

Article – e008

Machine Learning-Based Crop Yield Forecasting Using Environmental and Agricultural Parameters

Pavan Sahu¹, Prof. O.P. Karada²

¹ Student, MTech.(CSE), Sage University, Indore, India

² Professor, Sage University, Indore, India

Received: 17/07/2026

Revision Received: 18/08/2026

Accepted: 28/08/2026

ABSTRACT

Crop yield forecasting is an important component of precision agriculture, as reliable predictions can support agricultural planning, resource allocation, and food-security decisions. However, crop productivity is influenced by several interacting environmental and agricultural factors, making accurate prediction a challenging task. This study develops a machine learning framework for crop yield forecasting by considering rainfall, average temperature, pesticide usage, crop type, geographical area, and year-wise agricultural information. A processed dataset derived from publicly available FAO and World Bank sources was used for the experimental analysis.

Seven regression algorithms, namely Linear Regression, Random Forest, Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), Decision Tree, and Bagging Regressor, were trained and comparatively evaluated. Their performance was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2), and 5-Fold Cross-Validation. The experimental results show that ensemble-based models provide stronger predictive performance than conventional regression approaches. Among the evaluated models, the Bagging Regressor produced the highest test R^2 score of 98.59% and a mean cross-validation score of 98.80%, while Random Forest and XGBoost also demonstrated strong predictive performance.

The main contribution of this study is a comparative crop yield forecasting framework that combines agricultural and environmental variables with multi-model evaluation, cross-validation, feature-importance analysis, and multiple error metrics. The results demonstrate the potential of ensemble machine learning methods for crop yield estimation and provide a foundation for their future application in data-driven smart agriculture systems.

KEYWORDS: Crop Yield Prediction, Machine Learning, Smart Agriculture, Ensemble Learning, Random Forest, XGBoost, Precision Agriculture

I. INTRODUCTION

Agricultural production is influenced by a combination of climatic, environmental, and cultivation-related conditions. Reliable crop yield forecasting can therefore assist in agricultural planning, resource allocation, supply-chain management, irrigation decisions, and other data-driven farming activities. However, obtaining accurate yield estimates is difficult because crop productivity varies with factors such as rainfall, temperature, pesticide usage, geographical conditions, crop characteristics, and changes across agricultural years.

Conventional crop forecasting methods commonly depend on historical observations and statistical relationships between agricultural variables and production. Such approaches can provide useful baseline estimates, but their ability to represent complex and nonlinear interactions among multiple predictors may be limited. The increasing availability of agricultural datasets has consequently encouraged the use of Artificial Intelligence (AI) and Machine Learning (ML) techniques for predictive agricultural applications.

Machine learning models can identify relationships within multidimensional datasets without requiring all relationships between input variables and crop yield to be explicitly defined in advance. In particular, ensemble-based methods can combine the predictions of multiple learners and may provide improved predictive performance for datasets containing nonlinear relationships. These characteristics make machine learning a suitable approach for investigating the combined influence of environmental and agricultural variables on crop productivity.

Previous research has investigated crop yield prediction using different combinations of weather observations, soil-related information, remote-sensing data, and agricultural records. However, studies may differ considerably in the number of input variables and learning algorithms considered. In addition, evaluating several regression models under a common dataset and validation procedure can provide a clearer comparison of their predictive behavior. This creates a need for comparative analysis that considers multiple machine learning approaches together with consistent performance evaluation.

In this study, a machine learning-based framework is developed for crop yield forecasting using environmental and agricultural parameters. The experimental dataset contains crop yield information together with rainfall, pesticide usage, average temperature, crop type, geographical area, and year-wise agricultural records obtained from publicly available repositories.

Seven regression algorithms are considered in the comparative analysis: Linear Regression, Random Forest, Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), Decision Tree, and Bagging Regressor. Their predictive performance is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2), and 5-Fold Cross-Validation.

The experimental results show that ensemble-based approaches provide the strongest performance among the evaluated models. The Bagging Regressor obtains the highest test R^2 score of 98.59% and a mean cross-validation score of 98.80%, while Random Forest and XGBoost also achieve comparatively high predictive performance. These results provide a comparative basis for selecting machine learning models for crop yield estimation using the selected agricultural and environmental variables.

The main contributions of this study are as follows:

- A machine learning framework is developed for crop yield forecasting using combined environmental and agricultural parameters.
- Seven regression algorithms are implemented and compared using a common processed agricultural dataset.
- Model performance is assessed using multiple error metrics together with 5-Fold Cross-Validation.
- Feature-importance analysis is used to examine the relative contribution of the selected predictors to crop yield estimation.

The remainder of this paper is organized as follows. Section II reviews related research on machine learning-based crop yield prediction. Section III describes the dataset and proposed methodology. Section IV presents the experimental setup and implementation procedure. Section V discusses the experimental results and comparative analysis. Finally, Section VI presents the conclusions and directions for future work.

II. RELATED WORK

The use of machine learning in agriculture has expanded considerably as researchers seek data-driven methods for estimating crop productivity. Crop yield prediction studies have considered a range of information sources, including weather observations, agricultural records, environmental variables, and remote-sensing data. The choice of predictors and learning algorithms can substantially influence the resulting prediction performance.

Jeong *et al.* [1] examined Random Forest-based approaches for crop yield prediction at global and regional scales. Their work demonstrated the applicability of ensemble tree models to agricultural yield estimation using environmental and agricultural information. Everingham *et al.* [2] investigated machine learning methods for sugarcane yield prediction and reported that machine learning-based approaches can provide accurate estimates of agricultural productivity.

Other studies have investigated boosting-based methods for crop yield prediction. Shahhosseini *et al.* [3] evaluated machine learning approaches for corn yield estimation using satellite and climate-related information. Their findings indicate that machine learning models can capture relationships between environmental conditions and crop yield that are difficult to represent using simpler regression approaches.

Deep learning has also been explored for agricultural yield estimation. Khaki and Wang [4] investigated deep neural network models for crop yield prediction and demonstrated the potential of data-driven learning approaches for estimating agricultural outcomes from multiple input variables. Such studies highlight the increasing use of computational learning methods for handling complex agricultural datasets.

Gandhi *et al.* [5] investigated machine learning-based rice crop yield prediction, further demonstrating the application of predictive learning techniques to crop-specific agricultural forecasting. Collectively, these studies indicate that environmental and agricultural variables can provide useful information for crop yield estimation.

Despite these developments, differences remain in the datasets, input variables, algorithms, and evaluation procedures adopted across crop yield prediction studies. A comparative evaluation using a common processed dataset and consistent validation procedure can therefore provide useful insight into the relative performance of different regression algorithms. In this study, seven machine learning models are evaluated using multiple error measures and 5-Fold Cross-Validation, with additional feature-importance analysis used to examine the contribution of the selected predictors.

III. PROPOSED METHODOLOGY

A. Dataset Description

The dataset used in this research was obtained from publicly available agricultural and environmental repositories. The final processed dataset was constructed by integrating information collected from multiple sources, including the Food and Agriculture Organization (FAO) and the World Bank climate database. The resulting dataset combines agricultural production records with environmental variables used for crop yield prediction. The dataset contains approximately **28,242 observations** and includes both categorical and numerical attributes. The main features are **Area**, representing the geographical region or country; **Item**, representing the crop category; **Year**, representing the agricultural production year; **Average Rainfall (mm/year)**; **Pesticides Usage (tonnes)**; and **Average Temperature (°C)**. **Crop Yield (hg/ha)** was selected as the target variable. The dataset contains multiple crop categories and geographical regions represented in the original public dataset. These features were selected because they provide information related to environmental conditions, agricultural inputs, geographical characteristics, and temporal variation in crop production.

B. Data Preprocessing

Raw agricultural data contain categorical and numerical variables that require preprocessing before machine learning implementation. During preprocessing, unnecessary or redundant columns were removed from the processed dataset. The categorical variables **Area** and **Item** were converted into numerical representations using Label Encoding. Feature correlation analysis was subsequently performed to examine relationships among the selected predictor variables. The prepared dataset was then divided into training and testing subsets using a **70:30 train-test split with *random_state* = 42**. The target variable, Crop Yield, was separated from the predictor variables before model training. The preprocessing and model-development workflow is illustrated in Fig. 1.

C. Machine Learning Framework

The implemented framework follows a supervised regression workflow in which the selected agricultural and environmental variables are used to estimate crop yield.



Figure 1. Data-flow Diagram

The system architecture consists of the following stages:

Dataset Collection → Data Preprocessing → Feature Engineering → Model Training → Performance Evaluation → Crop Yield Prediction

Multiple machine learning algorithms were implemented to perform comparative evaluation of prediction performance.

The following models were investigated:

- [1] Linear Regression
- [2] Random Forest Regressor
- [3] Gradient Boosting Regressor
- [4] XGBoost Regressor
- [5] K-Nearest Neighbors (KNN)
- [6] Decision Tree Regressor
- [7] Bagging Regressor

The use of multiple algorithms enables systematic comparison between traditional regression techniques, instance-based learning methods, decision tree models, and ensemble learning approaches.

D. Performance Evaluation Metrics

To assess predictive performance, multiple evaluation metrics were employed.

1) Mean Squared Error (MSE)

Mean Squared Error measures the average squared difference between predicted and actual yield values

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

where:

y_i = actual yield value

$\hat{y}_i$ = predicted yield value

n = number of observations.

2) Mean Absolute Error (MAE)

MAE calculates the average magnitude of prediction error.

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

3) Mean Absolute Percentage Error (MAPE)

MAPE evaluates prediction accuracy in percentage form.

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

4) Coefficient of Determination (R² Score)

The R² score measures how effectively the predictive model explains variance in crop yield observations.

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

Higher R^2 values indicate better predictive capability.

E. Cross-Validation Strategy

To assess model stability beyond a single train–test partition, **5-Fold Cross-Validation** was applied. The dataset was divided into five folds, with four folds used for model training and the remaining fold used for validation in each iteration. The validation performance obtained from the five iterations was averaged to calculate the final cross-validation score. This procedure provides an additional assessment of model performance across different partitions of the available observations.

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

A. Experimental Environment

The framework was implemented in Python using a Jupyter Notebook environment. Pandas and NumPy were used for data preparation and numerical operations, while Matplotlib and Seaborn supported visualization. Scikit-Learn and XGBoost were used for model development and evaluation. The overall workflow comprised data preprocessing, model training, prediction, visualization, and performance assessment.

The primary software libraries used in this study include:

1. **Pandas** — data handling and preprocessing
2. **NumPy** — numerical computation
3. **Matplotlib** — graphical visualization
4. **Seaborn** — statistical plotting
5. **Scikit-Learn** — machine learning algorithms and evaluation
6. **XGBoost** — gradient boosting implementation

The experimentation workflow involved dataset preprocessing, model training, prediction generation, visualization, and performance analysis.

B. Dataset Configuration

The processed dataset used in this research contains approximately **28,242 agricultural observations** collected from integrated environmental and agricultural repositories.

The dataset includes both **categorical** and **numerical** variables.

Table I

Dataset Features Used for Prediction

Feature	Description
Area	Geographic region / country
Item	Crop category
Year	Agricultural production year
Average Rainfall	Annual rainfall measurement (mm/year)
Pesticides Usage	Agricultural pesticide consumption
Average Temperature	Average environmental temperature
Crop Yield	Target prediction variable

The dependent variable selected for supervised learning was **Crop Yield (hg/ha)**.

C. Model Configuration

The seven regression algorithms were implemented using their respective machine learning libraries. The principal model configurations used in the experiments are summarized in Table II. Where specific parameters were not explicitly configured, the default settings of the corresponding implementation were retained. No extensive automated hyperparameter optimization was performed, as the primary objective of the study was to provide a comparative evaluation of the selected regression algorithms under the stated experimental configuration.

Table II — Model Configurations Used in the Experiments

Model	Main Configuration
Linear Regression	Default parameters
Random Forest Regressor	random_state = 42; remaining parameters default
Gradient Boosting Regressor	n_estimators = 100, learning_rate = 0.1, max_depth = 3, random_state = 42
XGBoost Regressor	random_state = 42; remaining parameters default
KNN Regressor	n_neighbors = 5; remaining parameters default
Decision Tree Regressor	random_state = 42; remaining parameters default
Bagging Regressor	n_estimators = 150, random_state = 42

Model Configurations Used in the Experiments

D. Implementation Workflow

The implementation pipeline adopted in this research consists of multiple sequential stages.

Step 1 — Data Loading

The processed dataset was imported into the Python environment and inspected for structural consistency.

Step 2 — Data Preprocessing

Preprocessing operations included:

- removal of redundant attributes,
- encoding categorical variables,
- feature preparation,
- train–test partitioning.

A **70:30 train–test split** was used for experimental evaluation.

Step 3 — Model Training

Seven regression algorithms were implemented and trained independently:

- Linear Regression
- Random Forest

- Gradient Boosting
- XGBoost
- KNN
- Decision Tree
- Bagging Regressor

Each model was trained using the prepared feature matrix and target variable.

Step 4 — Prediction and Evaluation

After training, prediction outputs were generated on the testing dataset.

Performance analysis was performed using:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- R^2 Score
- 5-Fold Cross-Validation

E. Visualization and Analytical Framework

Visualization techniques were employed to interpret dataset behavior and model performance.

The analytical framework included:

1) Correlation Heatmap

A correlation heatmap was generated to examine dependencies among environmental and agricultural variables.

2) Feature Importance Analysis

Feature importance analysis was performed using the Random Forest model to identify influential predictors affecting crop yield.

3) Comparative Model Evaluation

A comparative bar plot was constructed to compare model prediction performance using R^2 scores.

4) Actual vs Predicted Analysis

Scatter plot visualization was used to evaluate prediction consistency between observed and estimated crop yield values.

These visualizations support interpretation of the dataset characteristics and comparative model performance.

F. Validation Strategy

The dataset was divided into training and testing subsets using a **70:30 train-test split**, with `random_state = 42` used to ensure reproducibility of the partition. The models were trained using the training subset, and their predictive performance was subsequently evaluated on the held-out testing subset. In addition to the single train-test evaluation, **5-Fold Cross-Validation** was performed for each of the seven machine learning models. In each fold, four partitions were used for training and the remaining partition was used for validation. The mean score across the five folds was calculated for comparative analysis. The cross-validation procedure was used to examine the consistency of model performance across different partitions of the available observations.

The train-test and cross-validation procedures were performed at the observation level rather than using a grouped split by geographical area, crop category, or year. Therefore, related observations corresponding to the same or similar crop-area combinations across different years may occur in different partitions. The reported results should consequently be interpreted as predictive performance under the experimental random-split setting used in this study, rather than as performance on completely unseen geographical regions or future years. A stricter temporal or group-based evaluation represents an important direction for future investigation.

V. RESULTS AND DISCUSSION

A. Comparative Model Performance Analysis

Seven machine learning algorithms were implemented and evaluated for crop yield prediction using environmental and agricultural parameters.

The experimental evaluation was conducted using **R² Score**, **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, and **Mean Absolute Percentage Error (MAPE)**.

TABLE III

Comparative Performance of Machine Learning Models

Model	R ² Score	MSE	MAE	MAPE
Linear Regression	0.0737	6293718529	60955.32	2.4195
Random Forest	0.9856	97654323	3480.84	0.1026
Gradient Boost	0.8311	1147345777	21184.66	0.5968
XGBoost	0.9743	174504240	7073.09	0.1984
KNN	0.2882	4836388296	47716.36	1.6312
Decision Tree	0.9762	161890347	3559.27	0.0961
Bagging Regressor	0.9859	95933983	3450.51	0.1012

The results indicate substantial performance variation among the evaluated algorithms.

The **Bagging Regressor** achieved the highest predictive performance with an **R² score of 98.59%**, accompanied by the lowest MSE and MAE values.

The Random Forest model demonstrated nearly equivalent performance, achieving an R² score of 98.56%. This result indicates that ensemble tree-based models performed strongly on the selected dataset. The XGBoost model also achieved strong prediction capability, with an R² score of 97.43% and relatively low prediction error.

In contrast, Linear Regression and KNN exhibited substantially lower predictive performance. This suggests that the relationships represented in the selected dataset were not captured as effectively by these models.

B. Cross-Validation Performance Evaluation

To assess model robustness and generalization capability, **5-Fold Cross-Validation** was conducted.

TABLE IV

Mean Cross-Validation Scores

Model	Mean CV Score
Linear Regression	0.0731
Random Forest	0.9875
Gradient Boost	0.8390
XGBoost	0.9775
KNN	0.3333
Decision Tree	0.9807
Bagging Regressor	0.9880

The cross-validation results show that the ensemble models maintained high mean scores across the five validation folds. The Bagging Regressor achieved the highest mean cross-validation score of **0.9880**, followed by Random Forest with **0.9875** and XGBoost with **0.9775**. These results are consistent with the relative performance observed in the held-out testing evaluation. However, because the folds were generated from the available observations rather than grouped by crop, geographical area, or year, the cross-validation results should be interpreted as evidence of consistency across the sampled partitions rather than as validation on completely independent regions or future years.

C. Correlation Analysis of Features

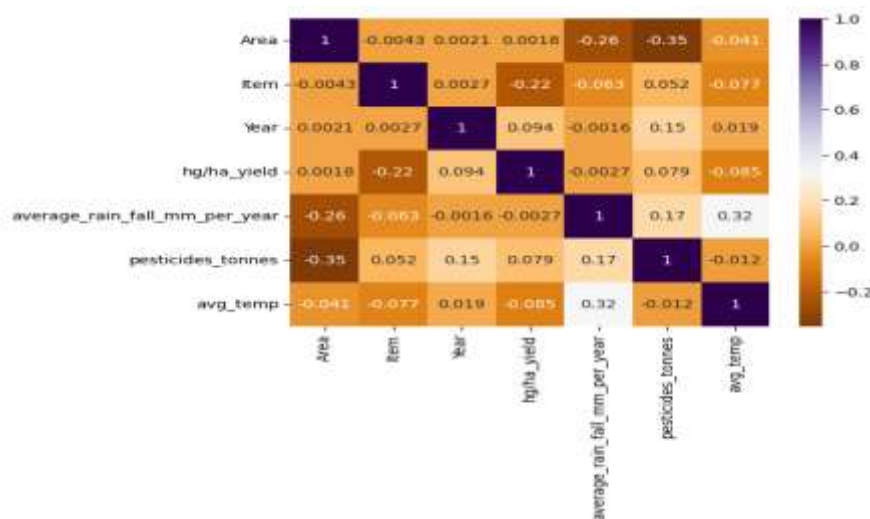


Figure 2. Correlation heatmap of environmental and agricultural features.

The correlation heatmap was generated to investigate relationships among predictor variables.

The correlation heatmap was used to examine pairwise relationships among the selected variables. The observed correlations provide an overview of the degree of association between the environmental and agricultural features used in the analysis.

The correlation analysis provides an overview of the relationships among the selected agricultural and environmental variables.

D. Feature Importance Analysis

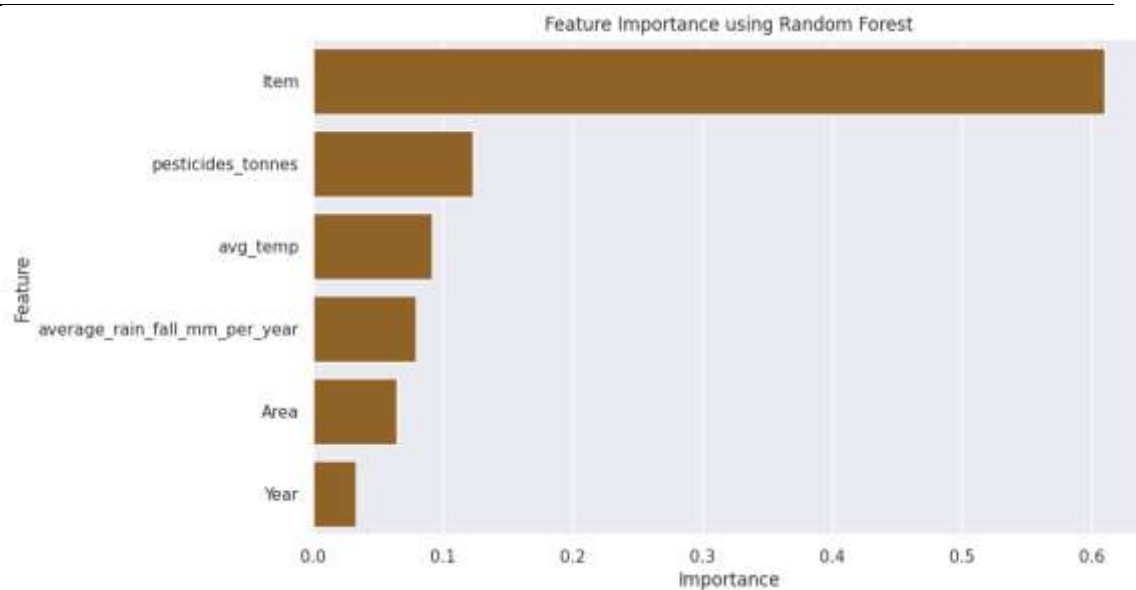


Figure 3. Random Forest feature importance analysis.

Feature importance analysis was conducted using the Random Forest model.

The results reveal that **crop category (Item)** contributes most significantly to crop yield prediction.

Additional influential variables include:

1. pesticide usage,
2. average temperature,
3. rainfall,
4. geographical region.

The **Year** attribute demonstrated comparatively lower influence.

These findings indicate that both environmental conditions and crop-specific characteristics were among the more influential predictors used by the Random Forest model.

E. Comparative Visualization of Model Performance

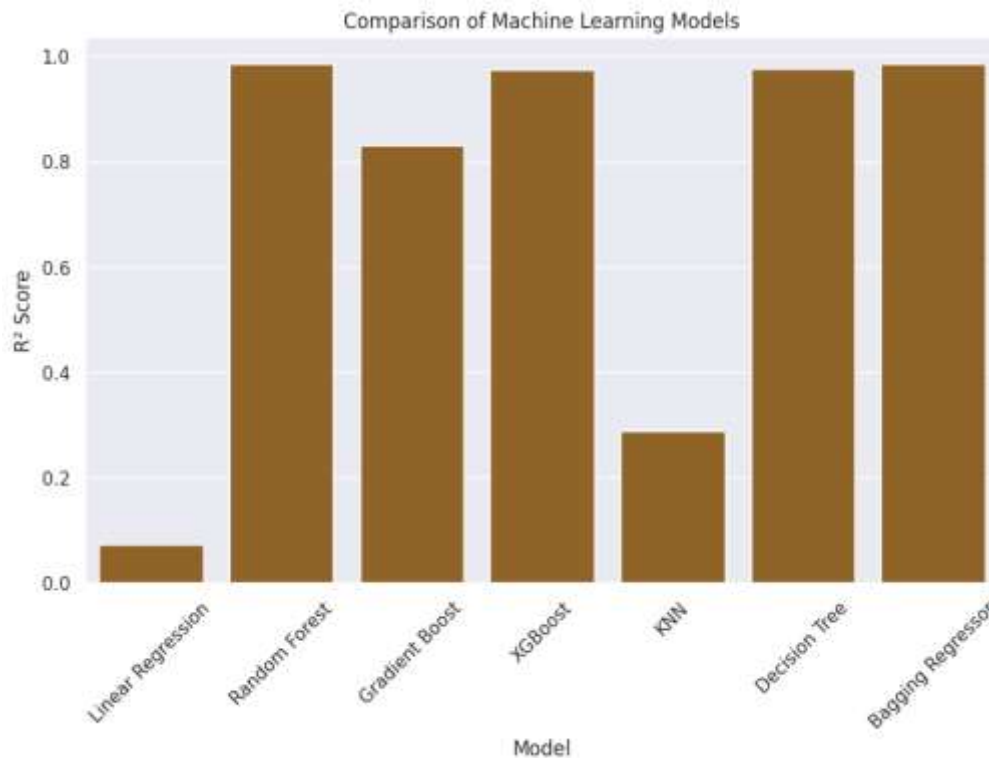


Figure 4. Comparative R² score analysis of machine learning models.

The comparative R² plot shows that the ensemble-based models achieved higher test scores than the Linear Regression and KNN models in this experiment.

Bagging Regressor, Random Forest, and XGBoost consistently outperformed traditional methods.

The comparative visualization highlights the stronger performance of the ensemble-based models on the selected dataset. Bagging Regressor, Random Forest, and XGBoost achieved substantially higher R² scores than Linear Regression and KNN, indicating that the ensemble approaches were better suited to the relationships represented by the selected agricultural and environmental features.

F. Actual vs Predicted Yield Analysis

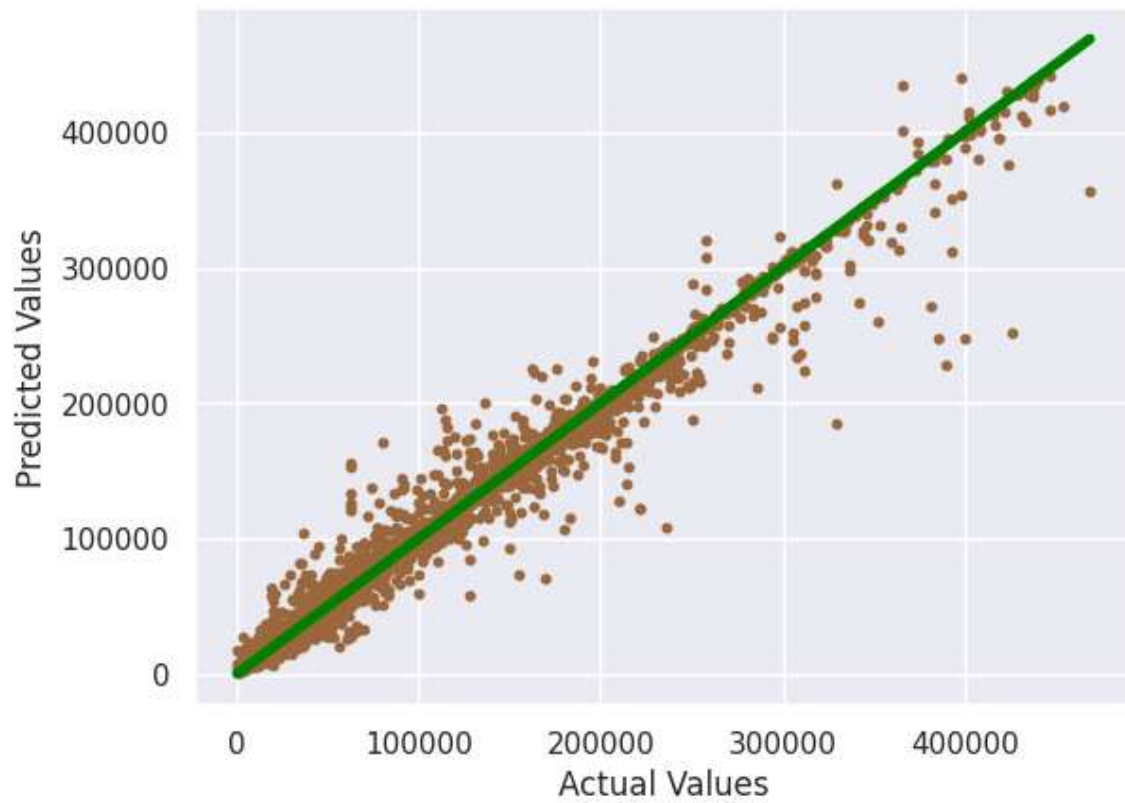


Figure 5. Actual versus predicted crop yield values.

The scatter plot illustrates the relationship between observed and predicted crop yield values.

The scatter plot shows that many predicted values are positioned close to the observed-value trend, indicating a strong correspondence between actual and predicted yields for a substantial portion of the test observations.

The deviations visible for some observations may reflect variability that is not fully represented by the selected predictors.

Overall, the visualization provides additional evidence of the predictive performance observed in the quantitative evaluation.

G. Discussion

The experimental findings demonstrate substantial differences in predictive performance among the evaluated machine learning models. The **Bagging Regressor achieved the highest test R^2 score of 0.9859**, while the Random Forest model achieved a closely comparable R^2 score of 0.9856. XGBoost also demonstrated strong performance with an R^2 score of 0.9743.

The relatively high predictive performance of the ensemble models can be associated with the characteristics of the selected dataset and the ability of tree-based ensemble methods to model nonlinear relationships and interactions among the input variables. The predictors include crop category, geographical area, year, rainfall, pesticide usage, and average temperature, which collectively provide substantial information about variations in the historical crop-yield observations. The Random Forest feature-importance analysis further indicates that crop category (Item), pesticide usage, average temperature, rainfall, and geographical region contribute to the prediction, while Year shows comparatively lower importance.

The authors also examined the experimental pipeline with respect to the possibility of target leakage. **Crop Yield was used as the dependent variable and was not included among the predictor features used for model training.** The reported test results were obtained from the specified 70:30 observation-level train-test split. However, because the experimental split was not grouped by geographical area, crop category, or year, related observations from the same or similar crop-area combinations may occur across the training and testing subsets. Therefore, the high R^2 values should be interpreted within the scope of the evaluation procedure used in this study and should not be considered equivalent to prediction on completely unseen regions or future agricultural years.

The 5-Fold Cross-Validation results provide an additional comparison of model performance across different partitions of the available observations. The Bagging Regressor obtained a mean cross-validation score of **0.9880**, followed by Random Forest at **0.9875** and XGBoost at **0.9775**. These results are consistent with the ranking observed in the held-out test evaluation.

The superior performance of the Bagging Regressor and Random Forest can be attributed to their ensemble-based tree-learning structures. Both approaches combine predictions from multiple decision trees, enabling them to represent nonlinear relationships and interactions among environmental, agricultural, geographical, and crop-related variables. The Bagging Regressor used **150 estimators** in the present experiment and aggregates predictions obtained from multiple estimators, which can reduce prediction variance. Random Forest similarly combines multiple randomized decision trees and benefits from ensemble aggregation. These characteristics are suitable for datasets in which the relationship between agricultural variables and crop yield is not adequately represented by a simple linear function. In contrast, Linear Regression provides a substantially more restricted functional form, while KNN depends strongly on the representation and structure of the feature space. The lower performance of these models in the present experiment therefore indicates that the ensemble tree-based approaches were better suited to the relationships represented in the selected dataset.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

This study developed and evaluated a machine learning-based framework for crop yield prediction using agricultural and environmental variables. A processed dataset containing crop, geographical, climatic, pesticide, and production information was used to compare seven regression algorithms: Linear Regression, Random Forest, Gradient Boosting, XGBoost, K-Nearest Neighbors, Decision Tree, and Bagging Regressor.

Model performance was assessed using R^2 , MSE, MAE, MAPE, and 5-Fold Cross-Validation. Among the evaluated models, the Bagging Regressor achieved the highest test R^2 of 98.59% and a mean cross-validation score of 98.80%. Random Forest and XGBoost also produced strong predictive results, while Linear Regression and KNN showed comparatively lower performance.

The results indicate that ensemble-based models were well suited to the relationships represented in the selected agricultural dataset. The findings also demonstrate that combining environmental and agricultural variables can provide useful information for crop yield estimation. However, the reported performance is based on the selected dataset and experimental configuration; validation using independent and more geographically diverse datasets would be necessary to assess the framework's performance in practical agricultural environments.

Overall, the study provides a comparative machine learning framework for crop yield forecasting and establishes a basis for further development of data-driven agricultural prediction systems.

B. Future Work

Although the proposed models produced strong predictive results, the current study is based on processed historical agricultural and environmental data. Several extensions can therefore be explored in future work.

1. Real-Time IoT Sensor Integration:

Future implementations could incorporate micro-weather sensors to collect localized measurements such as temperature, humidity, soil moisture, and rainfall for dynamic yield forecasting.

2. Satellite and Remote-Sensing Data:

Satellite imagery, vegetation indices such as NDVI, and geospatial information could be incorporated to provide additional spatial and crop-condition information.

3. Deep Learning Approaches:

Sequence-oriented and nonlinear deep learning architectures, including LSTM, ANN, and Transformer-based models, could be investigated for improved temporal crop forecasting.

4. Cloud-Based Deployment:

The framework could be extended to a cloud-based architecture capable of processing larger agricultural datasets and supporting scalable prediction services.

5. Mobile Decision-Support Applications:

A farmer-oriented application could be developed to present yield estimates, environmental information, and model-based insights in an accessible form.

6. Resource Optimization:

Future systems could combine yield forecasting with irrigation scheduling and other resource-management functions to support more efficient agricultural decision-making.

These extensions could transform the current offline prediction framework into a more adaptive system capable of incorporating real-time, spatial, and temporal agricultural information.

REFERENCES

- [1] J. H. Jeong, J. P. Resop, N. D. Mueller, J. C. Fleisher, K. Yun, E. E. Butler, et al., “Random Forests for Global and Regional Crop Yield Predictions,” PLoS ONE, vol. 11, no. 6, pp. 1–15, 2016.
- [2] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, “Accurate Prediction of Sugarcane Yield Using Machine Learning,” Agronomy for Sustainable Development, vol. 36, no. 2, pp. 1–13, 2016.
- [3] M. Shahhosseini, G. Hu, and S. Huber, “Machine Learning Approaches for Corn Yield Prediction Using Satellite and Climate Data,” Scientific Reports, vol. 10, pp. 1–15, 2020.
- [4] S. Khaki and L. Wang, “Crop Yield Prediction Using Deep Neural Networks,” Frontiers in Plant Science, vol. 10, pp. 1–13, 2019.
- [5] A. Gandhi, O. Armstrong, L. Petkar, and A. Tripathy, “Rice Crop Yield Prediction Using Machine Learning,” in Proc. IEEE Int. Conf. Technological Innovations in ICT for Agriculture and Rural Development, 2016, pp. 125–130.
- [6] F. Pedregosa et al., “Scikit-Learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [7] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
- [8] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.