A Comparative Analysis of Machine Learning Prediction Techniques for Crop Yield Prediction in India

A.P.S Manideep[1], Dr. Seema Kharb[2]
[1]National Institute of Technology, Delhi
[2]SRM University, Sonipat, Haryana
[1]181210004@nitdelhi.ac.in
[2] seema.kharb@srmuniversity.ac.in

Abstract
Nowadays, crop yield prediction is one of the most recent, interesting and challenging tasks due to its dependence on various variable parameters like environmental, weather, soil and climate factors. Machine learning has become one of the important tools for predicting crop yield. This paper presents a machine learning framework for crop yield prediction using crop and weather data. It also compares the performance of potential machine learning methods like regression, decision trees, random forest, support vector machine and gradient boosting to forecast the yield of 80 crops in India for the year 2001 to 2016 using historical data. Furthermore, it has been observed from the results that the root mean square (RMSE) of the random forest method is 9433.7 for the dataset.
Keywords: Machine learning, Crop Yield Prediction, Regression.

1. Introduction

Agriculture is one of the most important and crucial areas of interest for society. It is also among one of the critical objectives for the United Nations and is a part of the objective that is developing food security and decreasing the hunger rate [1]. Therefore, crop yield prediction, crop protection and land estimation are among the significant issues related to food production. In India, a large population is dependent on agriculture for their living. So, yield prediction will benefit farmers and help them to make financial and management decisions.

The yield of a particular crop depends upon many factors like climate conditions, rainfall, temperature, information related to soil like pH, soil type etc, crop information, nutrients, disinfectants, water quality etc.[2]–[5].

Accurate prediction about the history of the crop yield is an important thing that can be used for making decisions related to agricultural risk management. The farmer will check the yield of the crop before actually cultivating on the field. This gives farmers an idea about what is going to happen for a particular crop in that year. This could benefit them to attain greater crop productivity if the conditions are suitable. It could also help them to decrease loss due to unsuitable conditions.

Currently, most of the researchers are using machine learning algorithms for modelling and validating the challenges in precise agriculture[6]–[11]. Machine learning is a great step for making the future of farming so bright. There are abundant technologies that reduce risk, improve sustainability, and place the grower in the centre of predictively informed decisions. Machine Learning is key behind all these technologies. Through the application of Machine Learning, a farmer can log into a customized dashboard on a computer or tablet and access all the data related to his crop and problems related to the crops and the yield and also effective solutions for handling the problem. Machine Learning gives the grower the information about his or her own operation that changes how they look at farming.

Further, Most of the modern farming techniques involve using robots that are designed and programmed to handle various aspects in agriculture. Their sensors help them to collect data

and analyze the problems and implement a solution for the problem. These robots are really farmer friendly and also work faster than a human labourer.
Therefore, the focus of this paper is to investigate the impact of machine learning algorithms for the prediction of crop yield in India.
The organization of this paper as follows. Section 2, discussed the related work. Section 3, provides the block diagram and the methods used followed by implementation details in section 4. The results obtained are discussed in section 5 which are followed by conclusion and future work.

2. Literature Review

The growth of artificial intelligence has open up new opportunities in advancement of agriculture framework [12]–[16]. Artificial neural network (ANN) is the most widely used machine learning algorithm for crop-yield prediction [17], [18] and agriculture planning [19].Some of the articles that have been studying crop yield prediction in India has been summarized here.
In [20], the authors have studied the impact of various climatic factors on crop yield in Madhya Pradesh, India. They have developed a software tool named "Crop Advisor" that uses C4.5 algorithm to identify the most influencing parameters. In [21] the authors have used SVM and Naïve Bayes algorithms to design two ensemble methods AdaSVM and AdaNaive for rice yield prediction in Tamil Nadu, India. In [22], the authors have studied and analysed that random forest is a better method for yield prediction. They have considered the dataset of wheat, maize and potato for thirty years in US. In [23] the authors have proposed an framework using SNM for crop selection and in [24] the authors have analysed the use of machine learning for crop yield prediction in Jammu, India using soil parameters. In [25] the authors have analysed the paddy yield using weather and soil parameters.
In [26], the authors analyse the performance of deep and machine learning models by considering soil condition and climate conditions to predict the yield of crop. In [27], the authors used traditional crop modelling with machine learning methods to create a generic crop yield forecaster. The authors in [28] have conducted their study in Karnataka, India and uses crop yield and weather data to predict crop yield using both machine and deep learning algorithms. In [29] a hybrid regression model using reinforcement and random forest algorithms have been proposed and in [30] random forest techniques has been used for cotton yield prediction in Maharashtra, India.
It has been observed that major machine learning algorithms that have been used for crop yield prediction are artificial neural network, support vector, machine, decision tree, random forest and regression. Secondly, apart from crop yield there are many other factors that affect crop yield. The most widely used parameters are soil parameters, climate parameters and solar parameters.

Since, artificial neural network is the most widely used machine learning algorithm, so this article aims at investigating the performance of artificial neural network with Linear Regression, Decision Tree, Random Forest, Gradient Boosting Regressor and Linear support vector regressor (SVR).

3. Proposed methodology

In this work machine learning is used to predict the yield of crops in India. So, it is a regression task, as we need to predict the yield of the crop which is a continuous value. Hence this is a supervised learning task. Further, there is no need to train the system on the go and the training instances can be fed to the system and let the system learn from them. Even though we need to update the system with new instances, there is no need to do that instantly. Hence this is a batch learning algorithm where the system is fed with all the

training instances available. In this work the system was fed with training instances and a model is made from them by generalizing the instances. When we want to predict the yield value for a new and unknown instance, the system uses the model to do so. Therefore, this can also be considered as model based learning.

The overall flow of work has been shown in the Figure 1and the details of each sub-module are discussed in the following subsections.

3.1. Dataset Description

Four datasets containing the data of crop yield, rainfall, temperature and pesticide use by 100 countries were used and are taken from World Bank, FOASTAT and OGD platform India[31]–[33]. This work focuses on India only, so the data for India was extracted from all the four datasets from the year 2001 to 2016. A new dataset was created by combining all the data related to India which were extracted from the actual datasets. This new dataset contains 1280 instances of 80 different crops for 16 years and has following features, i.e., Type of the crop, crop yield, Average rainfall per annum, average temperature per annum, and total amount of pesticides used in tonnes. This dataset is used as the main dataset for the project. The dataset consist of both numerical and categorical attributes.

3.2. Data Pre-processing

In data pre-processing the raw data is transformed into a useful and efficient format. Various transformation pipelines were developed to deal with the null values, numerical attributes and the categorical attributes. The main intention behind the creation of transformation pipelines is that they can be reused in the future and also they are very flexible. In this step, the dataset was checked for any null values (if present). Unfortunately, the dataset is well organized and no null values were found. A simple imputer is designed to deal with the null values which replace the null values (if present) with the median of the column in which they are present. Also a Standard scaler is designed for standardizing the numerical attributes and a one hot encoder was created to encode the categorical attributes. All these transformers are combined into a single pipeline which when given the dataset, deals with null values, standardization of numerical attributes and encoding of categorical attributes and gives us the dataset that is ready to apply on the model.

3.3. Train and Test Set Split

For this work, the dataset has 1280 training instances (or examples). Out of these 1280 instances, the dataset was divided into two sets namely: the training set and the testing set. The training set has 80% of the total instances (1024) and the testing set has the rest 20% (256). This is done so because the model should not overfit the data. *Overfitting* signifies that the model is trained more on the data it has and it may not perform well on the instances that are new or unknown to the data. This is why, after splitting the whole dataset into the training set and the testing set, the testing set is kept aside from the focus of the model, so that the model never knows about the instances in the testing set. Now the training set is used for further analysis and insights are derived from the training set to see which kind of model fits the data perfectly. After experimenting with the training set and the model is finalized, the testing set is used to evaluate the model on how it is performing with unknown data

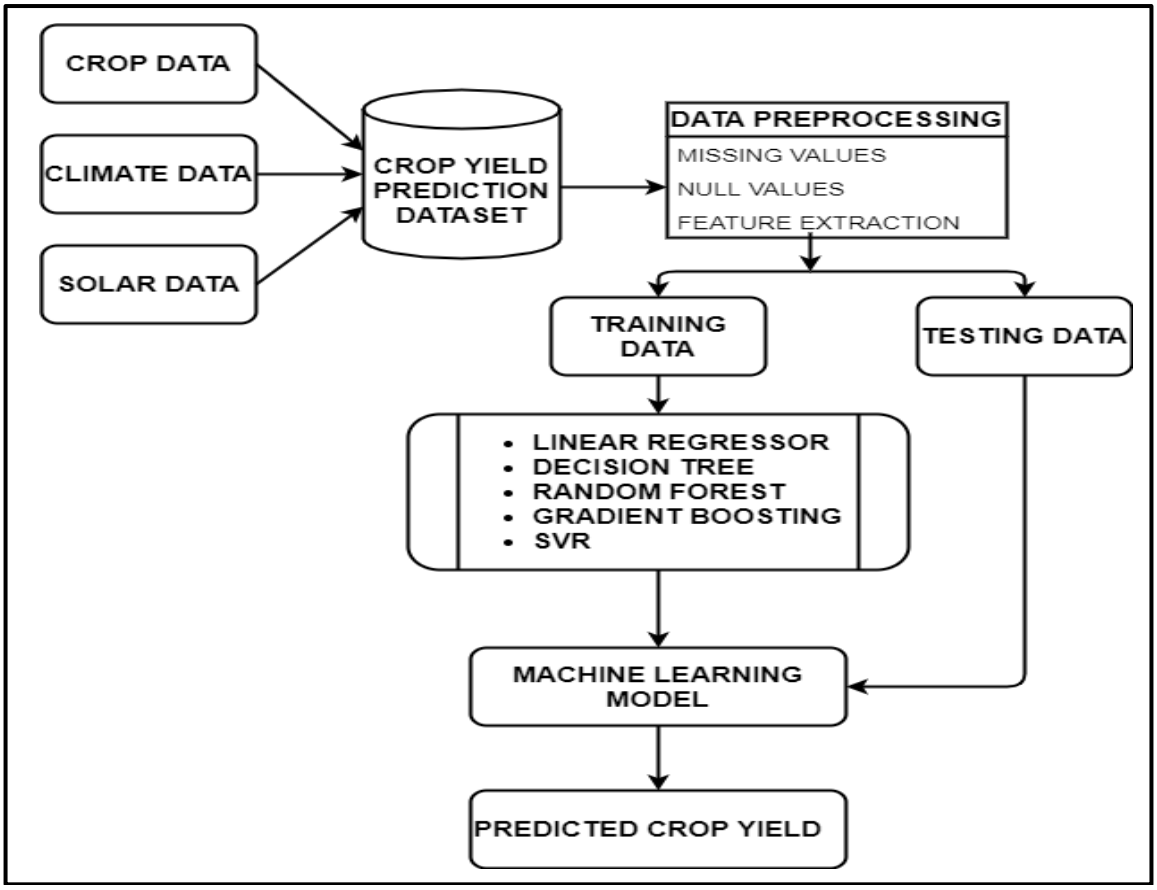


Figure 1: Block Diagram of Methodology

.

3.4. Train and Select the model

As crop yield prediction is a regression problem, different regression models such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor and Linear SVR have been used in this work. Every regression algorithm has its own way of solving regression tasks. These models have been described briefly in this sub-section.

Linear Regression. It is the linear machine learning model that assume linear relationship between the target vector and the input vector, under the conditions that target vector is continuous in nature and has a constant slope. It tries to fit a model which is a line and describes the relationship between the feature (input) and target vector. It is applicable if there is linear relation between target and feature vector, the data is normally distributed and features are independent to each other.

Decision Tree Regression. It is the predictive, tree based model that split the features based on the questions. The answer of these questions will be in the form of true and false. One of the condition to use decision tree is that the feature and target must be related (linearly or non-linearly).

*Random Forest Regression.*It is one of the most powerful and widely used ensemble learning method. It combines the concepts of bagging, in which the number of decision tress has been created by sample replacement, random selection of features and independency between the decision trees. If there is high dependency between the decision trees (correlation) then correspondingly the error rate for random forest is high. Another factor that

affect the performance of random forest is the strength of individual decision trees. An increase in the strength of trees result in lower error rate.

Gradient Boosting Regression. It is a predictive, additive ensemble method. One of the issues with weak learners is that there performance may vary even with a slight change in the data. So, boosting will help weak learners to filter out the samples that can be handled by them. In case of gradient boosting regression, the weak learner is decision tree and gradient descent method is used for loss minimization.

Linear Support Vector Regression. It is the generalization of support vector machine (SVM) for regression problems. SVR can be linear or non-linear based on the kernel function. In case of regression problem, in place of finding hyperplane, SVR find epsilon-insensitive region around the function that have at most epsilon-deviation, known as epsilon-tube.

The models were trained on the training set and evaluated using the cross validation technique to predict the crop yield.

3.5. Evaluating the model on test set

Choosing a model that performs well on the training data is not enough. The main aspect to consider while choosing a better model is how well the model generalizes, i.e., how well the model performs on instances that it has never seen. For this purpose all the models were evaluated using the testing set. As the testing set instances are completely new to the model, now the actual performance of each model can be examined.

4. Implementation details

This section describes the tool used and the performance parameters used for evaluation.

4.1. Tools Used

For the current project, Python is chosen as the programming language for all the implementations, starting from extracting the data, to evaluating the model. It has a huge library support for applications in the field of Machine Learning and Artificial Intelligence and this makes Python more suitable for solving problems in real world scenarios. Jupyter notebooks were used to write the code for the project. All the exploratory data analysis was done using Python libraries like NumPy, Pandas, Matplotlib, and Seaborn. The selection, training and evaluating the model was done using the Scikit-Learn library and its classes. No specific operating system is required as Python is a portable language.

4.2. Performance Parameters

The performance parameters used in this study are RMSE and R2 score.

As it is a supervised learning algorithm, the training instances have predefined labels. These labels are stored in the variable "y". A typical performance measure of regression tasks is "**Root mean square error (RMSE)**." It gives an idea of how much error the system typically makes in its predictions, with a higher weight for larger errors. It is a cost function measured on the set of instances using the model.

Typically, for an instance,

$error = y - \hat{y}$ = distance from the fitted regression line to the actual point

Where $y =$ actual point,

$\hat{y}$ = predicted point on the regression line.

$$MSE\ (Mean\ squared\ error) = \frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y_i})^2$$

Where, m = total number of instances in the dataset

$$RMSE\ (Root\ Mean\ squared\ error) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y_i})^2}$$

The more the RMSE value, the less efficient the model is.

One more evaluation metric named *R2 Score* is also used in this work which can be calculated as follows:

$$R2\ Score = R^2 = 1 - Relative\ Squared\ Error\ (RSE)$$

$$Relative\ Squared\ Error\ (RSE) = \frac{\sum_{i=1}^{m} (y_i - \hat{y_i})^2}{\sum_{i=1}^{m} \left(y_i - \underline{y}\right)^2}$$

Where $\underline{y}$ = mean value of all the y_i

R = Root of Relative squared error (also written as RRSE). It is a popular metric for accuracy of a model. It represents how close the data values are to the regression line. The higher the value of R^2 or (R), the better the model fits your data.

5. Results and Analysis

This section presents the results obtained by implementing the model discussed in the previous section. But before implementing the model we analyse the data thoroughly to get the insight related to the dataset. Therefore, this section firstly provides the statistical results and followed by the performance analysis of the proposed model.

5.1. Statistical Analysis

In order to check whether linear regression can be applicable to the current problem or not, statistical analysis has been performed. It has been performed to get an insight of the about the pattern of data. The results has been summarized in Table 1.

Table 1: Statistical Description the Training Dataset

	Year	Yield Value	avg_rain mm	avg_temp c	Pesticide Tonnes
count	1024	1024	1024	1024	1024
mean	2008.593	76198.419	86.955	24.709	41938.596
std	4.624	105374.091	7.668	0.244	11752.421
min	2001.000	1635.000	71.867	24.382	14485.330
25%	2005.000	10521.500	80.834	24.526	35342.000
50%	2009.000	28473.500	87.519	24.620	42482.560
75%	2013.000	96050.250	93.860	24.864	52980.000
max	2016.000	716343.000	100.849	25.239	60280.000

The Table 2shows the correlation matrix of various numerical attributes in the training set. It has been observed that the co-relation among input parameters is very low that is they are independent to each other.

Table 2: Correlation matrix for the training set

	Year	Yield Value	avg_rain mm	avg_temp_c	Pesticide Tonnes
Year	1	0.060909	0.44291	-0.012329	0.498088
Yield Value	0.060909	1	0.012935	0.004316	0.03566
avg_rain_mm	0.44291	0.012935	1	-0.126368	0.003161
avg_temp_c	-0.012329	0.004316	-0.126368	1	-0.196001
Pesticide Tonnes	0.498088	0.03566	0.003161	-0.196001	1

The following in Figure 2 heat map shows the correlation between the numerical attributes of the dataset.

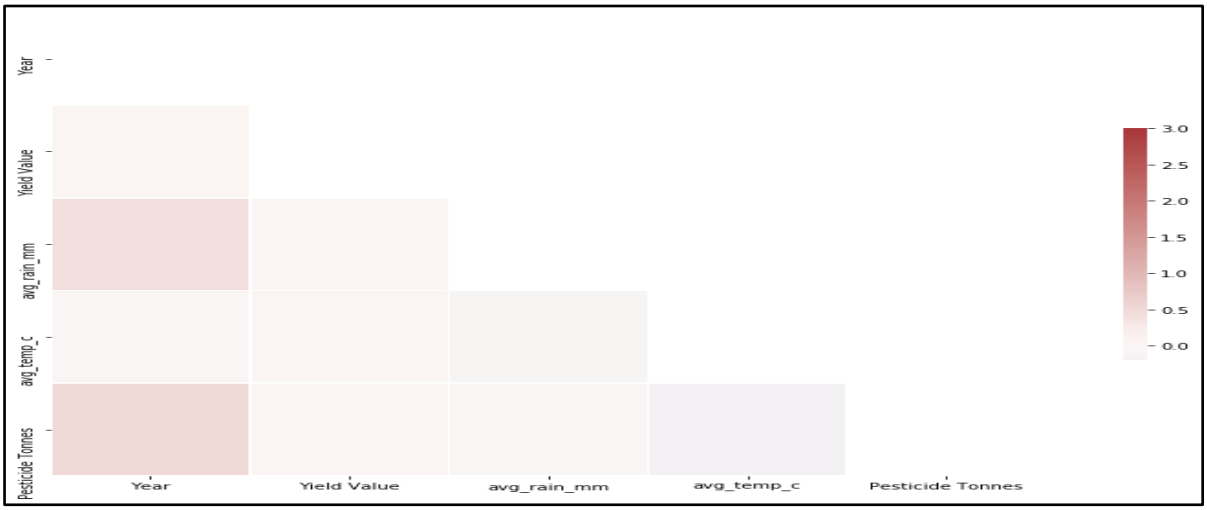


Figure 2: Heat map showing the correlation between numerical attributes of training set

In the heat map in Figure 2 shows that the intensity of the squares is not so high and from this we can infer that the linear relation between the 'Yield Value' and other numerical attributes is not enough to be considered as a dependency of one variable on the others.

To confirm this, we visualize the scatter matrix of the numerical attributes in the training set and plotted graphs between various numerical attributes with others. The **Figure 3** shows the plotted graph of the scatter matrix of the training set.

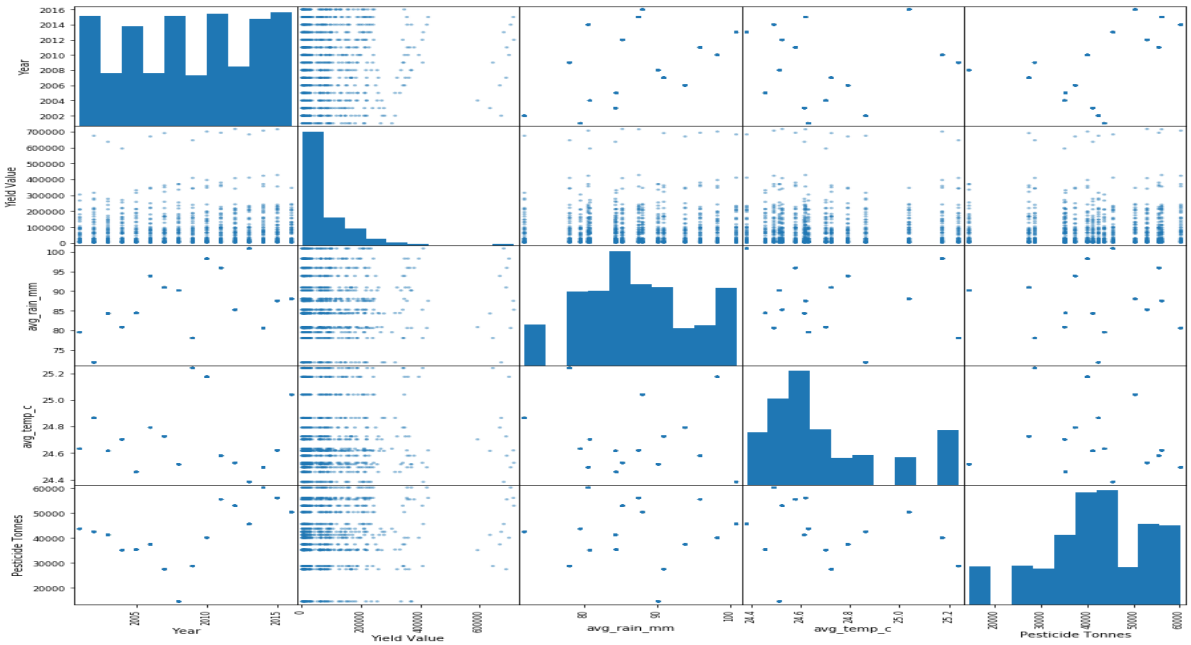
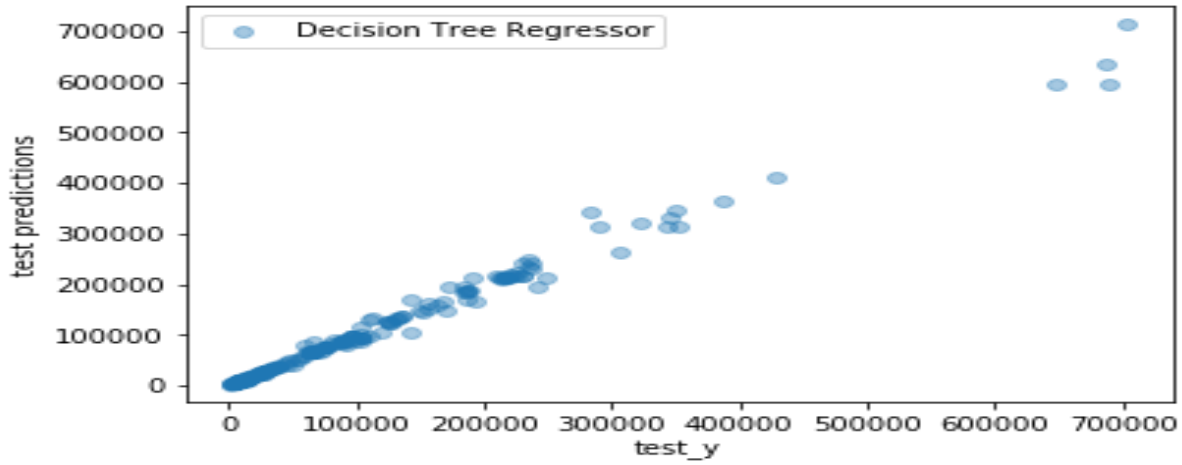


Figure 3: Scatter matrix of the Training Set

From the scatter matrix visualization, it has been observed that the plots of the 'Yield Value' vs various numerical attributes (2nd row in the above graph) are almost vertical in nature. This shows that the yield of the crop is not too linearly related with any of the other factors. So, a single attribute is not too important in predicting the crop yield but the linear combination might do the thing. Keeping this in mind, we can proceed for the next step of data preprocessing.

5.2. Plots of test_predictions vs test_y for the models on test set

The following plotsFigure 4



-Figure 8 show the predicted values vs actual values of the testing set by the models used in the project.

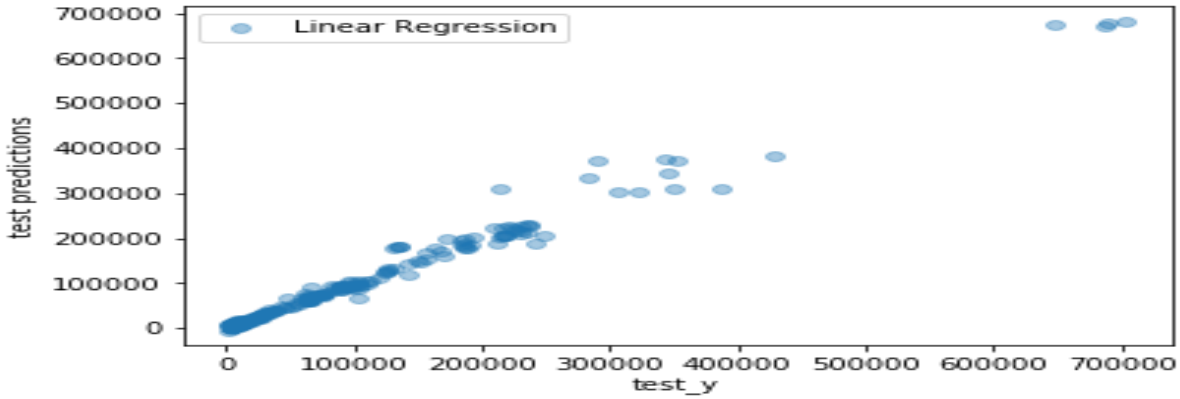


Figure 4: Linear Regression performance on Test Set

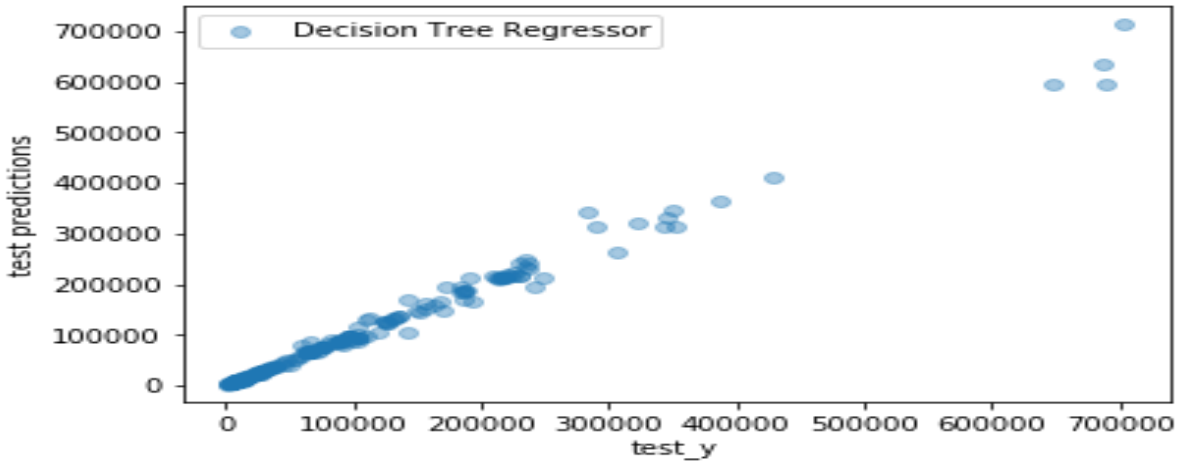


Figure 5: Decision Tree Regressor performance on test set

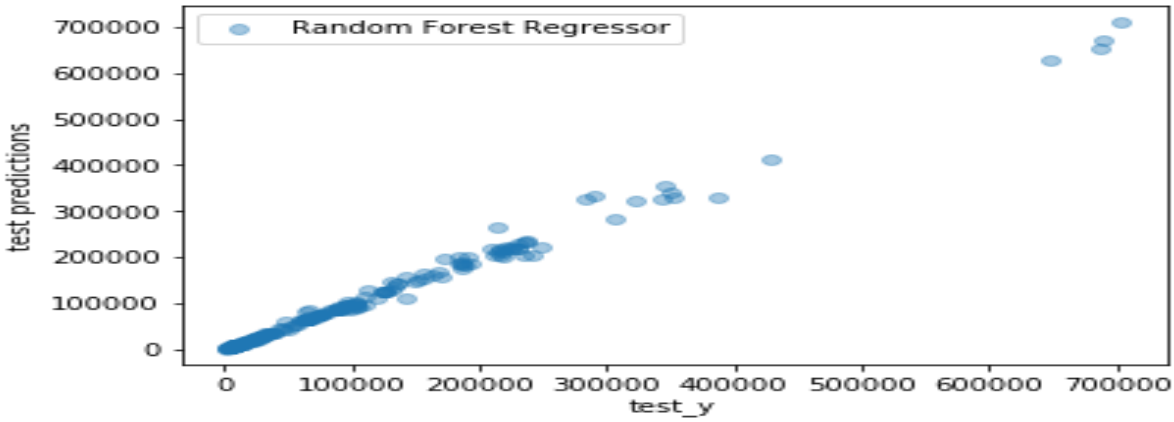


Figure 6: Random Forest Regressor performance on test set

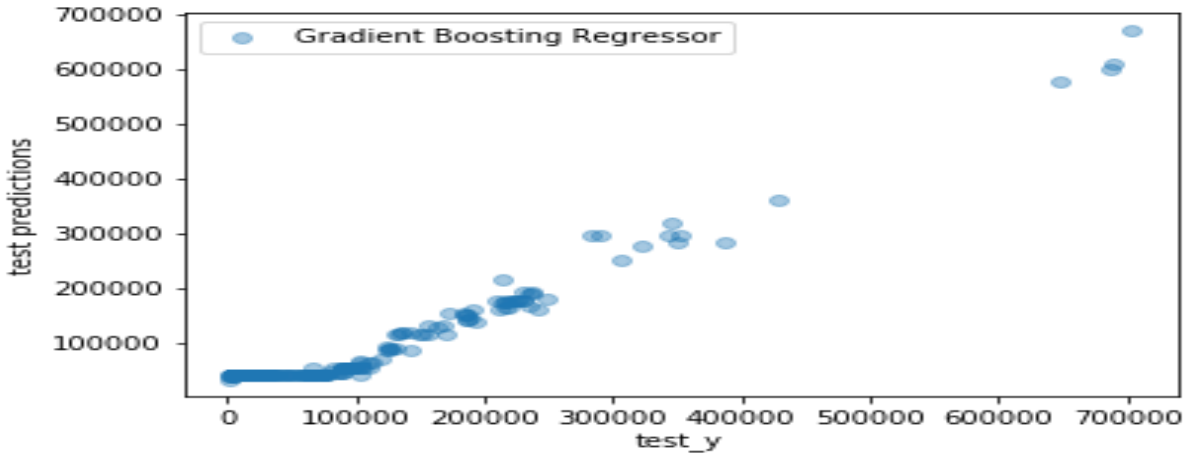


Figure 7: Gradient Boosting Regressor performance on test set

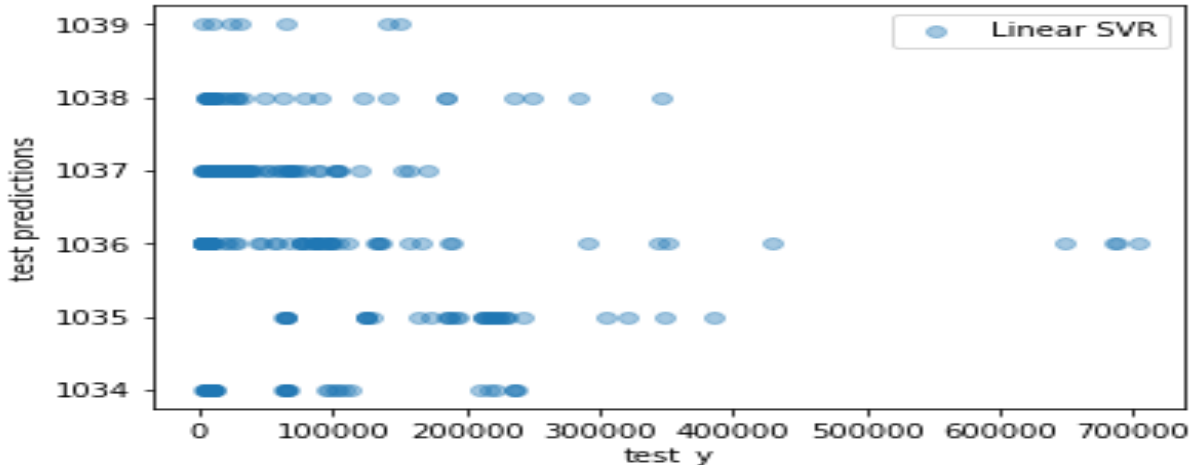


Figure 8: Linear SVR Performance on test set

All the predicted vs real plots show the predictions made by the models on the test set. It has been observed from the predicted vs actual plots that all the models have fitted well to the dataset and the error has been minimized. The Random Forest Regressor performed well on the testing set with an R2 score 0.993. The plot of the *test_predictions*vs *test_y*of Random Forest Regressor looks almost like a straight line which shows that it is a good fit. The plots of Decision Tree Regressor and Linear Regression also look similar but are slightly disoriented at some data points which shows the presence of outliers in the dataset. The Gradient Boosting Regressor didn't do well at the starting instances but as the training proceeds it eventually converges.It can be seen from theLinear SVR's graph that it is completely scatter and is not suitable for crop yield prediction may be non linear SVR may work better in this case.

The quantitative analysis of the evaluation parameters for training and testing phase has been summarised in Table 3 and Table 4 respectively. For an efficient model the RMSE value should be low as it is a measure of error while R2 score should be high. It can be observed that among all the models that were trained and experimented, the Random Forest Regressor shows better performance on the training set when compared to other models like Linear Regression, Gradient Boosting Regressor, Decision Tree Regressor, and Support Vector Regressor. Although the Decision Tree Regressorhas the value of R2 score as 1 which shows a perfect fit but it also has higher RMSE value. Hence, it can be infer that decision tree

regressor has been suffering from overfitting of the data. The Support Vector Regressorand the Gradient Boosting Regressor also did not perform well and has a high error. The basic Linear Regression model did well when compared to Gradient Boosting Regressor and Support Vector Regressor but Random Forest Regressor and the Decision Tree Regressor performed better than the Linear Regression in terms of RMSE and R2 scores.

Table 3: Evaluation scores on the training set

Model	RMSE	R2 Score	Remarks
Linear Regression	13901.388	0.985	Performed well but not well as compared to others
Decision Tree Regressor	13389.953	1.000	RMSE score was second least but the model has badly overfitted the training data
Random Forest Regressor	10370.288	0.998	Best performance with least RMSE and good R2 scores
Gradient Boosting Regressor	34645.196	0.905	Not a very good performance with low R2 score and high RMSE score
Support Vector Regressor	112629.648	-0.205	Bad performance with very high RMSE and negative R2 score

Finding the model that generalize well on the testing set is a very important part of a machine learning project. Out of all the models that were evaluated on the testing set, the Random Forest Regressor performed well on the testing set and could generalize well when compared to the other regression models used. The models performed the same as they did in the case of the training set. The Decision Tree Regressor, did not do better than the Random Forest Regressor but performed better than Linear Regression, Gradient Boosting Regression, and Linear Support Vector Regression.The Random Forest Regressor seemed to perform well in terms of both RMSE and R2 score. The Table 4 shows the scores of the models on the testing set.

Table 4: Evaluation scores on the testing set

Model	RMSE	R2 Score	Remarks
Linear Regression	14845.563	0.983	Performed well but not well as compared to others
Decision Tree Regressor	12041.631	0.989	RMSE score was second least but the model has badly overfitted the training data
Random Forest Regressor	9433.477	0.993	Best performance with least RMSE and good R2 scores
Gradient Boosting Regressor	35589.270	0.903	Not a very good performance with low R2 score and high RMSE score
Support Vector Regressor	143942.443	-0.572	Worst performance with very high RMSE and negative R2 score

6. Conclusion and Future Work

This work analyses the performance of machine learning models for crop yield prediction in India. For this the most commonly used machine learning algorithms have been implemented on the collected dataset and it has been observed that the although the Decision Tree Regressor is also a very good regressor but is a weak learner and is more prone to overfitting of data especially on training sets.

The Linear Regression is the basic regression algorithm which can be used only when the dataset is simple. It has been observed that that performance of linear regression has been affected by non-linearity of data, presence of outlier and high correlation among the features. Therefore, it may not work well for complex datasets. For this algorithms like Ridge Regression and Lasso Regression, which are more regularized forms of Linear Regression can be used when the dataset contains a large number of features or unwanted features.

The Random Forest Regressor performed well on both training and testing sets because of its non-linear and ensemble nature. Ensemble Learning is a type of machine learning strategy where a model that aggregates predictions of a group of predictors is chosen over a model with a best individual predictor. The group of predictors is called an ensemble and this technique is called ensemble learning. A Random Forest is an ensemble of Decision Trees. Despite its simplicity, Random Forest is the most powerful machine learning algorithm available today. A number of Decision Trees operate inside a random forest and an aggregate of all these Decision Trees make the Random Forest algorithm more efficient and powerful. All the decision trees inside the random forest can together detect more patterns in the data as compared to individual models that have only one predictor. One of the issue with this method is that it can't extrapolate and therefore is not able to handle the issue of covariate shift. The random forest in comparison to other techniques is more accurate, fast, simple, support penalization and provides a good estimate of error, correlation and strength.

As the future aspects, we will work to evaluate non-linear techniques and more advanced deep learning methods can be used as they have the advantage during feature selection phase in machine learning.

.

References

[1] S. Li, S. Peng, W. Chen, and X. Lu, "INCOME: Practical land monitoring in precision agriculture with sensor networks," *Comput. Commun.*, vol. 36, no. 4, pp. 459–467, 2013, doi: 10.1016/j.comcom.2012.10.011.

[2] A. D. Jones, F. M. Ngure, G. Pelto, and S. L. Young, "What are we assessing when we measure food security? A compendium and review of current metrics," *Advances in Nutrition*, vol. 4, no. 5. pp. 481–505, 2013, doi: 10.3945/an.113.004119.

[3] G. E. O. Ogutu, W. H. P. Franssen, I. Supit, P. Omondi, and R. W. A. Hutjes, "Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts," *Agric. For. Meteorol.*, vol. 250–251, pp. 243–261, 2018, doi: 10.1016/j.agrformet.2017.12.256.

[4] M. E. Holzman, F. Carmona, R. Rivas, and R. Niclòs, "Early assessment of crop yield from remotely sensed water stress and solar radiation data," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 297–308, 2018, doi: 10.1016/j.isprsjprs.2018.03.014.

[5] A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, "Machine Learning for High-Throughput Stress Phenotyping in Plants," *Trends Plant Sci.*, vol. 21, no. 2, pp. 110–124, 2016, doi: 10.1016/j.tplants.2015.10.015.

[6] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[7] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," *Comput. Electron. Agric.*, vol. 155, pp. 257–282, 2018, doi: 10.1016/j.compag.2018.10.024.

[8] T. U. Rehman, M. S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, "Current and future applications of statistical machine learning algorithms for agricultural machine vision systems," *Comput. Electron. Agric.*, vol. 156, pp. 585–605, 2019, doi: 10.1016/j.compag.2018.12.006.

[9] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Comput. Electron. Agric.*, vol. 121, pp. 57–65, 2016, doi: 10.1016/j.compag.2015.11.018.

[10] Y. Cai *et al.*, "A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach," *Remote Sens. Environ.*, vol. 210, pp. 35–47, 2018, doi: 10.1016/j.rse.2018.02.045.

[11] D. Elavarasan and P. M. Durairaj Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," *IEEE Access*, vol. 8, pp. 86886–86901, 2020, doi: 10.1109/ACCESS.2020.2992480.

[12] A. V. Samsonovich, "Socially emotional brain-inspired cognitive architecture framework for artificial intelligence," *Cogn. Syst. Res.*, vol. 60, pp. 57–76, 2020, doi: 10.1016/j.cogsys.2019.12.002.

[13] K. Ryan, P. Agrawal, and S. Franklin, "The pattern theory of self in artificial general intelligence: A theoretical framework for modeling self in biologically inspired cognitive architectures," *Cogn. Syst. Res.*, vol. 62, pp. 44–56, 2020, doi: 10.1016/j.cogsys.2019.09.018.

[14] R. Wason, "Deep learning: Evolution and expansion," *Cognitive Systems Research*, vol. 52. pp. 701–708, 2018, doi: 10.1016/j.cogsys.2018.08.023.

[15] X. Zhu, M. Zhu, and H. Ren, "Method of plant leaf recognition based on improved deep convolutional neural network," *Cogn. Syst. Res.*, vol. 52, pp. 223–233, 2018, doi: 10.1016/j.cogsys.2018.06.008.

[16] S. Zhang, W. Huang, and C. Zhang, "Three-channel convolutional neural networks for vegetable leaf disease recognition," *Cogn. Syst. Res.*, vol. 53, pp. 31–41, 2019, doi: 10.1016/j.cogsys.2018.04.006.

[17] K. Abrougui, K. Gabsi, B. Mercatoris, C. Khemis, R. Amami, and S. Chehaibi, "Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR)," *Soil Tillage Res.*, vol. 190, pp. 202–208, 2019, doi: 10.1016/j.still.2019.01.011.

[18] A. Haghverdi, R. A. Washington-Allen, and B. G. Leib, "Prediction of cotton lint yield from phenology of crop indices using artificial neural networks," *Comput. Electron. Agric.*, vol. 152, pp. 186–197, 2018, doi: 10.1016/j.compag.2018.07.021.

[19] J. Byakatonda, B. P. Parida, P. K. Kenabatho, and D. B. Moalafhi, "Influence of climate variability and length of rainy season on crop yields in semiarid Botswana," *Agric. For. Meteorol.*, vol. 248, pp. 130–144, 2018, doi: 10.1016/j.agrformet.2017.09.016.

[20] S. Veenadhari, B. Misra, and C. D. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters," Oct. 2014, doi: 10.1109/ICCCI.2014.6921718.

[21] N. Balakrishnan and G. Muthukumarasamy, "Crop Production - Ensemble Machine Learning Model for Prediction," *Int. J. Comput. Sci. Softw. Eng.*, vol. 5, no. 7, pp. 148–153, Jul. 2016, Accessed: Jun. 05, 2021. [Online]. Available:

https://www.proquest.com/openview/d678b09e570d574b3a0daebe036cc0d5/1?pq-origsite=gscholar&cbl=2044552.

[22] J. H. Jeong *et al.*, "Random forests for global and regional crop yield predictions," *PLoS One*, vol. 11, no. 6, Jun. 2016, doi: 10.1371/journal.pone.0156571.

[23] G. Suresh, D. A. S. Kumar, D. S. Lekashri, D. R. Manikandan, and C.-O. Head, "Efficient Crop Yield Recommendation System Using Machine Learning For Digital Farming," *Int. J. Mod. Agric.*, vol. 10, no. 1, p. 2021, Mar. 2021, Accessed: Jun. 05, 2021. [Online]. Available: http://modern-journals.com/index.php/ijma/article/view/688.

[24] V. Singh, A. Sarwar, and Sharma Vinod, "Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach.," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 1254–1259, Jun. 2017, Accessed: Jun. 05, 2021. [Online]. Available: https://web.b.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=09765697&AN=124636618&h=M3zi1D3lXDHUNOKK2NvE6ucsjRum607%2FZDT2zcLmlzqxbB4%2BaPnM4ucxLMD1ARVpDASuZAAo7og7eXwC4hvBZA%3D%3D&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3Fdirect%3Dtrue%26profile%3Dehost%26scope%3Dsite%26authtype%3Dcrawler%26jrnl%3D09765697%26AN%3D124636618.

[25] R. B. Guruprasad, K. Saurav, and S. Randhawa, "Machine Learning Methodologies for Paddy Yield Estimation in India: a Case Study," Nov. 2019, pp. 7254–7257, doi: 10.1109/igarss.2019.8900339.

[26] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," in *Journal of Physics: Conference Series*, 2021, vol. 1714, no. 1, doi: 10.1088/1742-6596/1714/1/012012.

[27] D. Paudel, H. Boogaard, A. de Wit, … S. J.-A., and undefined 2021, "Machine learning for large-scale crop yield forecasting," *Elsevier*, Accessed: Jun. 06, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308521X20308775.

[28] S. A. Shetty, T. Padmashree, B. M. Sagar, and N. K. Cauvery, "Performance Analysis on Machine Learning Algorithms with Deep Learning Model for Crop Yield Prediction," in *Springer*, 2021, pp. 739–750.

[29] D. Elavarasan and P. M. D. R. Vincent, "A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters," *J. Ambient Intell. Humaniz. Comput.*, 2021, doi: 10.1007/s12652-020-02752-y.

[30] N. R. Prasad, N. R. Patel, and A. Danodia, "Crop yield prediction in cotton for regional level using random forest approach," *Spat. Inf. Res.*, vol. 29, no. 2, pp. 195–206, 2021, doi: 10.1007/s41324-020-00346-6.

[31] "FAOSTAT." http://www.fao.org/faostat/en/#data/QC (accessed Jun. 06, 2021).

[32] Banco Mundial, "Indicators, Data," *World Development Indicators*, 2019. https://data.worldbank.org/indicator (accessed Jun. 06, 2021).

[33] Government of India (GOI), "Open Government Data (OGD) Platform India," *Website of Open Government Data (OGD) Platform India*, 2016. https://data.gov.in/ (accessed Jun. 06, 2021).