IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 489-497 March 2026-Sept 2026 www.csjournals.com

Machine Learning-Based Crop Yield Prediction: A Comparison of Regression Algorithms

Sagar Duhan, Sandeep Rathee, Dr. Vishal verma
MCA Student, Asst Professor, Professor
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

A large proportion of people depend on agriculture to provide both work and food security, making it an important component of the global economy. Crop yield depends on several points, such as temperature, rainfall, humidity, fertilizer usage, irrigation methods, soil type, and weather conditions. Accurately calculating crop yield crucial for farmers, agricultural specialists, policy makers, enabling them to make sensible choices for crop planning, resource management, or sustainable farming methods. Traditional methods of estimating crop yield are often time-consuming and less accurate, as they fail to adequately analyze the complex interrelationships between various agricultural factors. ML has emerged as a powerful technology capable of processing large agricultural datasets to generate accurate yield predictions. This research uses ML techniques to predict crop yield, based on a review and analysis of findings from five selected research papers. The study emphasizes machine learning methods exclusively and does not include deep learning approaches. It offers a number of machine learning methods, such linear regression, Xgboost, decision trees, and random forest compares them based on predictive capability, accuracy, computational efficiency, and suitability for agricultural applications. These studies examine key factors influencing crop yield, including weather conditions, soil characteristics, crop type, fertilizer usage, and irrigation methods. Data collection from agricultural datasets, data preprocessing, feature selection, model training, and performance evaluation are all covered in the presented methodology. To enhance data quality prior to training, techniques such as handle missing values, and feature engineering were applied. Having a root mean square error of 0.4993, mean absolute error of 0.3983, and R2 score of 0.9132, linear regression was shown to be the best technique for estimating crop yield among the trained models.

Keywords: Crop Yield Prediction, Machine Learning, Agriculture, XG Boost, Decision tree, Random Forest, Linear Regression, Data Preprocessing, Feature Selection, Precision Agriculture.

Introduction

Technology has brought about significant changes in agriculture, enabling us to make predictions that assist farmers, the government, and businesses in making informed decisions regarding agricultural production, thereby benefiting everyone. Modern precision agriculture requires

IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 489-497 March 2026-Sept 2026 www.csjournals.com

accurate models to prevent food shortages [1]. There are different climates and terrains in India due to which this task is a bit difficult but we are trying to solve it with modern methods. Agriculture is considered the backbone of the global economy with the advent of technology, we are transitioning from traditional farming to precision farming.

This shift encourages every individual to cultivate using modern methods an approach in which every drop of water and every seed is meticulously accounted for. This system also assists farmers in adopting better crop rotation practices and maintaining soil health, thereby enabling them to boost their yields in harmony with nature [2]. In today's digital era, old crop prediction models are failing due to climate change and changing the soil complexity. Although, the new deep learning models give good results, but their working method has become a hidden logic which no one understands.

Literature Review

Md. Abu Jabed et al. [2024] examined a review crop yield prediction in agriculture. They use machine learning algorithm like Random Forest (RF), XG Boost and this model concludes perform best with over all accuracy of 93% [1].

Abebe Gatie Betew et al. [2026] examined a review crop yield prediction in agriculture. Effective agricultural planning and global food security have been said to depend on this concept. They use machine learning algorithms including gradient boosting, support vector machine, and random forest. According to this approach, accuracy above 90% is the best [2].

L.K. Subramaniam et al. [2024] utilize deep learning to predict the yields of Indian regional crops. Its primary objective is to enhance farmers' income and improve import-export strategies. They use Weight-Tuned Deep Convolution Neural Network (WTDCNN). This model achieved best accuracy 98.96% [3].

M.M. Islam et al. [2024] this paper utilizes to estimate the production of the 10 most widely consumed crops in the world, a machine learning model was applied to data from 101 countries. They use XgBoost ensemble emerged as the top-performing. This model achieved approximately 97.45% [4].

S. Yash Pravesh et al. [2024] introduced a DL Model, LSTM, combining k-means clustering & transformer network for global crop yield prediction using historical climate data across 101 countries. The model successfully captures climate driven variations outerperforming baseline regression models with exceptional R^2 score of 0.951 [5].

Dataset description

The dataset download from Kaggle [6]. It contains 998,227 rows and 10 attributes. It provides information regarding crops specifically, the types of crops suitable for different locations. The dataset includes a target variable **yield per hectare** which indicates the crop production output.

Table 1: Dataset Description

Sr. No.	Column Name	Data Type	Variable Type	Description	Example Values
1	Region	Categorical	Independent(Feature)	Geographical area where crop is cultivated.	East, West, North, South
2	Soil Type	Categorical	Independent(Feature)	Specific classification of soil in that region.	Sandy, Clay, Loam, Slit
3	Crop	Categorical	Independent(Feature)	Type of crop grown during the specific agriculture cycle.	Wheat, Rice, Cotton, Maize
4	Weather Condition	Categorical	Independent(Feature)	Specific classification of weather for crops growth.	Cloudy, Sunny, Rainy
5	Temperature	Numerical	Independent(Feature)	Average atmospheric temperature record in degree Celsius.	25,30,35,40
6	Irrigation	Numerical	Independent(Feature)	In this we can see how many times irrigation has been done.	1,3,5,7
7	Fertilizer	Numerical	Independent(Feature)	Quantity of organic or chemical fertilizer is used per hectare to boost soil nutrients.	20,30,40,50
8	Rainfall	Numerical	Independent(Feature)	Total annual and seasonal rainfall record in mm.	100, 200, 400, 500, 300

IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 489-497 March 2026-Sept 2026 www.csjournals.com

9	Harvested days	Numerical	Independent(Feature)	Total days annually and seasonally harvesting of crops.	50,70,90,110
10	Yield	Numerical	Dependent(Target)	In this measured in tons per hectare.	1.5, 2.5, 3.5, 4.5

Methodology

This flowchart illustrates the entire methodology adopted in this research. It outlines the study's process, ranging from data collection to the final conclusions. Figure.1 depicts the workflow of this study:

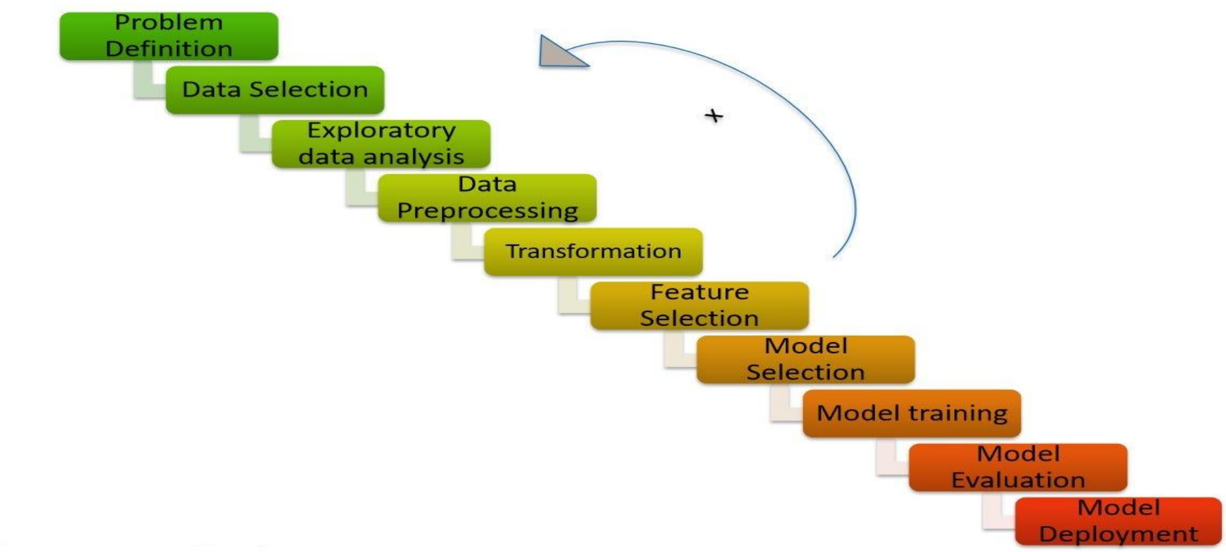

Fig.1 The flow chart of present study

Data Collection: The dataset was downloaded from Kaggle [6]. It contains 9,98,227 rows and 10 attributes.

- Data Preprocessing: Various data processing techniques were employed to improve the quality of the dataset.
- Handling missing values: Missing values were identified and imputed using the mode, mean, and median.
- Removing duplicate records: To avoid redundancy in the dataset, duplicate entries were checked for and removed.
- Data cleaning: Incorrect and unnecessary data within the dataset was either corrected or removed.

- Feature engineering: Feature engineering was performed to convert raw data into a best format for ML algorithms. Categorical data was converted into numerical values using **one-hot encoding and label encoding**.

Feature and target selection

Feature selection was used to select useful features and remove unnecessary features with the goal to improve the ML model's accuracy. The dataset was split into input features (X) and the output feature (y).

Input features (X):

- soil type
- crop
- Irrigation used
- Rainfall mm
- weather condition
- temperature used
- Fertilizer used
- region
- Days to harvest

Essential input features used for model training:

- rainfall in mm
- Irrigation
- fertilizer

Output features (y):

- Yield per hectare

Result

The result finding main three areas: correlation matrix analysis and model performance comparisons.

Correlation Matrix

A correlation matrix is used to analyze the relationships between different features in the dataset. It helps identify strongly and weakly correlated variables and assists in feature selection.

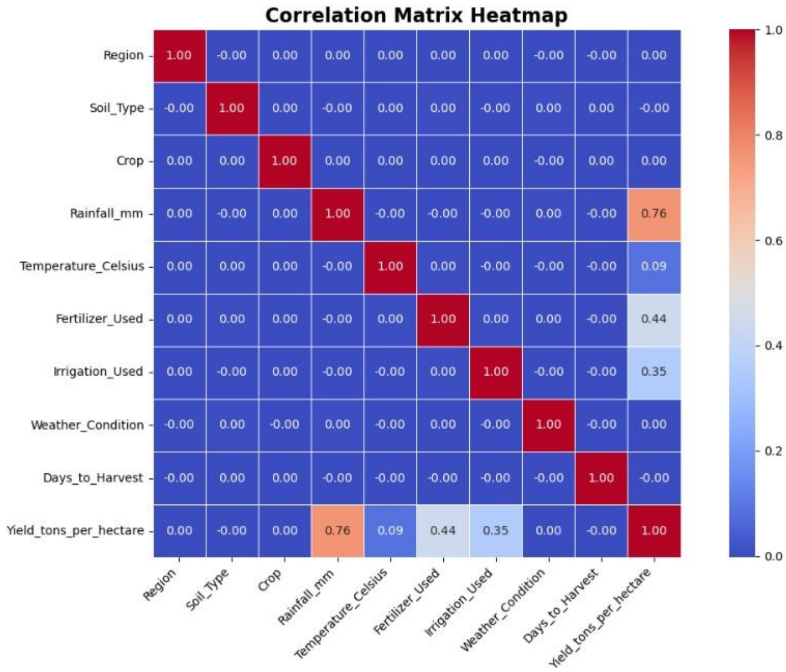

Fig.2: Correlation Matrix

Features with higher correlation values are considered more important for training the model. A heat map visually represents the relationships between the attributes.

XGBoost Feature Importance

The importance scores of different features for the Xgboost model are shown in this graph. Utilized fertilizer has the highest importance score of 0.58, making it a significant feature. Rainfall mm comes in third place with a score of roughly 0.15, followed by irrigation used with a score of 0.28. The model's prediction is barely affected by the remaining characteristics, such as temperature, days to harvest, crop, and soil type.

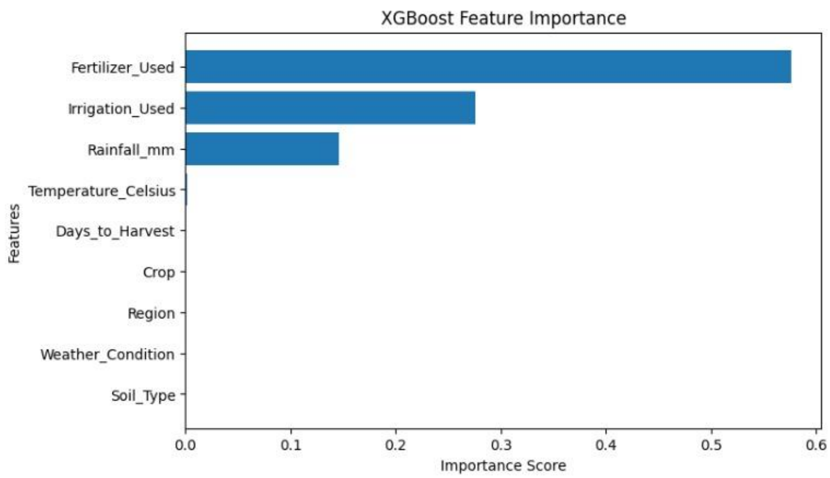

Fig.3: Xgboost Features

Random Forest Feature Importance

This Random forest model displays feature importance. The results differ slightly from the XG boost model here, Rainfall mm is the primary feature, with the highest importance score of 0.64. Fertilizer used stands at 0.21, and Irrigation used is approximately 0.14. Other features, such as Region, Soil type, and Weather condition, have negligible importance in this model.

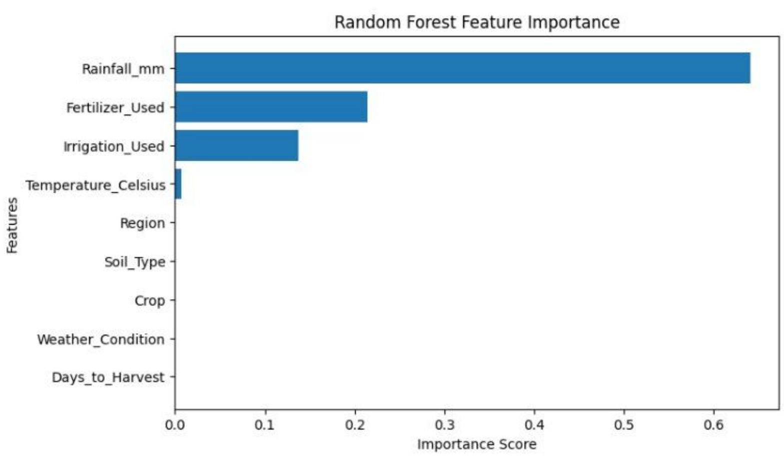

Fig.4: Random Forest Feature

Model Comparisons

This bar chart compares the performance of various ML models such as **Random forest, Decision tree, Linear regression, and Xgboost** based on their R^2 scores. Random forest, XGBoost, and linear regression all demonstrate excellent and nearly equal performance with R^2 scores around 0.90 in contrast the Decision Tree model performs slightly lower than the others, with an R^2 score of 0.82.

IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 489-497 March 2026-Sept 2026 www.csjournals.com

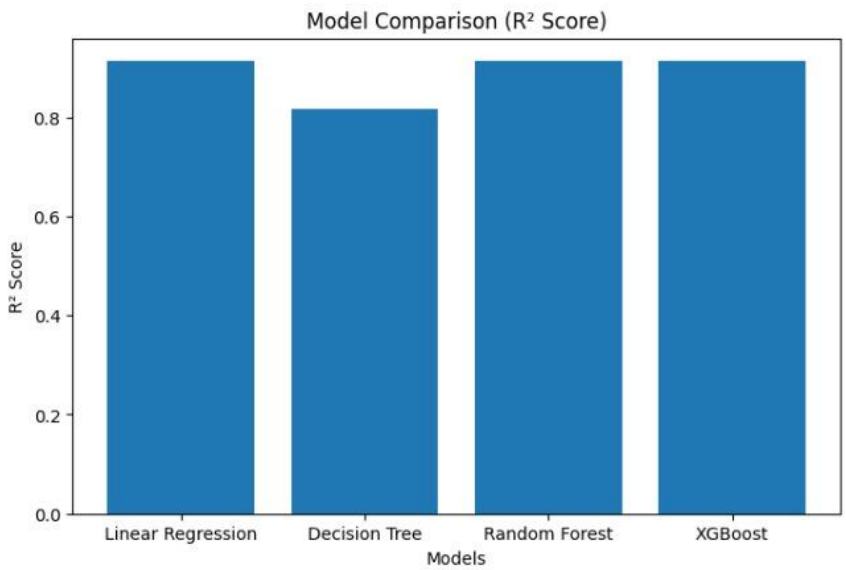

Fig.5: Model Comparisons

Performance Result

As you can see in this table, the results for the methods described earlier are presented. This table allows us to compare decision tree, random forest, linear regression, and xgboost to find out which yields the best results.

Table 2: Performance comparisons of model

Model	MAE	MSE	RMSE	R^2 Score
Linear Regression	0.3983	0.2493	0.4993	0.9132
Decision Tree	0.5825	0.5313	0.7289	0.8151
Random Forest	0.4114	0.2658	0.5156	0.9075
XGBoost	0.3996	0.2509	0.5009	0.9127

Conclusion

According to the results obtained, linear regression had a low prediction error and a high R^2 score of 0.9132. XG boost and Random forest also produced accurate results, while Decision tree showed comparatively lower performance. According to the correlation matrix, rainfall has the greatest impact on crop productivity, followed by irrigation and fertilizer. In overall, machine learning methods can improve crop yield prediction and help farmers to take smarter decisions to increase production in agriculture.

References:

[1] Md. Abu Jabed and Masrah Azrifah Azmi Murad, "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," Heliyon, vol. 10, Art. No. e40836, Nov. 2024, doi:10.1016/j.heliyon.2024.e40836

[2] Abebe Gatie Betew, Girma Gebresenbet, Daniel Ayalew Mengistu, Geta Kidanemariama Gelaw and Abdulkerim Mohammed Yibre, "Historical and contemporary crop yield prediction models: Key lessons and innovation," smart agricultural technology, vol. 13, Art. No. 101672, 2026, doi:10.1016/j.atech.2025.101672

[3] Leelavathi Kandasamy Subramaniam and Rjasenathipathi Marimuthu, "Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crop," e-Prime – Advances in electrical engineering, electronics and energy, vol. 8, Art. No. 100611, 2024, doi:10.1016/j.prime.2024.100611

[4] Mohmmad M. Islam, Majed Alharthi, Rotana S. Alkadi, Rafiqul Islam and Aabdul Kadar Mohammad Masum, "crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia," AIMS Agriculture and Food, vol. 9, no. 4, pp. 980-1003, 2024, doi:10.3934/agrfood.2024053

[5] S. Yash Pravesh, Nakshatra Garg, Ravik Arora, Sudhanshu Singh and S. Siva Sankari, "Predictive modeling of crop yield using deep learning based transformer with climate change effects," Multidisciplinary technovation, vol. 6, no. 6, pp. 223-240, 2024, doi:10.54392/irjmt24616

[6] Kaggle, "crop yield prediction," [online]. Available: https://www.kaggle.com/.Accessed: Jan.24, 20026