \quad (1)$$

The main aim of linear regression is to find the best values for 'a' and 'b' variables to represent the model.

B. Decision Tree Regression

The Random forest is built based on large collection of decision trees. Generally, decision trees is described on the basis of questions on each properties of the datasets. As more properties will be there as more the tree will be precise. Increasing the number of trees will lead us to more accurate result. This is an interpretable model because it makes classifications much similar to us. For example we ask a sequence of queries on any topics until we will get the perfect answers on that. Similarly Decision tree also ask too many question on the basis of features till it get the best predicted result.

C. Support Vector Regression

Support Vector Regression algorithm is almost similar to Support Vector Machine algorithm. But SVR is little different from SVM. Support Vector Regression is a regression algorithm in which we work with continuous values.

The main difference between Support Vector Regression and simple regression is that in simple regression we are trying to minimize the error rate while in Support Vector Regression we are trying to fit the error within a certain threshold.

In Support Vector Machine algorithm we are using two lines as a boundary lines but in Support Vector Regression we are using one hyper plane instead of two boundary line.

D. XG Boost Algorithm

XG Boost means Extreme Gradient Boosting Algorithm. The main reason behind using this are as follows:-

- Execution Speed: - XG Boost is faster than all others gradient boosting. According to the survey done by Szilard Pafka, in almost every case XG Boost is faster from all other performance from R, Python Spark and H2O. Its accuracy level is also high and it is also memory efficient.
- Model Performance: - XG Boost mainly dominates on the datasets which are either well-structured or tabular in form. This helps in the predictive modeling problems which are based on classification and regression.

E. Random Forest Algorithm

Random forest algorithm [15] represents the random collection of small trees generated from the given dataset. Random Forest Algorithm is used for both classification and regression. Random forest classifier has the ability to handle the missing values and the garbage values in the dataset. The pseudocode for random forest algorithm has two following stages:

I. Creation of Random forest.

II. Predicting the output of the test data from the created Random forest classifier.

I. Creation of Random Forest:

i. Randomly prefer "x" characteristics from total "z" characteristics. Where, x << z

ii. From the "x" characteristics, calculate the node "f" using the best middle point.

iii. Divide the node into daughter nodes using the best division.

iv. Repeat steps I to III until "l" number of nodes arrived.

v. Construct forest by replaying steps I to IV for "y" times to construct "y" number of trees.

In the starting of random forest algorithm, it starts with accidently or aimlessly selecting "x" characteristics out of total "z" features. After that, in the next stage we are going to dig out the node (root) by using the best division strategy by the randomly selected "k" features. In next stage, we are going to use the same best split approach for calculating the daughter nodes. Then after we will repeat the first three steps till we create the tree with a root node and having leaf node as the target. In last, we are repeating stage I to IV to construct "y" randomly constructed trees. This randomly created trees will create the random forest.

II. Random Forest Prediction Pseudocode:

To execute forecasting using the trained random forest algorithm the following pseudocode is used.

- Select the test qualities and utilize the rules of every incidentally created decision tree to forecast the outcome and stock the concluded result.
- Determine the votes for every concluded result.
- Recognize the highly voted concluded result as the last conclusion from the random forest algorithm.

To implement the forecasting with the help of the trained random forest algorithm firstly, we have to go through the test features through the rules of every aimlessly manufactured trees. Let's see we build 100 random decision trees to build the random forest. Every random forest will forecast distinct result for the same test feature. Then by examining every forecasted result votes will be calculated.

Let's consider the 100 random decision trees are forecasting some 3 different destination a, b, c then votes of 'a' is none other than but out of 100 random decision tree how many trees forecasting is a. Similarly for other two target (b, c). If 'a' is getting more number of votes. Let's imagine out of 100 random decision tree 60 trees are forecasting the target will be a. Then finally, random forest will return a as the forecasted target. The target which has maximum number of voting will be forecasted as the result.

The concept behind voting is to achieve the majority voting so that on basis of that we can predict the result. And later, random forest algorithm work on the voting received from the decision tree. Random forest, is much like to its name that implies, having a large amount of particular decision trees which perform together. Each particular tree in the random forest divides out a class forecasting and the class with the more number of votes becomes our model's prediction.

III. LITERATURE SURVEY

Shivnath Ghosh et al [4] uses Back Propagation Neural Network, Supervised Learning to find the suitable correlation percentage of the properties such as organic material, plant nutrients, micronutrients which affect the growth of the crops.

According to him BPN is considered to be an important component in the near future for the intervention of technological support to agricultural sector.

The learning part of BPN is completed in three major steps:-

- Feed forward Phase.
- Back Propagating of the errors.
- Adjusting the weights.

Zhihao et al [5] uses SVM techniques and RVM techniques to forecast the moisture of the soil. They uses some electronic devices for this namely "MicaZ mote" and "VH400". They used Machine Learning Algorithm which need a large data size. So they took Illinois historical data. It achieves 15% error rates and high correlations of 95%.

Vinciya et al [6] uses the Multiple Linear Regression (MLR) techniques to analysis the crop yield prediction. They also used decision tree algorithm for structured prediction and supervised learning algorithm for classification. They used three main algorithm for their research. These are as follows:

1. Decision tree algorithm (Structured Prediction)
2. Classification (Supervised Learning Algorithm)
3. Prediction (Multiple Linear Regressions)

- Decision Tree Algorithm:

In Decision Tree, there is a structured prediction or structured learning that involve predicting structured objects based on supervised machine learning. It don't take the input data like scalar discrete or real values. Let's take a prediction

$$\hat{y} = \arg\max \{y \in GEN(c)\} (d^T \Phi(c, y)) \quad (2)$$

Update d, from

$$\hat{y} \text{ to } g: d=d+h(-\Phi(c, \hat{y})+ \Phi(c, g)) \quad (3)$$

h is learning rate.

- Classification (Supervised Learning Algorithm)

The risk $R_{(k)}$ of function k is defined as the expected loss of k. This can be estimated from the training data as

$$R_{emp}(k) = 1/N \sum L(y_i, k(c_i)) \quad (4)$$

- Multiple Linear Regressions:

The values fit in $\hat{y}_i$ is given by the equation

$$f_0+ f_1 f_{i1}+ ... + f_p c_{ip} \quad (5)$$

And the residuals e_i that is the difference between the observed and fitted values computes the value of $y_i - \hat{y}_i$. The calculation of the residuals is equal to zero (0). The variance σ_2 which is also known as the mean-squared error (MSE) is estimated by

$$s^2 = \frac{\sum ei2}{n-p-1}, \quad (6)$$

And the square root of MSE is the estimate of the standard error.

Ying Ding et al [7] used Model Predictive Control (MPC) methods. According to him Model Predictive Control method is more suitable than Process Control Model which is originated in industries. This is because it can effectively address non-linear and large time delay system. They described about how the MPC get developed in these days:-

1. Classical MPC (1960)
2. Improved MPC (2007)
3. Latest MPC (2011)

There are more classification in each period of MPC with its benefits and drawbacks.

Anuja Chandgude et al [8] used Machine Learning, sensors, artificial neural network, etc. to find the growth description, crop prediction, crop disease, etc. The data prediction they used artificial neural network in which the types available are:-

- Perceptron
- Multi-layered Perceptron
- Recurrent Neural Network
- Self Organizing Maps

Components used for architecture design are as follows:-
Sensors

- Microcontroller
- XBee Module
- IoT Module
- Prediction

Subhadra Mishra et al [9] checks environmental factors which affects the crop mostly. For example, weather condition, soil properties, fertilizers, irrigation, etc. They used machine learning techniques which are as follows:-

- Artificial Neural Network
- Information Fuzzy Network
- Decision Tree
- Regression Analysis
- Clustering
- Principal Component Analysis
- Bayesian Belief Network
- Time Series Analysis
- Markov Chain Model

Noran S.Ouf et al [10] has gone through many process like crop disease prediction, crop yield prediction, weather forecasting, smart identification. In crop yield prediction they used Back Propagation Network and multi linear regression. They considered several parameters like precipitation amount, crop biomass, soil evaporation, transpiration fertilizers, etc. They also used decision tree, support vector machine, hidden Markov model, etc. for machine learning. They differentiated machine learning in supervised and unsupervised.

Fabrizio Balducci et al [11] and his team has gone through some phases:-

- Data Sources
- Machine learning Task Design
- Software Tools

In first phase they are collecting the datasets from three different sources. These sources are:-

- CNR(National Research Council Scientific Datasets)
- Istat Statistical Datasets
- IoT Sensors

In second phase for Machine Learning Task Design they have portioned in 5 tasks:-

- Forecasting future Datasets(Istat)
- Comparison b/w Machine Learning Algorithm on vanished data(CNR Scientific Datasets)

- Reorganizing of vanished data from observation terminals manipulating neural network & linear and polynomial regression models(IoT sensor Datasets)
- Reorganizing of deleted Data from observation terminals manipulating the decision tree, & polynomial & K-Nearest Neighbors (KNN) algorithm (IoT Sensor Dataset)
- Disclosure of damaged observation terminals by Sensor Values (IoT Sensor Datasets)

IV. EXPERIMENTAL RESULTS

The main aims of this research work is to predict the yield of crop in a specific climate condition. At first, we are going to collect the data and on that datasets we are applying four algorithms to check the accuracy of the algorithm. We are taking linear regression, Decision tree regression, Random Forest Algorithm and XG Boost algorithm. We collected the data from the government website www.data.gov.in like maximum temperature minimum temperature, season wise rainfall, area of the cultivated land, and production of the rice. I arranged the data for the year 2000-2014 and it is also arranged in district wise for the selected four state Tamil Nadu, Andhra Pradesh, Kerala, Karnataka.

The below Fig.1 shows the comparison of R^2 of the five algorithm which we applied on the collected datasets. After applying algorithm we get the value of R^2 and these values are shown in below graph. For linear regression the value of R^2 is 0.8885 i.e. the accuracy level for linear regression is 88.85%. Similarly for Decision Tree Regression the value of R^2 is 0.8922 i.e. the accuracy level for decision tree regression is 89.22%. Similarly, for Random Forest Algorithm and XG Boost Algorithm the value of R^2 is 0.9314 and 0.9391 respectively i.e. the accuracy level will be 93.14% and 93.91% respectively. For R^2 how much the high the value of R^2 that much accurate that algorithm will be. So, from above result the XG Boost Algorithm has the best value so it will be the best algorithm for predicting the crop yield.

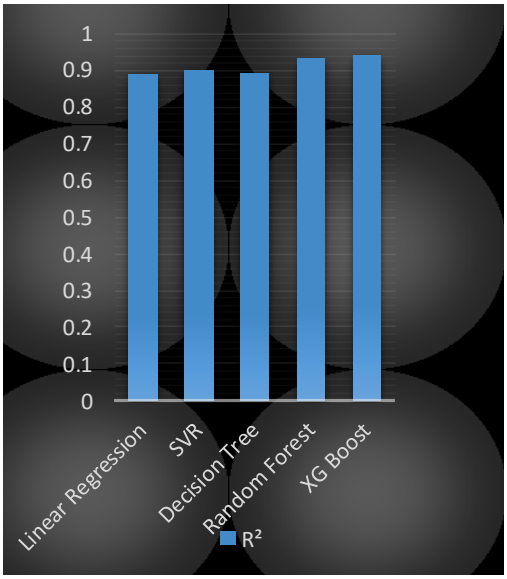

Figure 1: - Comparison of R^2

The below Fig.2 is the comparison graph of Mean Square Error (MSE) of four algorithms. Since for normalized value is not used for SVR its MSE value is not compared with other values. The value of MSE for Linear Regression is 2972466269.2335, while for Decision Tree Algorithm its value is 2847222194.9642, for Random Forest Algorithm its value is 2178239719.7533 while for XG Boost Algorithm its value is 1999378847.4874 which is least from all the used algorithm. As we know that how much less the MSE value of an algorithm that much high accuracy and perfect is that algorithm. So from this result also, the XG Boost Algorithm is the best one.

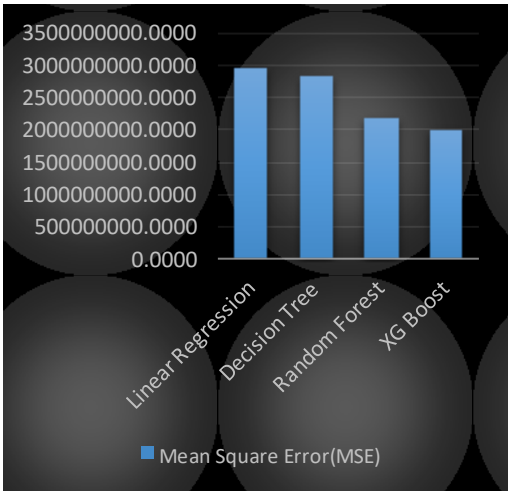

Figure 2: - Comparison of Mean Square Error (MSE)

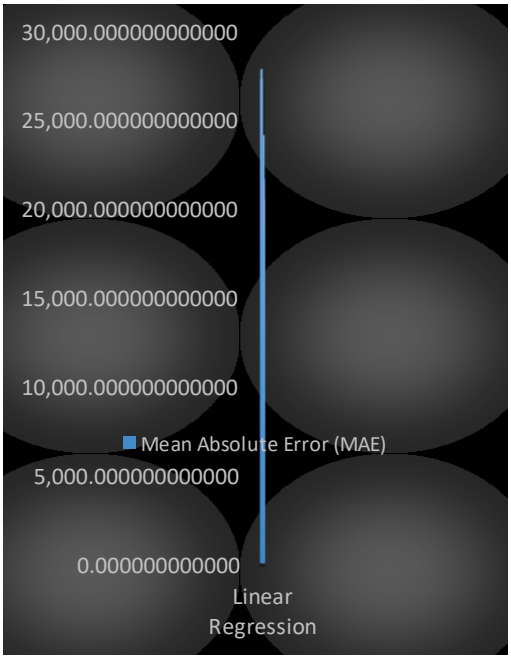

Figure 3: - Comparison of Mean Absolute Error (MAE)

The above Fig.3 is the comparison graph of Mean Absolute Error (MAE) of five algorithm. The value of MAE for Linear Regression is 27883.626823467002, for SVR its value is 27402.1696, while for Decision

Tree Algorithm its value is 24187.307644628097 while for Random Forest Algorithm its value is 21699.01751446281 while for XG Boost Algorithm its value is 20613.23612695489 which is least from all the used algorithm. Similar as MSE in this case also as lower the MAE value that much higher will be the accuracy of that algorithm.

Table 1: - Comparison of R^2, MSE, MAE

Model	R^2	MSE	MAE
Linear Regression	0.8885	2972466269.2335	27883.6268
SVR	0.8991	NA	27402.1696
Decision Tree Algorithm	0.8922	2847222194.9642	24187.3076
Random Forest Algorithm	0.9314	2178239719.7533	21699.0175
XG Boost Algorithm	0.9391	1999378847.4874	20613.2361

V. CONCLUSION

From the above results we can easily find that the accuracy level of XG Boost algorithm is higher in every case either it is R2 or MSE or MAE. The data used for building the model are gathered form data.gov.in for the year from 2000 to 2014. Area of cultivation, maximum temperature, minimum temperature and rainfall are the input variable and production is the output or dependent variable. The proposed XG Boost algorithm is compared with linear regression, SVR, Decision Tree and Random forest. XG Boost model gives better performance than all other models. In future, it is planned to optimize the hyper parameters of each traditional algorithms and should be tested with new data sets.

REFERENCES

1. https://www.omicsonline.org/open-access/agriculture-role-on-indian-economy-2151-6219-1000176.php?aid=62176
2. https://m.jagranjosh.com/general-knowledge/what-is-the-sectorwise-contribution-in-gdp-of-india-1519797705-1
3. http://statisticstimes.com/economy/sectorwise-gdp-contribution-of-india.php
4. Shivnath Ghosh, Santanu Koley, "Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks" IJRITCC, vol. 2, Issue 2,292-297,2014.
5. Zhihao Hong,Z. Kalbarczyk,R. K. Iyer, "A Data-Driven Approach to Soil Moisture Collection and Prediction" IEEE Xplore,vol. 2, Issue 2,292-297,2016.
6. P. Vinciya, Dr. A. Valarmathi, "Agriculture Analysis for Next Generation High Tech Farming in Data Mining" IJARCSSE,vol. 6, Issue 5, 2016.
7. Ying Ding, Liang Wang, Yongwei Li, Daoliang Li, "Model predictive control and its application in agriculture: A review" Computers and Electronics in Agriculture, June 2018.
8. Anuja Chandgude, Nikita Harpale, Diksha Jadhav, Punam Pawar, Suhas M. Patil, "A Review on Machine Learning Algorithm Used For Crop Monitoring System in Agriculture", IRJET April 2018.
9. Subhadra Mishra, Debahuti Mishra and Gour Hari Santra, "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper", Indian Journal of Science and Technology, October 2016.
10. Noran S. Ouf, "A Review on the Relevant Applications of Machine Learning in Agriculture", IJIREEICE, August 2018.
11. Fabrizio Balducci, Donato Impedovo and Giuseppe Pirlo, "Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement" Machines, 1 September 2018.
12. Hemant kumar Wani, Nilima Ashtankar. An Appropriate Model Predicting Pest/Disease of Crops Using Machine Learning Algorithm. [C]//ICACCS 2017.
13. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?
14. https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/
15. https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

AUTHORS PROFILE

Rohit Ravi, has completed his Bachelor of Technology in computer Science and Engineering from Skyline Institute of Engineering & Technology in year 2016-17 with first division. Now he is a final year student of Master of Technology in Computer Science and Engineering of SRM Institute of Science & Technology. This is the first time for him when he is going to publish a paper on his survey under the guidance of his guide Dr. B. Baranidharan (Associate Professor) regarding his project in the Scopus journal IJRTE. He is doing survey in the field of agriculture to predict the yield of the crop and how to enhance the productivity of crop using technology.

Dr.B.Baranidharan, has completed his Master of Technology in Computer Science and Engineering from SRM IST, Chennai and PhD in Wireless Sensor Networks (specialization) from SASTRA Deemed University, Thanjavur. Currently, he is working as Associate Professor in the department of CSE, SRM IST. He is having more than 10 years of academic experience and have published 22 papers in various International Journals and Conferences. Earlier his research involved about designing new clustering architecture for Wireless Sensor Networks and Internet of Things using various computational techniques. His current research includes Artificial Intelligence, Machine learning, Deep learning and Internet of Things.