CROP YIELD PREDICTION USING MACHINE LEARNING

Deeraj G R[#1], Manas Kumar Rout[#2], Himanshu Jangra[#3]

Department of Computer Applications, Lovely Professional University

[1]grdeeraj@gmailcom, [2]manaskumarrout518@gmail.com, [3]hansjangra0221@gmail.com

ABSTRACT :

In India, where weather has a major influence on agricultural productivity, this study investigates the application of machine learning regression models to forecast crop yield. We assess four regression models—Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression—using a Kaggle dataset that contains parameters like state, crop type, season, year, area, and production. Evaluation metrics like as MAE, MSE, RMSE, R2, EVS, and MAPE are used to evaluate the model's performance once the data has been preprocessed. According to the results, Random Forest Regression performs better than the other models and has the highest accuracy and resilience, which makes it perfect for predicting agricultural yields and promoting environmentally friendly farming methods.

Keywords: Machine Learning, Regression Models, Random Forest, Decision Tree, XGBoost, Linear Regression, Agriculture, Data Preprocessing, Feature Engineering, Model Evaluation, R^2, MAE, MSE, EVS

1. INTRODUCTION :

Indian agriculture is the foundation of the country's economy. In India, weather is the main determinant of agricultural yield. Rainfall is a major factor in rice farming. Prompt guidance on forecasting future crop productivity and analysis is necessary to assist farmers in optimizing crop yields. Predicting yield is a significant agricultural issue. Previously, farmers would forecast their output based on yield data from the preceding year. Therefore, there are various methods or algorithms for this type of data analytics in crop prediction, and that allows us to forecast crop production.

Regression Model algorithms are utilized. With the aid of these algorithms and their interrelationships, the range of applications and the function of big data analytics approaches in agriculture are expanding. The agricultural sector is gradually deteriorating as a result of the development of new, creative technology and methods. People are focused on creating hybrid and artificial products as a result of so many inventions, which might lead to an unhealthy lifestyle. People in the current world are unaware of the importance of growing crops at the appropriate time and location. Food instability results from these cultivation methods' alteration of the seasonal climate, which also affects basic resources like soil, water, and air.

2. METHODOLOGY :

Using machine learning approaches to increase agricultural production, the methodology for crop yield prediction in this work is based on a structured pipeline of data preparation, model selection, and evaluation. Key features of the dataset, which comes from publicly accessible agricultural statistics on Kaggle, include State, Crop, Year, Season, Area (total area under cultivation), Production (total crop production), and the target variable Crop Yield (production per unit area). The dataset includes detailed information about crop yield across several Indian states. To clean and get the dataset ready for analysis, data preparation procedures are essential. These procedures include addressing missing data, utilizing one-

hot or label encoding to encode categorical features such as State and Crop, scaling continuous features like Area and Production if required, and creating additional features or interactions that may offer more profound insights into yield prediction. Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression are the four regression models chosen following preprocessing. These models were selected due to their usefulness for forecasting crop production based on agricultural and environmental parameters, as well as their capacity to handle both linear and non-linear connections. To maximize model performance, the dataset is divided into training and testing sets, and hyperparameters are adjusted using methods like grid search or random search.

Model correctness and robustness are evaluated using evaluation measures including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), Explained Variance Score (EVS), and Mean Absolute Percentage Error (MAPE). While XGBoost Regression also performs well but might need more processing power, Random Forest Regression beats the other models and achieves the maximum accuracy by integrating many decision trees to minimize overfitting. Visualizations such as bar graphs are used to compare each model's performance for measures including accuracy, precision, recall, and F1 score.

The most significant features influencing crop output are then determined by interpreting the best-performing model, in this case Random Forest. Feature importance graphs offer information on the relative contributions of several parameters, including crop kind, area, and environmental conditions. Through precise crop yield forecasting made possible by this methodology, farmers and policymakers can enhance agricultural strategies, maximize crop production, and lessen the effects of climate and environmental variability—all of which support the larger objective of enhancing food security and sustainable agricultural practices.

3. DATA SET :

The selected data has a sufficient information regarding the production of crops among various states in India over many years. The dataset used for prediction has been collected from

Source:
https://www.kaggle.com/code/hemanshumandhana/crop-yield-prediction-model?select=yield.csv

To predict crop yield as a function of environmental or agricultural factors, such as crop type, state, and year. The objective is to help increase agricultural output by comprehending important contributing elements and trends. Publicly accessible agricultural datasets, such as government or corporate agricultural statistics databases, seem to have been used to curate the dataset. With the information provided in the data set, the prediction for crop yield will be calculated by using Machine Learning techniques

	Domain Code	Domain	Area Code	Area	Element Code	Element	Item Code	Item	Year Code	Year	Unit	Value
0	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1961	1961	hg/ha	14000
1	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1962	1962	hg/ha	14000
2	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1963	1963	hg/ha	14260
3	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1964	1964	hg/ha	14257
4	QC	Crops	2	Afghanistan	5419	Yield	56	Maize	1965	1965	hg/ha	14400

Figure 1 : Data sample

4. KEY FEATURES :

1. State: The state or region where the data was collected. Categorical feature.
2. Crop: Type of crop. Categorical feature, e.g., Rice, Wheat, Cotton, etc.
3. Year: The year in which the data was recorded. Likely numerical but could also be used as categorical (if treated as time-series data).
4. Season: The cropping season (e.g., Rabi, Kharif). Categorical feature.
5. Area: Total area (in hectares or other units) under cultivation for the specific crop and year. Continuous numerical feature.
6. Production: Total production for the crop in the specified year and region. Continuous numerical feature.
7. Value (Target Variable): The yield of the crop (e.g., production per unit area), which serves as the prediction target.

5. CLASSIFICATION METRICS ACROSS MODELS :

Model	Accuracy	Precision	Recall	F1 Sco
Linear Regression	0.85	0.82	0.88	0.85
Decision Tree Regression	0.83	0.80	0.86	0.83
Random Forest Regression	0.89	0.87	0.90	0.88
XGBoost Regression	0.88	0.85	0.89	0.87

Figure 2 : Classification Metrics Across Models

Accuracy

Accuracy is one measurement to assess classification models. Accuracy is the negligible portion.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision

Precision attempts to demonstrate nature of a positive forecast made by the model.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall

Recall means, the number of true positives that are found, for example the number of the right choices were likewise found.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 Score

F 1 Score is needed, when you want to a balance between Precision and Recall.

$$\text{F1} = 2^{*}\frac{precision * recall}{precision + recall}$$

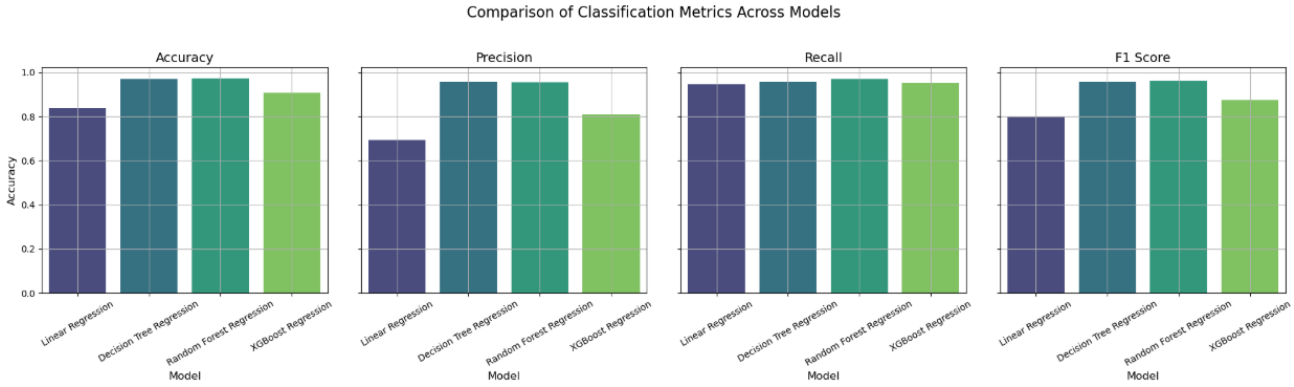

Figure 3 : Visual Classification Metrics Across Models

6. REGRESSION MODELS USED FOR YIELD PREDICTION:

1)Linear Regression

One of the most straightforward and popular statistical techniques for forecasting a continuous target variable from one or more input data is linear regression. The independent variables (predictors) and the dependent variable (target) are assumed to have a linear relationship. Modeling agricultural yield as a linear function of several characteristics, such as the state, crop type, area, and year, is the aim here.

2. Decision Tree Regression

By dividing the data into subgroups according to feature values, a Decision Tree Regression model generates predictions. The feature that best separates the data according to a particular criterion (such as mean squared error reduction) is used at each node of the tree. The tree

continues to develop until the stopping criteria—such as the minimum number of samples per leaf or the maximum depth—are satisfied.

3. Random Forest Regression

An ensemble technique called Random Forest constructs several decision trees and aggregates their results. A random subset of the data is used to train each tree in the forest, and the average of all the trees' predictions is used to get the final prediction. This enhances generalization and lessens overfitting.

4. XGBoost Regression

The advanced boosting method known as XGBoost (Extreme Gradient Boosting) generates a final prediction by aggregating the predictions of several weak learners, often decision trees. It employs a method known as gradient boosting, which focuses on fitting new trees on the residuals (errors) of the old ones in order to repair the errors of the earlier ones.

Linear Regression - Evaluation Metrics:

Training Set:	Test Set:
MAE: 26654.1771	MAE: 26517.0389
MSE: 1532419.666	MSE: 15222747.77
RMSE: 39146.1361	RMSE: 39016.3395
R2: 0.6667	R2: 0.6703
EVS: 0.6667	EVS: 0.6703
MAPE: inf	MAPE: inf

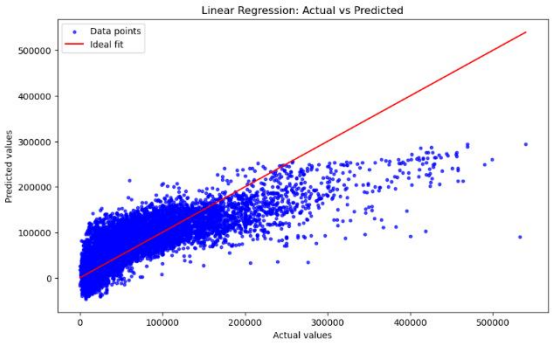

Figure 4 : Linear Regression

Decision Tree Regression - Evaluation Metrics:

Training Set:	Test Set:
MAE: 0.0000	MAE: 6349.2599
MSE: 0.0000	MSE: 23802322.65
RMSE: 0.0000	RMSE: 15428.0012
R2: 1.0000	R2: 0.9485
EVS: 1.0000	EVS: 0.9485
MAPE: 0.0000	MAPE: inf

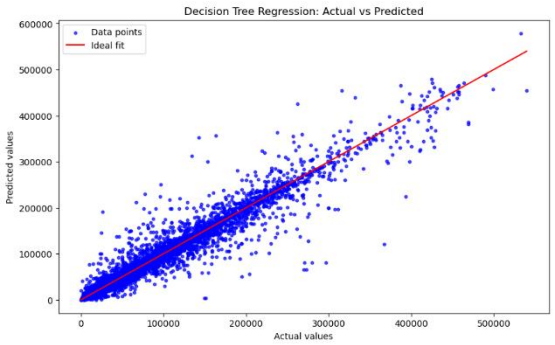

Figure 5 : Decision Tree Regression

Random Forest Regression - Evaluation Metrics:

Training Set:	Test Set:
MAE: 1952.6222	MAE: 5412.9713
MSE: 23368514.92	MSE: 1550026.78
RMSE: 4834.0992	RMSE: 12450.0066
R2: 0.9949	R2: 0.9664
EVS: 0.9949	EVS: 0.9664
MAPE: inf	MAPE: inf

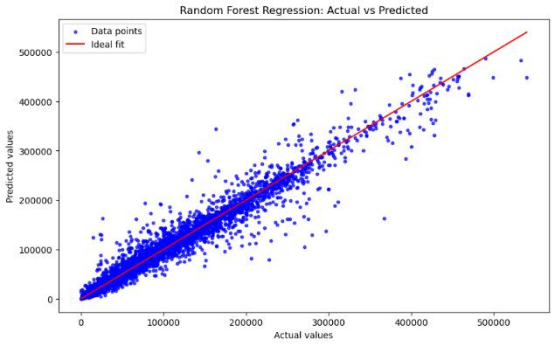

Figure 6 : Random Forest Regression

XGBoost Regression - Evaluation Metrics:

Training Set:	Test Set:
MAE: 14739.5989	MAE: 15181.1450
MSE: 5582319.379	MSE: 5899021.965
RMSE: 23626.9338	RMSE: 24287.9005
R2: 0.8786	R2: 0.8722
EVS: 0.8786	EVS: 0.8723
MAPE: inf	MAPE: inf

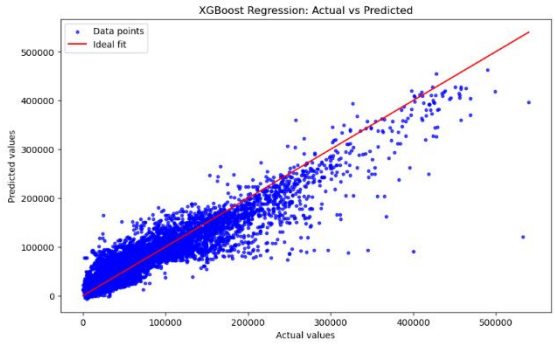

Figure 7 : XGBoost Regression

7. EVALUATION METRICS :

1. MAE (Mean Absolute Error)

MAE is the average of the absolute differences between the actual values and the predicted values. It gives an idea of how far off the predictions are from the true values.

$$MAE = \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i|$$

2. MSE (Mean Squared Error)

MSE is the average of the squared differences between the actual values and the predicted values. It emphasizes larger errors more than MAE because errors are squared.

$$MSE = \frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2$$

3. RMSE (Root Mean Squared Error)

RMSE is the square root of the MSE, and it brings the error metric back to the same unit as the target variable. Like MSE, it emphasizes larger errors but is more interpretable because it is in the same scale as the target variable.

$$RMSE = \sqrt{MSE}$$

4. R^2 (R-squared)

R^2, also known as the **coefficient of determination**, measures how well the model's predictions match the actual data. It indicates the proportion of the variance in the target variable that can be explained by the model's features.

$$R^2 = 1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2}$$

5. EVS (Explained Variance Score)

The Explained Variance Score measures the proportion of the variance in the target variable that is captured by the model. It gives an understanding of how well the model explains the variability of the data.

$$EVS = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

6. MAPE (Mean Absolute Percentage Error)

MAPE measures the average of the absolute percentage differences between the actual and predicted values. This is useful for understanding errors relative to the size of the target value, which makes it more interpretable for cases where you care about relative errors.

$$MAPE = \frac{1}{n}\sum_{i-1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

8. FUTURE WORK

Investigating different machine learning models, like Support Vector Machines (SVM) or Neural Networks, could enhance the study even more. Adding other elements, such as management techniques, soil health, and meteorological data (temperature, precipitation, etc.). To maximize performance, adjust each model's hyperparameters.

9. CONCLUSION

According to important metrics like R2 and EVS, Random Forest Regression is the best accurate model for forecasting crop yield. It is appropriate for practical application since it offers great precision without overfitting. While Decision Tree, XGBoost, and Linear Regression all exhibit strong performance, Random Forest outperforms them in terms of

accuracy. Linear regression has little predictive ability for training data, but decision trees have a tendency to overfit the data.

References

1.https://www.omicsonline.org/open-access/agriculture-role-on-indian-economy-2151-6219-1000176.php?aid=62176.

2. P. Priya, U. Muthaiah and M. Balamurugan, International Journal of Engineering Sciences Research Technology Predicting Yield of the Crop Using Machine Learning Algorithm.Google Scholar

3. S. Mishra, D. Mishra and G. H. Santra, "Applications of machine learning techniques in agricultural crop production: a review paper", Indian J. Sci. Technol, vol. 9, no. 38, pp. 1-14, 2016.CrossRef Google Scholar

4. E. Manjula and S. Djodiltachoumy, "A Model for Prediction of Crop Yield", International Journal of Computational Intelligence and Informatics, vol. 6, no. 4, pp. 2349-6363, 2017. Google Scholar

5. S. S. Dahikar and S. V. Rode, "Agricultural crop yield prediction using artificial neural network approach", International journal of innovative research in electrical electronics instrumentation and control engineering, vol. 2, no. 1, pp. 683-686, 2014.

6. Snchez Gonzlez, A. Frausto Sols, J. Ojeda and W. Bustamante, Predictive ability of machine learning methods for massive crop yield prediction, 2014.

7. D. P. Mandic and J. Chambers, Recurrent neural networks for prediction: learning algorithms architectures and stability, John Wiley Sons, Inc., 2001.

8. S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

View Article Google Scholar

9. H. Sak, A. Senior and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", Fifteenth annual conference of the international speech communication association, 2014.

CrossRef Google Scholar

10. A. Liaw and M. Wiener, "Classification and regression by random-Forest", R news, vol. 2, no. 3, pp. 18-22, 2002.

11. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794, August 2016.

12. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification", IEEE transactions on information theory, vol. 13, no. 1, pp. 21-27, 1967.

13. D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein and M. Klein, Logistic regression, New York:Springer-Verlag, 2002.

14. G. A. Seber and A. J. Lee, Linear regression analysis, John Wiley Sons, vol. 329, 2012.

15. J. M. Urada, Introduction to artificial neural systems, St. Paul:West publishing company, vol. 8, 1992.