Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR CROP YIELD PREDICTION

Ritesh Patil, M.Sc. Statistics & Data Science, Nilkamal School of Mathematics, Applied Statistics & Analytics, NMIMS (Mumbai), Maharashtra, 400056, India.
Vishakha Pawar, M.Sc. Statistics & Data Science, Nilkamal School of Mathematics, Applied Statistics & Analytics, NMIMS (Mumbai), Maharashtra, 400056, India.
Nikhil K. Pawanikar, Assistant Professor, Department of Computer Science, Nilkamal School of Mathematics, Applied Statistics & Analytics, NMIMS (Mumbai), Maharashtra, 400056, India.

ABSTRACT
Agriculture is one of the pillars of the Indian economy and accurate estimation of crop yield is important for planning resources and financial management. This research compares the performance of five machine learning algorithms Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and XGBoost on historical Indian district agriculture data on 15 Kharif crops for the period 1966 to 2017. Model performance was assessed in R^2 Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and a normalized combined score. Random Forest performed consistently better than other models, with better accuracy, stability, and generalizability. XGBoost was also strongly predictive, whereas KNN was moderately stable but could not handle high-dimensional data. Decision Tree had decent MAE scores but tended to overfit, and Linear Regression had limited performance for nonlinear agricultural data sets. These results highlight the potential of ensemble-based models, especially Random Forest, in improving crop output forecast and aiding AI-based farming loan eligibility systems

Keywords:
Crop Yield Prediction, Machine Learning, Random Forest, XGBoost, Decision Tree, KNN, Linear Regression, Model Comparison.

I. Introduction

Agriculture continues to be the mainstay of income for a large percentage of India's population, especially in rural areas. One of the ongoing issues in the sector is reliably forecasting crop which yields, are determined by a such multifaceted interaction of variables such as rainfall, temperature, soil type and fertilizer application. These changing conditions frequently constrain decision-making in a timely and informed manner for farmers and policymakers. Recent machine learning progress provides promising answers by allowing analysis of big agricultural and weather data sets to reveal concealed patterns and make credible predictions. In this paper, we utilize five machine learning algorithms Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and XGBoost to forecast the yields of 15 key Kharif crops such as rice, maize, soybean, and cotton. These crops are grown mainly during the monsoon season and are vital to Indian food security and rural economy.
The data set consists of historical weather and crop yield data gathered from different districts throughout India. Model performance was determined in terms of prediction accuracy and reliability. Our findings show that ensemble-based models, especially Random Forest, provide better performance, demonstrating the promise of machine learning to enable data-driven agricultural planning and financial systems.

II. Literature

Over the past decade, numerous studies have been conducted on applying machine learning (ML) techniques to predict crop yield and suggest crops. All these attempts are geared towards improving agricultural production and assisting farmers in making fact-based choices by evaluating a broad spectrum of environmental, climatic, and soil parameters.

Nigam et al. (2019) [1] explored multiple machine learning models including LSTM, simple RNN, Random Forest, and XGBoost to predict crop yield based on factors such as temperature, rainfall, and crop area. Their study emphasized the effectiveness of ensemble learning techniques, noting that Random Forest performed particularly well in modeling non-linear dependencies in agricultural datasets. Swathi and Sudha (2023) [2] proposed a soil nutrition-based crop classification model using various machine learning algorithms including Decision Tree, SVM, Naive Bayes, KNN, XGBoost, and Random Forest. Their analysis revealed that Extreme Gradient Boosting and Naive Bayes achieved the highest performance, with AUC scores of 0.994 and 0.993 respectively. Jambekar et al. (2018) [3] applied data mining techniques—specifically Multiple Linear Regression, Random Forest Regression, and Multivariate Adaptive Regression Splines (Earth)—to predict the production of rice, wheat, and maize in India. The study concluded that Earth outperformed the other models in terms of mean squared error and R^2 score, especially for rice and wheat datasets. Parameswari et al. (2021) [4] implemented crop recommendation models using rule-based classifiers like PART, JRip, and Decision Table. Their results showed that the PART algorithm achieved the highest precision (98.33%) and was computationally efficient, suggesting its viability in real-time agricultural decision support systems. Sangeetha and Shruthi (2020) [5] developed a comprehensive crop yield prediction system incorporating rainfall, pH value, and nutrient content. Their model focused not only on predicting yields but also on recommending optimal fertilizer usage. This dual-purpose approach enhanced yield outcomes and soil health management practices. Nishant et al. (2020) [6] introduced a novel crop yield prediction model utilizing advanced regression techniques such as Kernel Ridge, Lasso, and ENet. Their work also employed stacking regression to boost model performance and reduce prediction errors. The use of easily available parameters like district, season, and crop type made the system more user-friendly for farmers. Bhanumathi et al. (2019) [7] proposed a system that integrated soil and climate data to both predict crop yield and recommend fertilizer ratios. Their use of machine learning models, particularly Random Forest and ANN, provided actionable insights to farmers and supported better crop management decisions. Pandith et al. (2020) [8] evaluated multiple supervised learning models including K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Naïve Bayes, Logistic Regression, and Random Forest for mustard crop yield prediction using soil nutrient data from Jammu. They concluded that KNN and ANN provided the most accurate results, making them suitable for soil-based yield estimation tasks. Abdel-salam et al. (2024) [9] proposed a hybrid feature selection approach combined with an optimized Support Vector Regression (SVR) model. They used clustering (K-means) and a correlation-based filter (CFS), followed by recursive feature elimination (RFE) and an Improved Crayfish Optimization Algorithm (ICOA) for hyperparameter tuning. Their framework significantly enhanced accuracy and computational efficiency compared to traditional methods. Sujatha and Devi (2016) [10] explored crop yield forecasting using classification algorithms such as Naïve Bayes and J48. Their study utilized historical crop and climate data, and highlighted the importance of data preprocessing and attribute selection in improving the forecasting performance of classification models. Kamath et al. (2021) [11] employed Random Forest for yield forecasting using soil and weather attributes, demonstrating its superiority over other regression models including Multivariate Adaptive Regression Splines and Multiple Linear Regression. They concluded that Random Forest offers better precision in region-specific yield forecasting scenarios. Modi et al. (2021) [12] designed an SVM-based crop recommendation system using soil parameters such as N, P, K, and pH. Their system, implemented in Anaconda Navigator, aimed to maximize crop profitability and minimize farmer losses through accurate soil classification and prediction. Padmavathi et al. (2024) [13] conducted a comparative analysis of various ensemble learning algorithms Random Forest, Gradient Boost, XGBoost, AdaBoost, LightGBM, and CatBoost for both crop recommendation and yield prediction. They found that boosting methods, particularly XGBoost and LightGBM, performed well in high-dimensional and imbalanced datasets, offering superior accuracy and interpretability. Veenadhari (2011) [14] presented a broad review of data mining techniques in agriculture, emphasizing algorithms like ID3, k-means, k-NN, and SVM. The review underscored the effectiveness

of decision trees and neural networks in modeling non-linear relationships in crop productivity datasets. Maya Gopal and Bhargavi (2019) [15] conducted a comparative evaluation of ML algorithms, including Random Forest (RF), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN), for crop yield prediction. Their study underscored the importance of identifying optimal feature subsets, revealing that RF consistently delivered the highest accuracy across multiple feature configurations. Iniyan et al. (2023) [16] explored various regression techniques, including Lasso, Ridge, and Gradient Boosting, for predicting crop yields in Maharashtra. Their findings suggested that Gradient Boosting consistently outperformed other models in accuracy and robustness across varying datasets. Similarly, Jhajharia et al. (2023) [17] demonstrated the effectiveness of deep learning models, particularly Long Short-Term Memory (LSTM) networks, alongside classical ML algorithms like RF and Gradient Descent, for predicting crop yields in Rajasthan. RF emerged as the most accurate model, achieving an R^2 of 0.963. Nikhil et al. (2024) [18] evaluated the performance of tree-based models, including Extra Trees Regressor and RF, in predicting crop yields across South Indian states. Their results highlighted the superior predictive capability of Extra Trees Regressor, achieving an R^2 of 0.9615. Gosai et al. (2021) [19] proposed a crop recommendation system using ML algorithms such as Decision Trees, Naïve Bayes, and RF. Their system integrated IoT-based soil monitoring with ML models, providing actionable recommendations for optimal crop selection. Shook et al. (2021) [20] proposed a deep learning framework incorporating LSTM networks enhanced with temporal attention mechanisms to predict soybean yield across 28 U.S. states and Canadian provinces. Their approach effectively integrated genotype-relatedness measures with 30 weeks of multivariate weather data, outperforming baseline models such as SVR-RBF and LASSO in both accuracy and interpretability. Dey et al. (2024) [21] evaluated five ML algorithms XGBoost, SVM, Random Forest, KNN, and Decision Tree on a diverse dataset comprising NPK levels, pH, and climatic variables for both agricultural and horticultural crops. XGBoost achieved the best accuracy of prediction (up to 99.3% AUC) among crop categories, affirming its better capability to handle imbalanced and heterogenous agricultural data. Shahhosseini et al. (2021) [22] investigated hybrid models combining crop simulation (APSIM) outputs with machine learning algorithms like LightGBM, XGBoost, and Random Forest. Their hybrid APSIM+ML approach improved corn yield prediction accuracy in the US Corn Belt by 7–20%, with soil moisture and drought stress indicators identified as the most influential features for ML performance.

Even with advancements, issues still persist in handling imbalanced datasets, real-time environmental heterogeneity, crop variety and regional applicability. Many research studies propose the fusion of IoT and remote sensing data and embracing AI methods for dynamic yield estimation and crop suggestion. However, the potential of mapping ML with geospatial analytics and weather forecasting is an area that can be explored further.

Objectives

1. To develop machine learning models for forecasting the yield of 15 prominent Kharif crops based on historical agricultural and climatic data from Indian districts.
2. To analyse the predictability of five machine learning algorithms: Linear Regression, Decision Tree, KNN, Random Forest and XGBoost.
3. To compare the models in terms of regression metrics like R^2, RMSE, and MAE to measure accuracy, consistency and generalizability.

Data Description

The data employed in this research was obtained from ICRISAT (International Crops Research Institute For The Semi-Arid Tropics) and comprises 16,032 entries gathered from 311 districts of 20 Indian states between the years 1966 and 2017. It is centered around 15 Kharif crops and provides information on crop yields (kg/ha), climatic variables (rainfall, temperature, precipitation), fertilizer use (nitrogen, phosphate, potash). Geographical features like state and district names, year wise data

render the dataset appropriate for regional and time-series based analysis in agricultural yield forecasting.

III. Methodology

This paper will develop and compare various machine learning models for predicting crop yields with structured agricultural data. The general methodology is broken down into several stages, such as data preprocessing, feature engineering, model choosing, training, evaluation, and comparison. Special emphasis is laid on knowing how various machine learning models achieve yields of 15 main Kharif crops and predict the across different districts states of India.

3.1. Data Preprocessing and Feature Engineering

Data preprocessing step to confirm the quality and dataset consistency matters of the before the application of machine learning algorithms. The data consists of both numerical and categorical features, as well as target variables for crop yields.

1.1. Dealing with Categorical Variables: The data contains categorical variables like state name, district name, and type of soil. One-hot encoding was employed to convert these into binary vectors. The method works well for Decision Trees and Random Forest algorithms that can efficiently deal with sparse matrices.

1.2. Scaling of Numerical Features: Most machine learning algorithms, especially K-Nearest Neighbors (KNN) and Linear Regression, are scale-sensitive for the input data. Thus, numerical features like rainfall, temperature, and fertilizer application (nitrogen, phosphate, potash) were scaled using StandardScaler, which rescales the data to have mean 0 and standard deviation 1.

1.3. Log Transformation of Target Variables: Values of crop yield were right-skewed. To meet this and stabilize the variance, a log1p transformation (log(1+ x)) was used for the target variables. This assists in enhancing model performance as well as minimizing the influence of outliers.

1.4. Train-Test Split: Preprocessed data was divided into 70% training set and 30% test set. This provided enough data for the models to learn patterns while training yet enabling accurate assessment on unseen test data.

3.2 Feature and Target Variables

The independent variables (features) used for training the models included:

2.1 Temporal and Geographic Variables: Year, State Name, District Name

2.2 Environmental Variables: Average Rainfall (mm), Average Temperature (°C), Average Precipitation (mm)

2.3 Agricultural Inputs: Nitrogen, Phosphate, and Potash usage per hectare

2.4 Soil Characteristics: Soil type percentage (categorical)

The dependent variables (targets) were the crop yields (in kg/ha) for 15 selected Kharif crops, namely: Rice, Kharif Sorghum, Pearl Millet, Maize, Finger Millet, Pigeon Pea, Minor Pulses, Sesamum, Safflower, Castor, Sunflower, Soybean, Oilseeds, Sugarcane, and Cotton.
Each model was trained and tested independently for all 15 crop yield targets.

3.3 Machine Learning Models Implemented

Five supervised regression models were chosen for this comparison study depending on their popularity, interpretability, and capacity to represent linear and nonlinear relationships:

3.1 Linear Regression (LR): Served as a baseline model to compare with simple linear methods. It presumes a linear, proportionate relationship between the feature variables and the target variable.

3.2 Decision Tree Regressor: A decision rule-learning model based on trees. It can model nonlinear relationships and is simple to interpret but overfits.

3.3 K-Nearest Neighbors Regressor (KNN): A non-parametric, distance-based model that estimates yield depending on the similarity of a point to its neighbors. It is useful when the data has well-localized patterns.

3.4 Random Forest Regressor (RF): A technique that builds many decision trees and combines their predictions. It is resistant to overfitting, deals with feature interactions well, and gives feature importance scores.

3.5 XGBoost Regressor: A sophisticated boosting algorithm that is highly scalable and accurate. It sequentially combines multiple weak learners and optimizes the performance gradient descent methods. using Each of the models was developed using the Python Scikit-learn and XGBoost libraries and tested with the same training data to ensure fairness in comparison.

3.4 Model Training and Optimization

For every machine learning model, hyperparameter tuning was done as required with methods like grid search and cross-validation to enhance performance:

4.1 Linear Regression: No hyperparameters were adjusted.

4.2 KNN: The best number of neighbors (k) was tuned using grid search.

4.3 Decision Tree: Maximum depth and minimum samples per leaf were adjusted to prevent overfitting.

4.4 Random Forest: Number of trees, maximum depth, and feature subset size were tuned.

4.5 XGBoost: The learning rate, number zof estimators, and maximum depth were hyperparameters tuned with grid search.

There was a 5-fold cross-validation strategy used while training to minimize the risk of overfitting and to determine how well the models generalize.

3.5 Model Evaluation Metrics

To and compare model performance, the following were used: comprehensively evaluate multiple metrics

5.1 R^2 (Coefficient of Determination): How well the model accounts for variation in crop yield.

5.2 Adjusted R^2: Penalizes the inclusion of unnecessary predictors, giving a truer reading when there are multiple features.

5.3 RMSE (Root Mean Squared Shows Error): the square root of the mean squared predicted and actual values. Outliers are sensitive to it.

5.4 MAE (Mean Absolute Error): Averages the absolute differences between forecasted and actual yields.

Their average for all 15 crops was used across to rank each model to summarize and A combined normalized score was calculated to provide a general performance indicator that balances accuracy, stability and generalizability.

IV. Results & Discussion

The performances of five machine learning algorithms Random Forest, XGBoost, K-Nearest Neighbors (KNN), Decision Tree, and Linear Regression were evaluated for predicting the crop yield of 15 Kharif crops. The models were evaluated by four common regression evaluation measures: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE). The mean values of these measures over all target variables are given below.

Model	MSE	RMSE	R^2	MAE
Random Forest	0.1362	0.3604	0.7909	0.1530
XGBoost	0.1487	0.3766	0.7727	0.1817
KNN	0.1802	0.4178	0.6968	0.1720
Decision Tree	0.2407	0.4771	0.6359	0.1461
Linear Regression	0.2598	0.5016	0.5751	0.3052

Table1. Average performance metrics across all target variables

1. **Mean Squared Error (MSE):** measures MSE the mean of the squared differences between values. Lower MSE is preferable. Random Forest recorded the lowest actual and predicted average MSE

(0.1362), followed by XGBoost (0.1487). KNN, Decision Tree and Linear Regression had higher MSE values (0.1802, 0.2407, and 0.2598 respectively), reflecting comparatively lower predictive accuracy.

2. **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE and gives a measure of the average size of error in the same unit as the target variable. Random Forest once more worked best with an RMSE of 0.3604, followed by XGBoost (0.3766). The highest RMSE of Linear Regression (0.5016) proved its drawback in capturing non-linear patterns in the data.
3. **Coefficient of Determination (R^2):** R^2 measures the degree to which the variation in the dependent variable is explained by the independent variables. Random Forest produced the highest mean R^2 value of 0.7909, which implies high predictive capability. XGBoost was close behind at R^2 of 0.7727. KNN and Decision Tree produced moderate results (0.6968 and 0.6359 respectively), and Linear Regression performed lowest (0.5751), indicating weak model fit.
4. **Mean Absolute Error (MAE):** MAE estimates the average absolute size of errors in a sample of predictions, but not their direction. Surprisingly, Decision Tree had the smallest MAE (0.1461), followed closely by Random Forest (0.1530). Nevertheless, although it had low MAE, Decision Tree was not performing well on other scores, indicating it could commit small errors on average but is not stable or generalizable.

Linear Regression has the largest MAE (0.3052), as expected from its low accuracy and reliability.

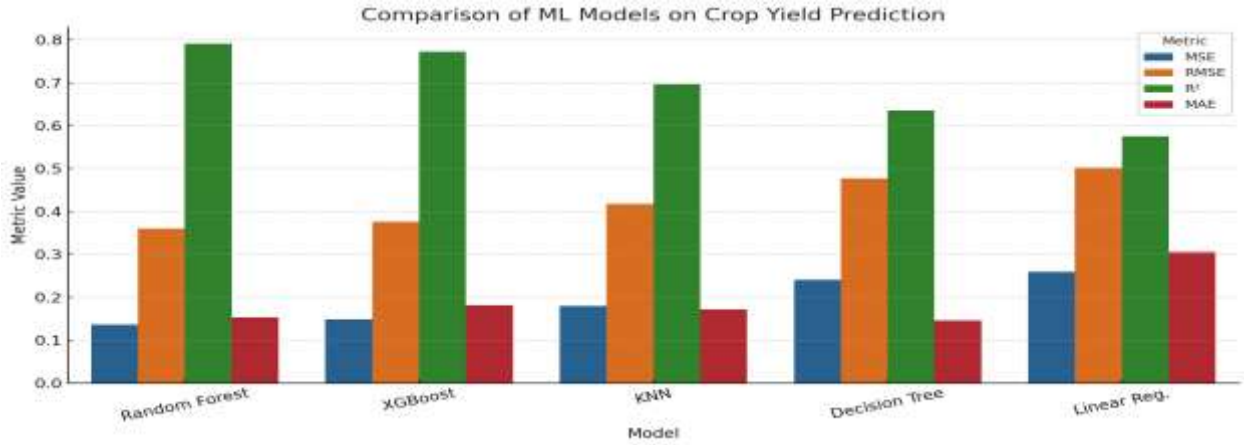

Fig 1. Comparison of ML models on crop yield prediction

Overall Comparison and Model Ranking

1. Random Forest Regressor performed the best of all other models in most of the metrics consistently, making it the strongest and most consistent model for predicting crop yield in this research.
2. XGBoost Regressor trailed closely, with competitive performance and slightly greater error margins.
3. KNN provided stable but less precise predictions, and might be better applied to localized patterns in the data.
4. Decision Tree Regressor, although reporting low MAE, performed poorly in MSE and R^2, which suggests overfitting and inconsistency.
5. Linear Regression was the poorest on all measures, highlighting its inability to deal with complicated, data.

Target Variable wise Evaluation Metrics:

1. R^2 Score by model and target variables (Fig 2)
2. RMSE by model and target variables (Fig 3)
3. MAE by models and target variables (Fig 4)

Some more key factors:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

1. Model ranking consistency across metrics and target variables. (Fig 5)
2. **Model ranking based on average ranks** (Table 2)**:** Random Forest consistently ranks best; Linear Regression consistently ranks worst, while XGBoost, KNN, and Decision Tree show varying degrees of consistency in their rankings across different crops and performance measures.
3. **Best performing model for each crop** (Table 3)**:** Random Forest is the most frequently chosen best model, followed by XGBoost and KNN, indicating their superior predictive capabilities for agricultural yields.
4. Checking the overfitting of the two best performing models. (Fig 8 & Fig 9)

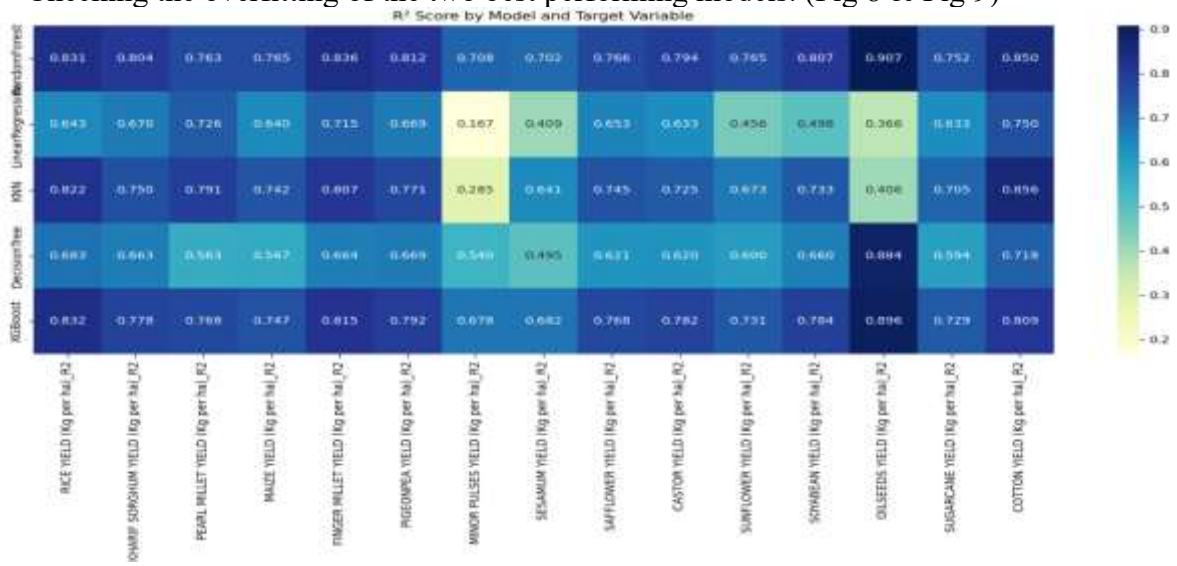

Fig 2. R^2 score by model and target variables

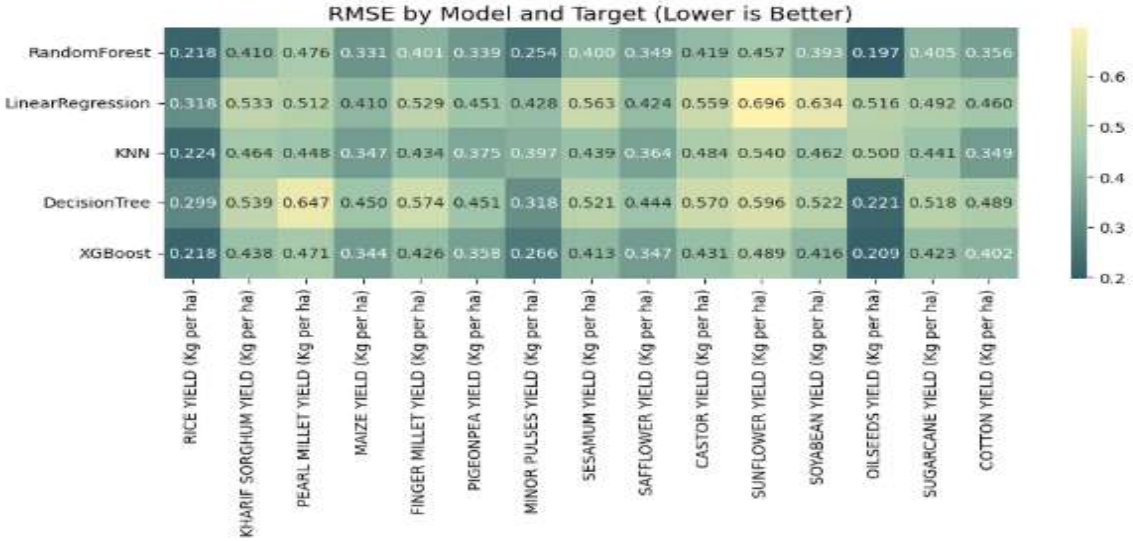

Fig 3. RMSE by model and target variables

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

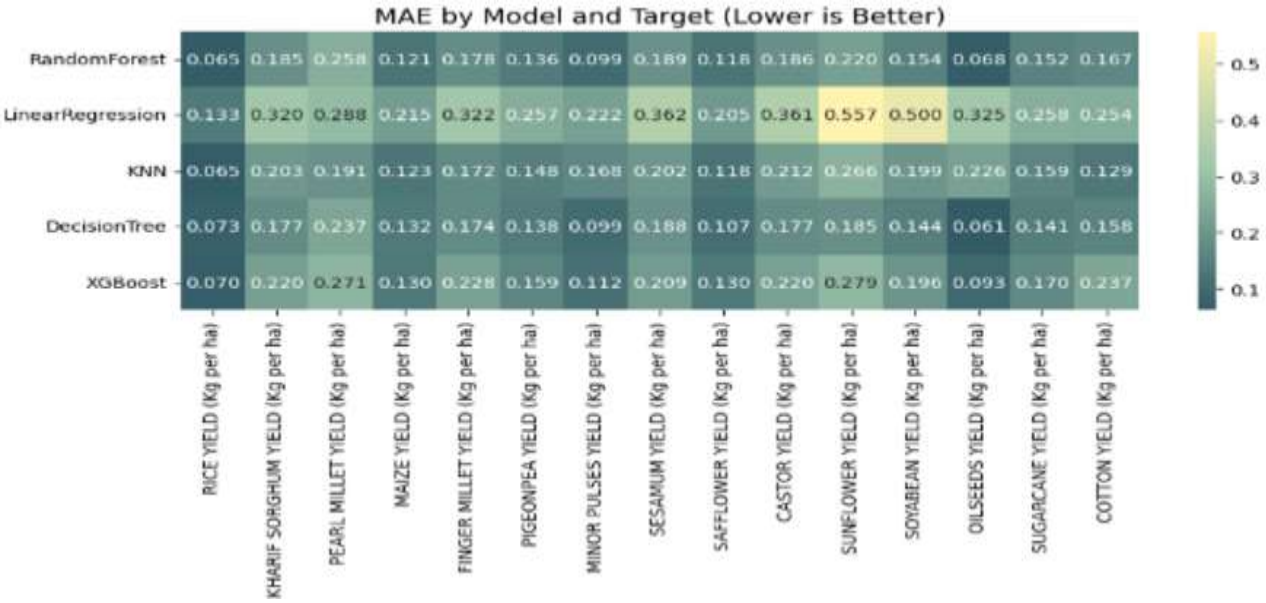

Fig 4. MAE by model and target variables

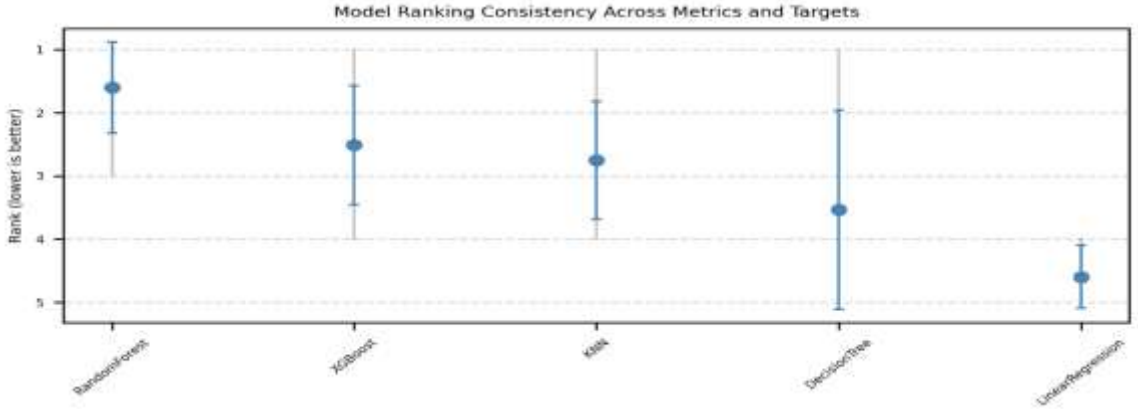

Fig 5. Model ranking consistency across metrics and target variables

	Mean_Rank	StdDev_Rank	Min_Rank	Max_Rank	Range
RandomForest	1.600000	0.719848	1.0	3.0	2.0
XGBoost	2.511111	0.944415	1.0	4.0	3.0
KNN	2.755556	0.933117	1.0	4.0	3.0
DecisionTree	3.533333	1.575379	1.0	5.0	4.0
LinearRegression	4.600000	0.495434	4.0	5.0	1.0

Table 2. Model ranking based on average ranks

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

	Best_Model	R2_Score
RICE YIELD (Kg per ha)	XGBoost	0.832287
KHARIF SORGHUM YIELD (Kg per ha)	RandomForest	0.804345
PEARL MILLET YIELD (Kg per ha)	KNN	0.790532
MAIZE YIELD (Kg per ha)	RandomForest	0.76507
FINGER MILLET YIELD (Kg per ha)	RandomForest	0.835883
PIGEONPEA YIELD (Kg per ha)	RandomForest	0.812443
MINOR PULSES YIELD (Kg per ha)	RandomForest	0.7077
SESAMUM YIELD (Kg per ha)	RandomForest	0.70192
SAFFLOWER YIELD (Kg per ha)	XGBoost	0.76838
CASTOR YIELD (Kg per ha)	RandomForest	0.793941
SUNFLOWER YIELD (Kg per ha)	RandomForest	0.765255
SOYABEAN YIELD (Kg per ha)	RandomForest	0.807068
OILSEEDS YIELD (Kg per ha)	RandomForest	0.907283
SUGARCANE YIELD (Kg per ha)	RandomForest	0.751636
COTTON YIELD (Kg per ha)	KNN	0.856139

Table 3. Best performing model for each crop

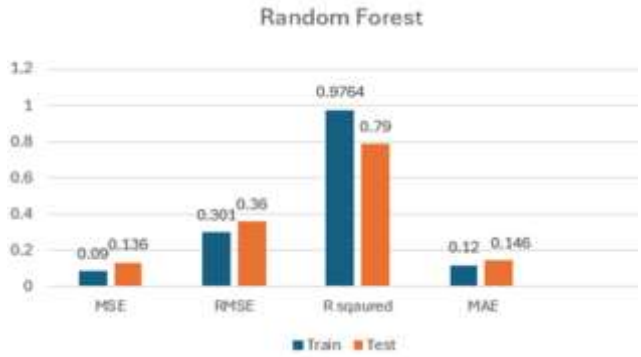

Fig 8. R^2 of Train & Test for Random Forest

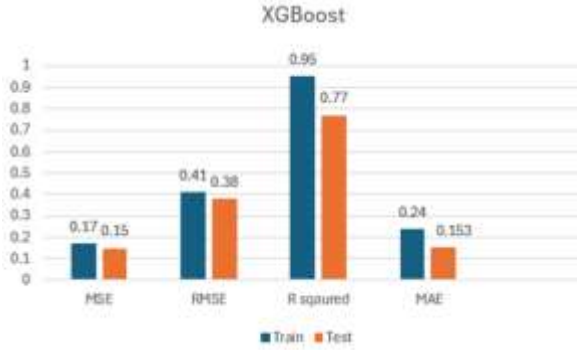

Fig 9. R^2 of Train & Test for XGBoost

V. Conclusion

In this research, we evaluated five different machine learning algorithms to determine how accurate they can be in forecasting the yield of Kharif crops based on historical agricultural and weather data. The algorithms we experimented with were Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and XGBoost. Our aim was to determine which algorithm provides the most accurate and credible results.

Upon testing all models on a single dataset and assessing them in terms of typical error measures such as MSE, RMSE, R^2, and MAE, Random Forest proved to be the best overall. It yielded the highest accuracy and lowest error in the majority of instances. XGBoost also performed well and rivaled Random Forest. KNN and Decision Tree yielded average performance, while Linear Regression yielded the least accurate predictions.

Based on this comparison, we can conclude that Random Forest is the best model to apply for crop yield prediction using such data. It is more efficient at dealing with the complexity of agricultural data compared to the other models. This can assist researchers and planners in selecting correct model for the agricultural industry. making improved decisions in Based on the performance measurements of the XGBoost and Random Forest models from the plots, we can see evidence of overfitting in both, with a greater impact in XGBoost. The XGBoost model has a high training R^2 of 0.95 but a significantly lower test R^2 of 0.77, suggesting that it fits the training data extremely well but generalizes worse to unseen data.

Additionally, the training error values (MSE = 0.17, MAE = 0.24) are less than the test errors (MSE = 0.15, MAE = 0.153), although the difference is not too great indicating slight overfitting. Conversely, the Random Forest model also has a good training R^2 of 0.9764 and a test R^2 of 0.79. Although the difference is approximately the same amount, the lesser overall train-test performance variation in

RMSE and MAE (Train RMSE = 0.301 vs. Test RMSE = 0.36, Train MAE = 0.12 vs. Test MAE = 0.146) indicates that Random Forest is less overfitted than XGBoost and generalizes slightly better. Overall, both models have good predictive ability but present overfitting symptoms especially XGBoost showing limitations in generalization to new data. To correct this, future research can consider applying deep learning techniques, which can provide better accuracy and stability, particularly when used in large-scale and complicated can take advantage of and datasets that feature extraction deeper architectures can handle.

References

[1] Nigam A., Garg S., Agrawal A., Agrawal P. (2019), "Crop yield prediction using machine learning algorithms." In 2019 Fifth International Conference on Image Information Processing (ICIIP) Nov 15 (pp. 125–30). IEEE.

[2] Swathi, T., Sudha, S. (2023), "Crop classification and prediction based on soil nutrition using machine learning methods." Int. j. inf. tecnol. 15, 2951–2960 (2023). https://doi.org/10.1007/s41870-023-01345-0

[3] S. Jambekar, S. Nema and Z. Saquib. (2018), "Prediction of Crop Production in India Using Data Mining Techniques." 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697446.

[4] P. Parameswari, N. Rajathi and K. J. Harshanaa. (2021), "Machine Learning Approaches for Crop Recommendation." 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675480.

[5] Sangeeta, Shruthi G. (2020), "Design and implementation of crop yield prediction model in agriculture." International Journal of Scientific & Technology Research 8.1 (2020): 544-549.

[6] P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber. (2020), "Crop Yield Prediction based on Indian Agriculture using Machine Learning." 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154036.

[7] S. Bhanumathi, M. Vineeth and N. Rohit. (2019), "Crop Yield Prediction and Efficient use of Fertilizers." 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0769-0773, doi: 10.1109/ICCSP.2019.8698087.

[8] Pandith, Vaishali & Kour, Haneet & Singh, Surjeet & Manhas, Dr & Sharma, Vinod. (2020). "Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis." Journal of scientific research. 64. 10.37398/JSR.2020.640254.

[9] Abdel-salam, M., Kumar, N. & Mahajan, S. (2024), "A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning." Neural Comput & Applic 36, 20723–20750. https://doi.org/10.1007/s00521-024-10226-x

[10] Sujatha R., Isakki P. (2016), "A study on crop yield forecasting using classification Techniques." In International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16) (pp. 1–4). IEEE.

[11] Pallavi Kamath, Pallavi Patil, Shrilatha S, Sushma, Sowmya S. (2021), "Crop yield forecasting using data mining." Global Transitions Proceedings, Volume 2, Issue 2, 2021, Pages 402-407, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2021.08.008.

[12] D. Modi, A. V. Sutagundar, V. Yalavigi and A. Aravatagimath. (2021), "Crop Recommendation Using Machine Learning Algorithm." 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702392.

[13] Anbarasan, Padmavathi & Gupta, Arnab & Prakash, Koppadi. (2024), "Crop Recommendation and Yield prediction Using Machine Learning based Approaches." 302-309. 10.1109/ICRTCST61793.2024.10578531.

[14] S.Veenadhari, Dr.BharatMisra, Dr.CDSingh. (2011), "Data mining techniques for predicting crop productivity–a review article."

[15] P. S., M. G., & R., B. (2019), "Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms." Applied Artificial Intelligence, 33(7), 621–642. https://doi.org/10.1080/08839514.2019.1592343

[16] S Iniyan, V Akhil Varma, Ch Teja Naidu. (2023), "Crop yield prediction using machine learning techniques." Advances in Engineering Software, Volume 175, 2023, 103326, ISSN 0965-9978, https://doi.org/10.1016/j.advengsoft.2022.103326.

[17] Kavita Jhajharia, Pratistha Mathur, Sanchit Jain, Sukriti Nijhawan, (2023), "Crop Yield Prediction using Machine Learning and Deep Learning Techniques." Procedia Computer Science, Volume 218, 2023, Pages 406-417, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.01.023.

[18] Nikhil, U.V.; Pandiyan, A.M.; Raja, S.P.; Stamenkovic, Z. (2024), "Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models." Computers 2024, 13, 137. https://doi.org/10.3390/ computers13060137

[19] Dhruvi Gosai, Chintal Raval, Rikin Nayak, Hardik Jayswal, Axat Patel, (2021), "Crop Recommendation System using Machine Learning" International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 7, Issue 3, pp.558-569, May-June-2021. Available at doi : https://doi.org/10.32628/CSEIT2173129

[20] Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK. (2021) "Crop yield prediction integrating genotype and weather variables using deep learning." PLoS One. 2021 Jun 17; 16(6): e0252402. doi: 10.1371/journal.pone.0252402. PMID: 34138872; PMCID: PMC8211294.

[21] Biplob Dey, Jannatul Ferdous, Romel Ahmed. (2024), "Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables." Heliyon, Volume 10, Issue 3, 2024, e25112, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2024.e25112.

[22] Shahhosseini, M., Hu, G., Huber, I. et al. (2021), "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt." Sci Rep 11, 1606 (2021). https://doi.org/10.1038/s41598-020-80820-1