International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Forecasting Crop Yield through Classification Approaches of Machine Learning

R. Akshara, Deepak Sanaka, Anusha Vaspari, J. Uday Kiran

Abstract*: Crop yields are critically dependent on weather. Data Mining and Machine Learning help a great deal in forecasting the yield data beforehand which would be beneficial to the farmers who could then plan the irrigation procedures according to the predicted yield. We make use of already available data to get the forecast figures. We propose to use supervised learning techniques specifically classification models on the available data to get the forecast figures. This paper focuses on wheat yield across the country with analysis done for mid 2017- mid 2018 data.

***Index Terms*: Data Mining, Machine Learning, Yield Forecasting.**

I. INTRODUCTION

Agriculture forms the spine of Indian economy. Agriculture in India is largely dependent on monsoons which are highly unpredictable. Historically, farmers are heavily reliant on the weather data alone to start the irrigation procedures. Weather data when combined with detailed analysis of previous yield numbers provides a solution for improving the current and forthcoming yields. Analysis of crop yield is made possible by the advancements in use of machine learning and data mining in agriculture.

II. MACHINE LEARNING IN AGRICULTURE

Machine Learning now spans across many sectors and nowadays including agriculture. Machine Learning in agriculture helps in improving productivity and the quality of crop. The seed retailers utilize this farming innovation to churn the information to create way better crops. Whereas the pest control companies are utilizing them to recognize the different bacteria's, bugs and vermins. The Machine Learning technologies are used to determine which corp and which conditions will produce the better yield. It will also determine which weather condition gives the highest returns.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
R. Akshara*, Assistant Professor, Department of Computer Science and Engineering, Vignan Institute of Technology and Science College in Hyderabad, India
J. Uday Kiran, B.Tech (CSE), Vignan Institute of Technology and Science, College in Hyderabad, India
Deepak Sanaka, B.Tech (CSE), Vignan Institute of Technology and Science, College in Hyderabad, India
Anusha Vaspari, B.Tech (CSE), Vignan Institute of Technology and Science, College in Hyderabad, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

III. RELATED WORK

In 2014, S.Veenadhari, Dr.Bharat Misra and Dr.CD Singh has used Crop Advisor for developing user friendly webpage for predicting the crop yield based on climatic parameters and C4.5 for finding the most influencing parameter. It concluded that the accuracy of predictions are above 75 per cent and it can be used by any user by providing the climatic data [5]. In The 2014, Miss.Snehal S.Dahikar, Dr.Sandeep V.Rode has used Artificial Neural Networks for predicting the crop yield by sensing various parameters of soil and atmosphere. It concluded that Artificial Neural Network is powerful tool for prediction of crop yield and analyse in matlab for better predictions [6]. In 2015, D Ramesh and B Vishnu Vardhan has data mining techniques we have found used data mining techniques for estimating the crop production. It concluded that the results of two methods namely multiple linear regression technique and density-based clustering technique are compared according to particular region to improve and authenticate the validity of crop yield prediction [7]. In 2017, Talha Siddique, Dipro Baruna, Zannatul Ferrous and Amithaba Chakrabarthy have used various algorithms namely multiple linear regression and k-nearest neighbor regression. It concluded that multiple linear regression gave the accurate predictions [8]. In 2018, Nanyang Ziu, Xu liu, Kai Hu and Ya Guo have provided deep learning algorithms which provides concepts, limitations, implementation and training process. It concluded that Deep learning can be used in agriculture information processing, agriculture production system optimal control, smart agriculture machinery equipment and agricultural economic system management [9]. In 2018, Fabrizio Balducci, Donato Impedevo and Giuseppe Pirlo have explained how to manage heterogeneous information and data coming from real datasets [10].

IV. DATASET ANALYSIS

To achieve Crop Yield Prediction, we have used a wheat yield dataset obtained from kaggle.com. There are a total 177492 records in the dataset. It contains the following attributes.
State, Latitude, Longitude, Date, apparentTemperatureMax, apparentTemperatureMin, cloudCover, dewPoint,humidity, precipIntensity, precipIntensityMax, precipProbability, precipAccumulation, precipTypeIsRain, precipTypeIsOther, pressure, temperatureMax, temperatureMin, visibility, windBearing, windSpeed, NDVI, DayInSeason, Yield

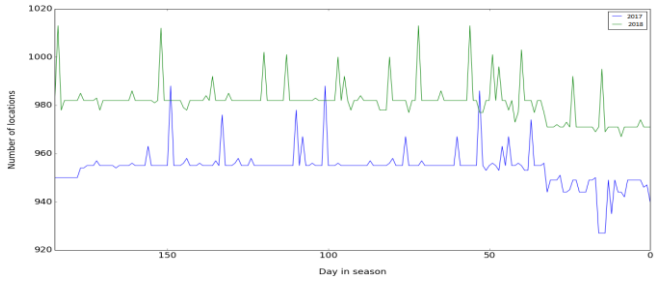

Fig.1, Plot of No. of Locations w.r.t. Day in season

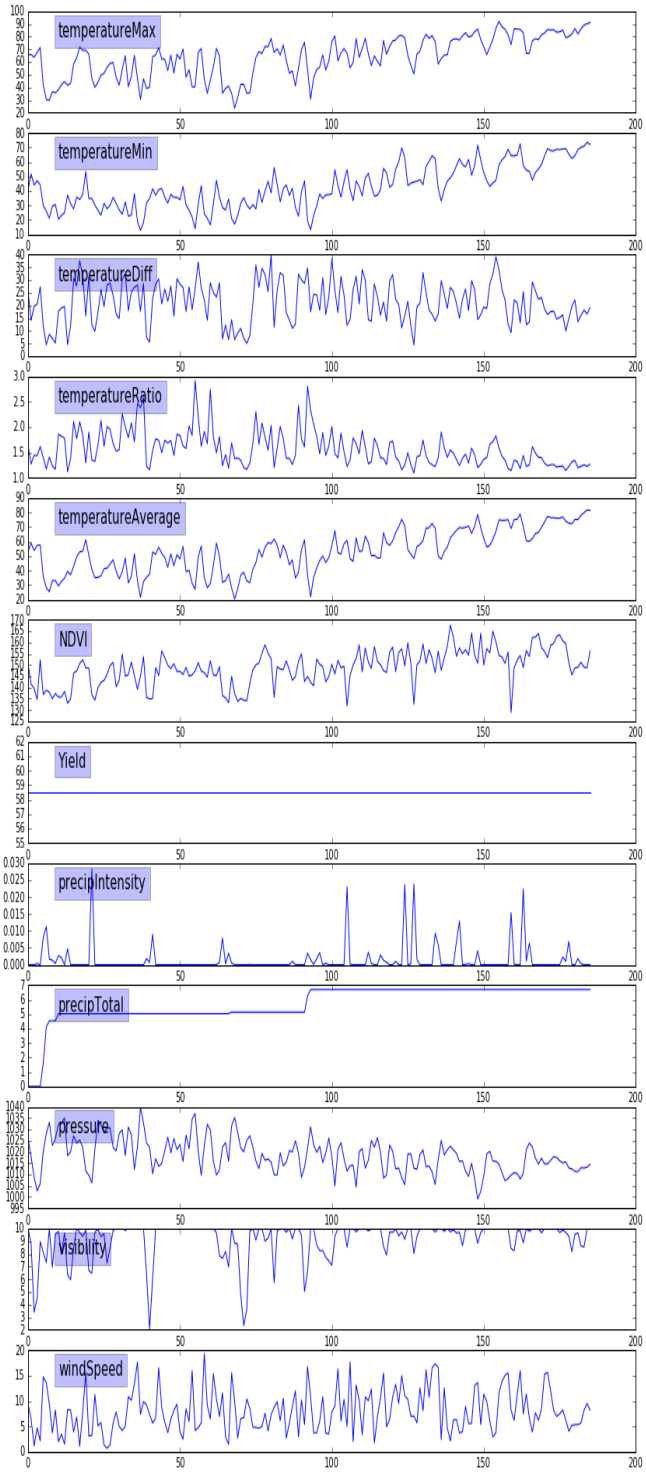

Fig.2. Plot of change in attributes w.r.t No. of Days

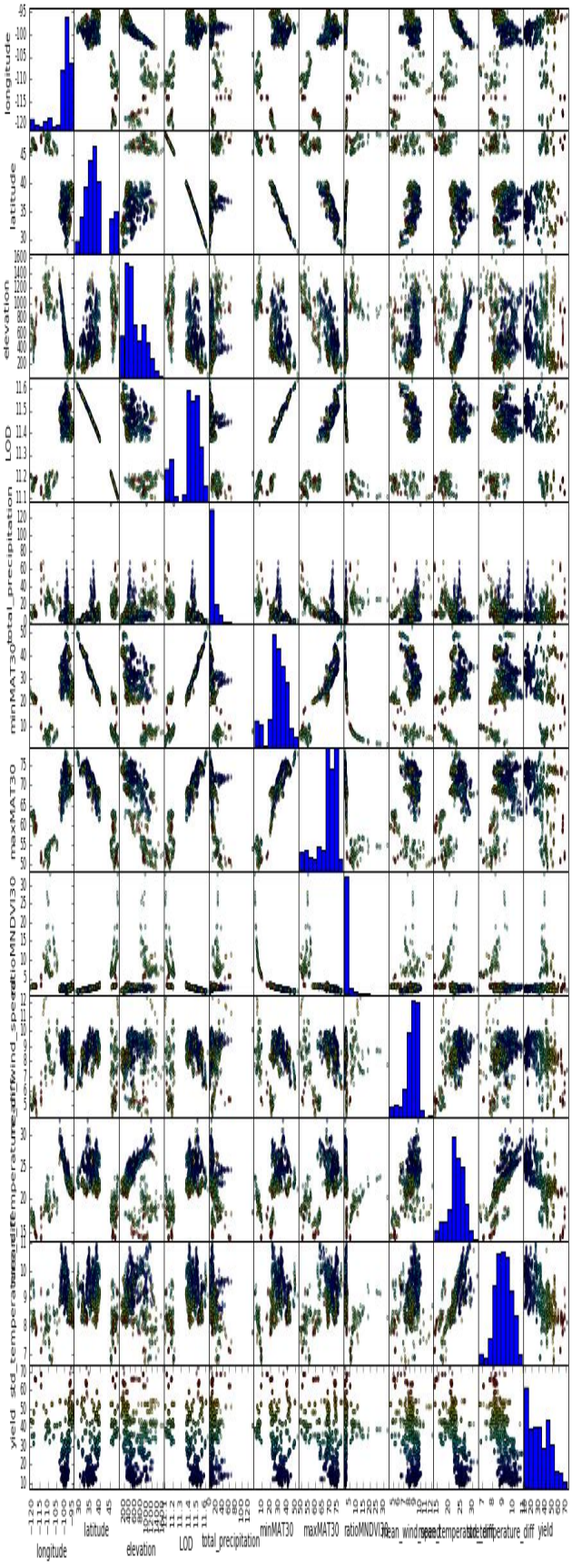

Fig.3. Correlation in the training set

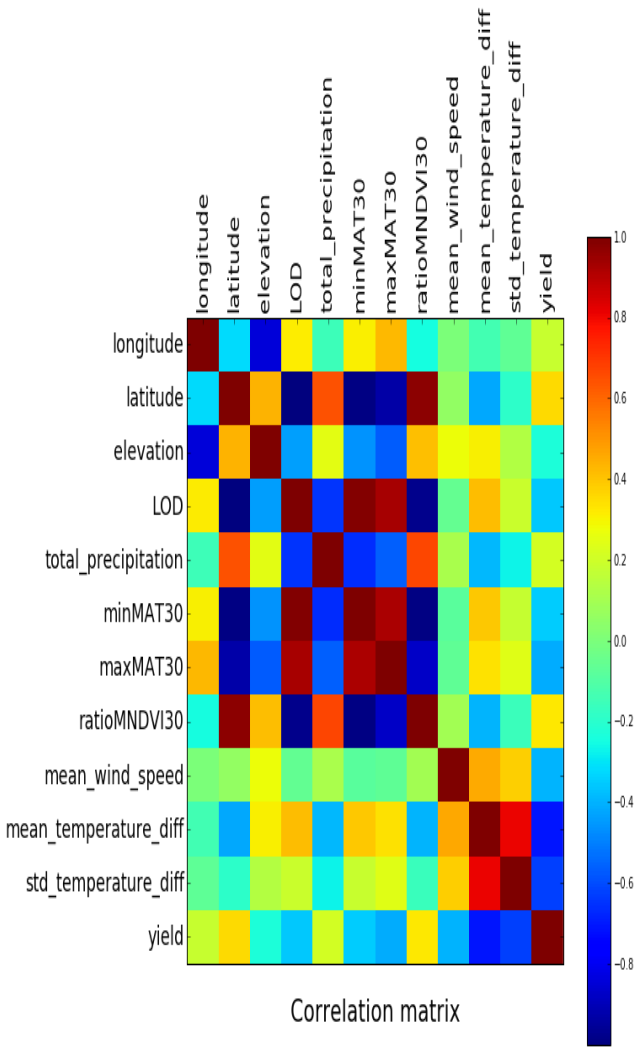

Fig.4. Correlation Matrix

V.CLASSIFICATION APPROACH FOR CROP YIELD PREDICTION

Classification is a supervised learning technique of machine learning which gives accurate results for predicting information such as yield prediction. Earlier crop prediction models used other classification techniques such as decision tree C4.5 which gave accuracy up to 76%. In our proposed work, we have used techniques such as K-NN classifier, Random Forest, Support Vector Machine, Gradient Boosted Decision Tree Regressor. The proposed model for the classification is shown in the below figure

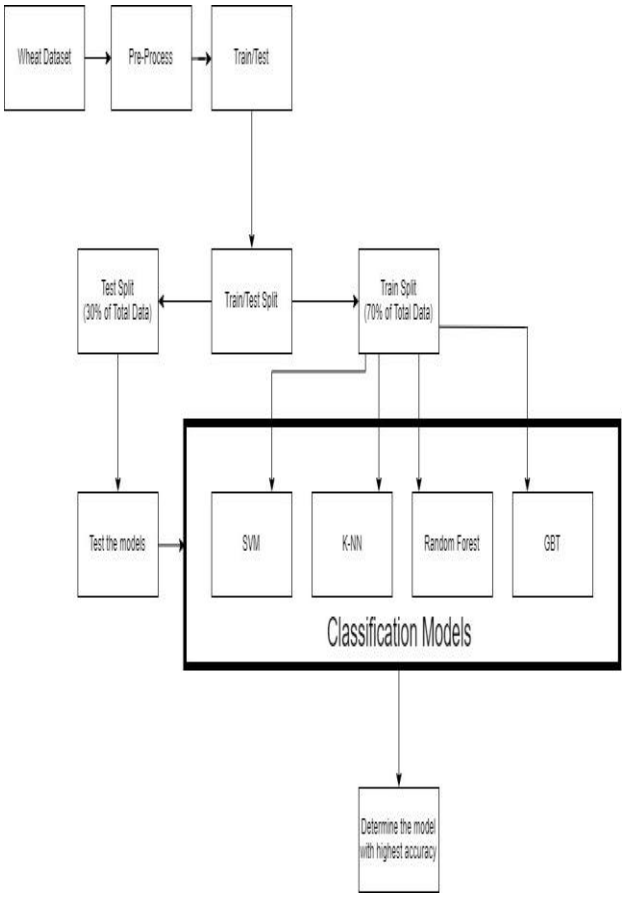

Fig.5 Proposed Model

VI. IMPLEMENTATION

Initially, the model is trained on a dataset. So that, based on this for a given input, we can predict the crop yield. In this model, the model is trained on 2017 dataset which has more fields and the crop yield is predicted for 2018. We have reserved 30% of the data for testing and the remaining 70% of the data for training. For classification, the algorithms that we have used are,

1. Support Vector Machine

In machine learning, support-vector machines (SVMs, moreover support-vector systems) are supervised learning models with related learning calculations that analyze data utilized for classification and regression analysis. It is a binary classification technique where it classifies the training instances into one of the two specified categories. SVMs can perform linear as well as non-linear classifications by using a technique known as kernel trick which implicitly maps inputs into high dimensional feature spaces.

SVM classifies inputs by constructing a hyperplane or a set of hyperplanes which act as a classification criteria. The equation of output of a linear SVM is $u = w.x - b$, where w is the normal vector of a hyperplane and x is the input vector

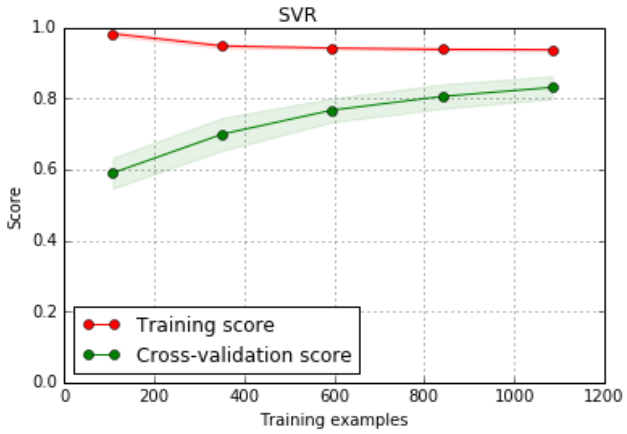

Fig. 6 Learning Curve for SVM

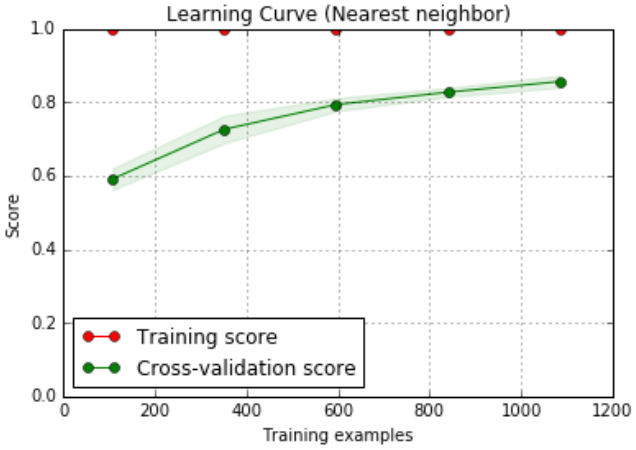

Fig. 8 Learning Curve for K-NN

Truth data

	Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Class 1	104721	72771	177492	59%
Class 2	72771	104721	177492	59%
Truth overall	177492	177492	354984	
User Accuracy (Recall)	59%	59%		

Fig.7. Confusion Matrix for SVM

Truth data

	Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Class 1	143769	33723	177492	81%
Class 2	33723	143769	177492	81%
Truth overall	177492	177492	354984	
User Accuracy (Recall)	81%	81%		

Fig. 9. Confusion Matrix for K-NN

2. K-Nearest Neighbor

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning and Data Mining. It is a supervised learning algorithm and has intense applications in pattern recognition, data mining and intrusion detection. In our model, K-NN has been used to predict crop yield. K-NN classifies the objects into classes. The object is mapped to the class most common among its k nearest neighbors. If k=1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. The distance to determine nearest neighbor is given by

$$\left(\sum_{i=1}^{k} \left(|x_i - y_i| \right)^q \right)^{1/q}$$

where q ranges from 0 to ∞ .

3. Random Forest

Random Forest is collection of multiple decision trees which is used for classification and regression. It constructs a multitude of decision trees during training randomly based on the input of that instance. The more the number of trees, the more robust the forest looks like. It gives the highest accuracy when there are maximum number of trees possible for the given instances. The chances of overfitting the model are less when there are more number of trees. It can also classify categorical variables. Basic parameters to Random Forest Classifier can be total number of trees to be generated and decision tree related parameters like minimum split, split criteria etc. The predictions for new samples can be done according to the equation

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

where x' is the denotation of new samples.

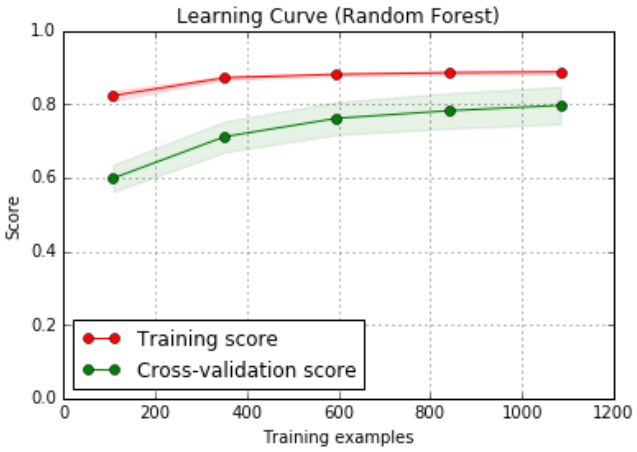

Fig. 10 Learning Curve for Random Forest

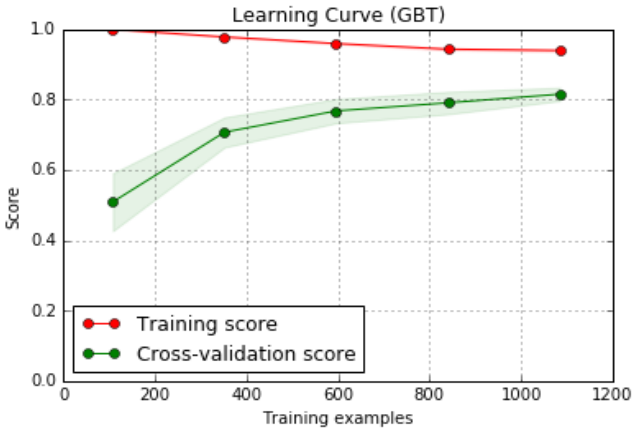

Fig.12. Learning Curve for GBT

Truth data

	Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Class 1	146431	31061	177492	82.5%
Class 2	31061	146431	177492	82.5%
Truth overall	177492	177492	354984	
User Accuracy (Recall)	82.5%	82.5%		

Fig. 11. Confusion Matrix for Random Forest

Truth data

	Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Class 1	148206	29286	177492	83.5%
Class 2	29286	148206	177492	83.5%
Truth overall	177492	177492	354984	
User Accuracy (Recall)	83.5%	83.5%		

Fig.13. Confusion Matrix for GBT

4. Gradient-Boosted Decision Tree Regressor

Unlike random forest which builds and binds a forest of random decision trees in parallel, gradient boosted decision trees build a series of trees. In the series, each tree is specifically trained to rectify the mistakes of the previous tree. Gradient boosted trees use a lot of shallow trees known as 'weak learners' in machine learning. It is built in a non-random way to create a model that makes lesser and lesser mistakes as more trees are added to the series. Making predictions through gradient-boosted trees is fast and doesn't use a lot of memory. The important factor which controls the complexity of model is the number of estimators. Another parameter known as learning rate determines the series of trees are built. High learning rate indicates that in a series a successor is putting strong emphasis on its predecessor. The main goal of gradient boosted decision trees is to keep creating new trees with minimum bias.

$$\begin{aligned} E\{(y-\hat{y})^2\} &= E\{(\epsilon + f(\mathbf{x}) - \hat{y})^2\} \\ &= E\{(\epsilon + f(\mathbf{x}) - E(\hat{y}) + -\hat{y} + E(\hat{y}))^2\} \\ &= \sigma^2 + (f(\mathbf{x}) - E(\hat{y}))^2 + \text{var}(\hat{y}) \\ &= \text{Irreducible error} + \text{bias}^2 + \text{variance} \end{aligned}$$

VII. RESULTS

After applying SVM, KNN, Random Forest and Gradient Boosted Decision Tree Regressor the following results are generated

Algorithm	SVM	K-NN	Random Forest	Gradient Boosted Tree Regressor
Accuracy	**59**	**81**	**82.3**	**83.5**

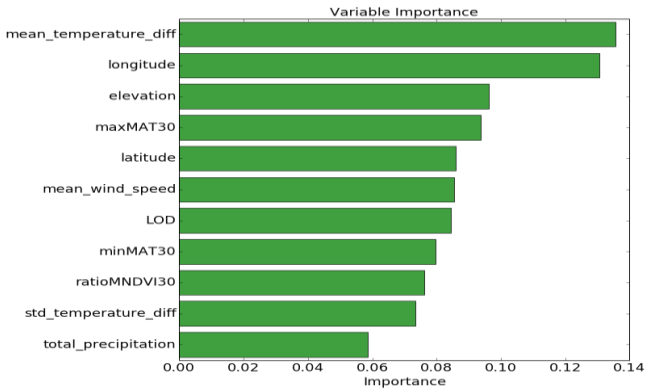

Fig.14. Features which made the most impact.

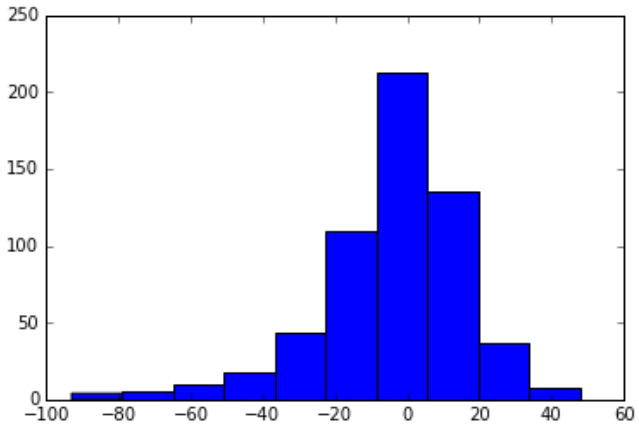

Fig.15. Histogram between test values and predicted values

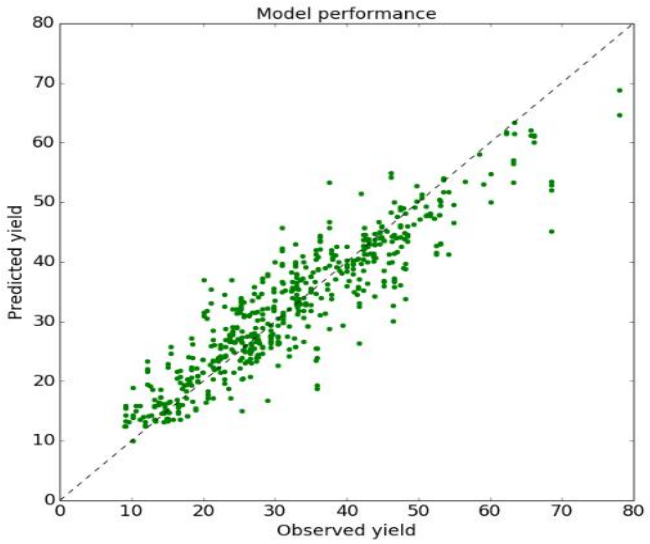

Fig.16. Model Performance for GBT

VIII. CONCLUSION

Yield of the crop can effectively be predicted by the classification models available in machine learning. We have applied four classification methods to predict the yield of the crop. The algorithms that we have used are SVM, K-Nearest Neighbor, Random Forest, Gradient Boosted Decision Tree Regressor. SVM has generated an accuracy of 59%, K-NN has generated an accuracy of 81% where as Random Forest has generated an accuracy of 82.5%. The model with most accuracy is generated by using Gradient Boosted Decision Tree Regressor (GBT) with an accuracy of 83.5%. By this we can conclude that GBT classifies and predicts crop yield better than any other models that we have used.

REFERENCES

1. Plant growth and soil moisture relationships, J. F Bierhuizen, Unesco/NS/AZ/476, Madrid Symposum Paper N0.37, Paris 1 September 1959.(https://unesdoc.unesco.org/ark:/48223/pf0000148851)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification And Regression Trees. Wadsworth.(https://books.google.co.in/books/about/Classification_and_regression_trees.html?id=uxPvAAAAMAAJ&redir_esc=y
3. Jun Wu, Anastasiya Olesnikova, Chi-Hwa Song, Won Don Lee (2009). The Development and Application of Decision Tree for Agriculture Data.(https://www.researchgate.net/publication/224382079_The_Development_and_Application_of_Decision_Tree_for_Agriculture_Data)
4. A Detailed Introduction to K-Nearest Neighbor (KNN) AlgorithmMay 17, 2010 by Saravanan Thirumuruganathan.
5. S. Veenadhari, Dr. Bharat Misra, Dr. CD Singh, Jan. 03-05, 2014, Machine learning approach for forecasting crop yield based on climatic parameters, Coimbatore,INDIA.(https://ieeexplore.ieee.org/document/6921718)
6. Miss.Snehal S. Dahikar, Dr.Sandeep V.Rode, January 2014, Agricultural crop yield prediction using artificial neural network approach, Vol.2 Issue 1.
7. D.Ramesh, B Vishnu Vardhan, January 2015, analysis of crop yield prediction using data mining techniques.
8. Talha Siddique, Dipro Baruna, Zannatul Ferdous, Amithaba Chakrabarty, 7-8 September 2017, Automated farming prediction.
9. Nanyang Ziu, Xu liu, Kai Hu, Ya Guo, July, 2018, Deep learning for smart agriculture.
10. Fabrizio Balducci, Donato Impedevo and Giuseppe Pirlo, 1 September 2018, Machine learning applications on agricultural datasets for smart farm enhancement.

AUTHORS PROFILE

R. Akshara, Assistant Professor, Department of Computer Science and Engineering, Vignan Institute of Technology and Science

Deepak Sanaka, B.Tech (CSE), Vignan Institute of Technology and Science

Anusha Vaspari, B.Tech (CSE), Vignan Institute of Technology and Science

J. Uday Kiran, B.Tech (CSE), Vignan Institute of Technology and Science