Crop Yield Prediction in Maharashtra Using Machine Learning: A Comparative Analysis of Multiple Models

Janhvi Kirtane[1*]; Mangal Patil[2]; Shinde Vinayak[3]; Naresh Kambale[4]; Nakul Satao[5]

[1,5]M. Tech Agricultural Engineering Scholar, D. Y. Patil Agriculture & Technical University Talsande, Kolhapur.
[2]Head of the Deparatment Faculty of Agricultural Engineering, School of Engineering and Technology, D. Y. Patil Agriculture & Technical University Talsande, Kolhapur
[3]Assistant Professor, Faculty of Agricultural Engineering, School of Engineering and Technology, D. Y. Patil Agriculture & Technical University Talsande, Kolhapur.
[4]Head of the Department, Faculty of Computer Science, School of Engineering and Technology, D. Y. Patil Agriculture & Technical University Talsande, Kolhapur.

Corresponding Author: Janhvi Kirtane[1*]

Publication Date: 2026/09/15

Abstract: This study aims to develop and evaluate machine learning models for crop yield prediction using environmental and soil-related factors. The dataset includes key variables such as rainfall, temperature, humidity, soil pH, and crop production data collected from different regions. Various machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, and XGBoost, were implemented and compared to identify the most effective model for accurate yield prediction. The results indicate that ensemble models outperform traditional approaches, with XGBoost achieving the highest performance with an R^2 score of 0.86 and RMSE 224.59, followed by Random Forest with 0.85 and RMSE 238.11. Linear Regression R^2 score 0.573 and RMSE 404.75 showed comparatively lower performance due to its limitation in capturing non-linear relationships in agricultural data. Feature importance analysis revealed that rainfall, humidity, and temperature are the most influential factors affecting crop yield. Crop-wise analysis further demonstrated that prediction accuracy varies across different crops, with chickpea showing higher accuracy compared to rice and maize, which exhibited greater variability. The comparison between actual and predicted values confirms that the models are capable of capturing overall yield trends with reasonable accuracy. The findings highlight the effectiveness of machine learning techniques, particularly ensemble methods, in improving crop yield prediction. This study provides useful insights for agricultural planning, resource management, and decision-making, contributing towards sustainable farming practices.

Keywords: *Crop Yield Prediction, Machine Learning, Random Forest, XGBoost, Feature Importance, Agricultural Data.*

How to Cite: Janhvi Kirtane; Mangal Patil; Shinde Vinayak; Naresh Kambale; Nakul Satao (2026) Crop Yield Prediction in Maharashtra Using Machine Learning: A Comparative Analysis of Multiple Models. *International Journal of Innovative Science and Research Technology*, 11(9), 180-192. https://doi.org/10.38124/ijisrt/26sep024

I. INTRODUCTION

Agriculture plays a crucial role in economic development and food security, particularly in countries where a large portion of the population depends on farming. However, crop yield is highly influenced by environmental and climatic factors such as rainfall, temperature, humidity, and soil conditions. These factors interact in complex and non-linear ways, making accurate prediction of crop yield a challenging problem. In addition, recent changes in climate patterns have increased uncertainty in agricultural production, further emphasizing the need for reliable prediction systems (Addu *et al*., 2024).

Traditional methods of crop yield estimation are generally based on statistical analysis and historical averages. While these approaches are simple to apply, they often fail to capture the dynamic relationships among multiple influencing variables. As highlighted by Shalini et al. (2025), conventional techniques are limited in handling large-scale

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

and heterogeneous datasets and are not well-suited for modeling sudden environmental variations, which reduces their prediction accuracy.

With the advancement of data-driven technologies, machine learning has emerged as an effective approach for crop yield prediction. Machine learning models can process large datasets and identify complex patterns that are difficult to model using traditional techniques. Studies by Ramesh and Kumaresan (2025) show that algorithms such as Decision Trees, Random Forest, and XGBoost are capable of capturing non-linear relationships between environmental factors and crop yield, resulting in improved prediction performance.

Several researchers have also emphasized the effectiveness of ensemble learning techniques in agricultural prediction tasks. Ensemble models combine multiple learning algorithms to improve generalization and reduce overfitting. For example, Random Forest aggregates multiple decision trees to produce stable predictions, while boosting techniques such as XGBoost iteratively improve model performance. According to Ramesh and Kumaresan (2025), ensemble approaches often outperform individual models due to their ability to capture diverse patterns in the data.

Environmental variables remain key determinants of crop productivity. Rainfall influences water availability, temperature affects crop growth and development stages, and humidity impacts moisture conditions in the soil and atmosphere. In addition, soil parameters such as pH and nutrient levels also play a significant role in determining crop yield. Addu *et al.* (2024) highlighted that these environmental and soil-related factors are essential inputs for building accurate prediction models.

Despite the progress in machine learning applications, achieving consistent prediction accuracy across different crops and regions remains a challenge. Variability in climate conditions, differences in crop characteristics, and limitations in data quality can affect model performance. Therefore, it is important to evaluate multiple models and analyze their effectiveness under different conditions.

In this context, the present study aims to apply and compare various machine learning models, including Linear Regression, Decision Tree, Random Forest, and XGBoost, for crop yield prediction using environmental and soil-related features. The study also focuses on identifying the most influential variables affecting crop yield and evaluating model performance across different crops. The findings of this study can support better decision-making in agriculture and contribute to improved productivity and resource management.

The main objective of this study is to develop and evaluate machine learning models for crop yield prediction using environmental and soil-related factors.

II. LITERATURE REVIEW

Anakha Venugopal et al. [1] proposed a mobile-based application for crop prediction that identifies suitable crops and estimates yield using meteorological data such as temperature, humidity, and wind speed. The study applied Logistic Regression, Naïve Bayes, and Random Forest algorithms, where Random Forest achieved the highest accuracy. However, the absence of soil-related parameters limited the model's ability to capture soil influence on yield.

Thomas van Klompenburg *et al.* [2] conducted a systematic literature review (SLR) to analyze commonly used algorithms and features in crop prediction. Their findings showed that rainfall, temperature, and soil type are the most frequently used attributes, while Artificial Neural Networks (ANN) are widely adopted models. The study also highlighted RMSE as the most commonly used evaluation metric, but it lacked discussion on data quality and interpretability challenges.

Sonal Agarwal *et al.* [3] proposed a hybrid model combining Support Vector Machine (SVM) with deep learning techniques such as Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN). The hybrid approach achieved improved accuracy compared to traditional models. However, the study did not address computational complexity and model efficiency.

Addu *et al.* [4] developed a machine learning-based crop yield prediction system using environmental parameters such as rainfall, temperature, humidity, and pH. The study applied multiple models including Decision Tree and Random Forest, demonstrating that machine learning can effectively assist farmers in crop planning and decision-making.

Shalini *et al.* [5] proposed a machine learning framework using LSTM, CatBoost, and XGBoost for crop yield forecasting. The study highlighted the importance of combining deep learning and ensemble techniques to handle complex agricultural datasets. The results showed improved prediction accuracy, although the system required high computational resources.

Ramesh and Kumaresan [6] introduced a stacked ensemble model that combines multiple machine learning algorithms such as Linear Regression, Random Forest, and XGBoost. The model achieved high prediction accuracy with improved R^2 scores and reduced error metrics, demonstrating the effectiveness of ensemble approaches in agricultural prediction.

Jeong *et al.* [7] applied Random Forest models for crop yield prediction using environmental and soil data. Their study showed that Random Forest outperforms traditional regression models due to its ability to handle high-dimensional data and complex feature interactions.

Zhang *et al.* [8] explored deep learning techniques using Convolutional Neural Networks (CNN) for crop yield

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

prediction based on satellite imagery. The model demonstrated strong spatial prediction capability but faced limitations in handling temporal variations.

Patel and Ramesh [9] used XGBoost for weather-driven crop yield prediction and reported significant improvements in prediction accuracy compared to traditional models. The study emphasized the importance of climatic variables in influencing crop productivity.

Ma *et al.* [10] applied Long Short-Term Memory (LSTM) networks for time-series crop yield prediction. The results showed that LSTM effectively captures temporal dependencies in agricultural data, outperforming traditional statistical models.

Cedric *et al*. [11] developed a machine learning-based system for predicting crop yields at the country level using climatic and agricultural data. The study demonstrated that models such as k-Nearest Neighbors can provide reliable predictions for crops like maize and rice. However, the model's performance was sensitive to regional variations and data availability.

Batool *et al*. [12] compared the Aqua Crop simulation model with machine learning techniques for predicting tea yield. Their results showed that XGBoost outperformed traditional simulation approaches by achieving lower error values, indicating its effectiveness even with limited datasets.

Oikonomidis *et al*. [13] evaluated multiple machine learning and hybrid deep learning models, including CNN combined with XGBoost and RNN. The study highlighted that hybrid models can improve prediction accuracy, but they require high computational resources and complex implementation.

Kamath *et al*. [14] emphasized the importance of Random Forest for analyzing large agricultural datasets. The study showed that Random Forest is efficient in extracting hidden patterns from data and provides better prediction performance compared to simple models.

Hasan *et al.* [15] proposed a hybrid model combining K-Nearest Neighbors and Random Forest for crop yield prediction. The model achieved better accuracy compared to individual algorithms, demonstrating the benefit of combining multiple techniques.

Boppudi [16] introduced an optimized deep ensemble model for crop yield prediction using feature selection and optimization algorithms. The study reported improved accuracy and reduced prediction error, although the approach increases model complexity.

Shahhosseini *et al.* [17] developed an ensemble framework for crop yield prediction using weather data across multiple regions. The study showed that combining multiple models improves prediction reliability, especially when dealing with seasonal variability.

Lu *et al.* [18] proposed a stacking ensemble model with attention mechanisms to improve prediction accuracy. The results demonstrated significant improvements in performance metrics such as RMSE and correlation coefficient, indicating the effectiveness of advanced ensemble techniques.

Khaki *et al*. [19] developed a deep learning framework combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for crop yield prediction. The model effectively captured both spatial and temporal features, resulting in improved prediction performance.

Mahdizadeh Gharakhanlou and Perez [20] applied ensemble machine learning models such as AdaBoost, Gradient Boosting, and Random Forest for yield prediction. Their findings showed that XGBoost achieved the best performance among all models, highlighting the strength of boosting techniques in agricultural applications.

III. MATERIALS AND METHODS

➤ *Study Area*

This study focuses on Maharashtra a major agricultural state in western India. Farming is an important activity in this state, but conditions are not the same everywhere. The Maharashtra state lies roughly between 15° and 22° north latitude and 72° to 80° east longitude. Because of its size and geography, Maharashtra shows clear differences in climate, soil types, and cropping patterns across districts.

The state can be broadly grouped into regions such as Konkan, Western Maharashtra, Marathwada, and Vidarbha. Each of these areas experiences different weather behavior. For example, the Konkan belt receives very high rainfall every year, often above 2000 mm, while Marathwada and parts of Vidarbha face irregular and sometimes low rainfall. This uneven distribution of rain creates uncertainty in farming and directly affects crop output.

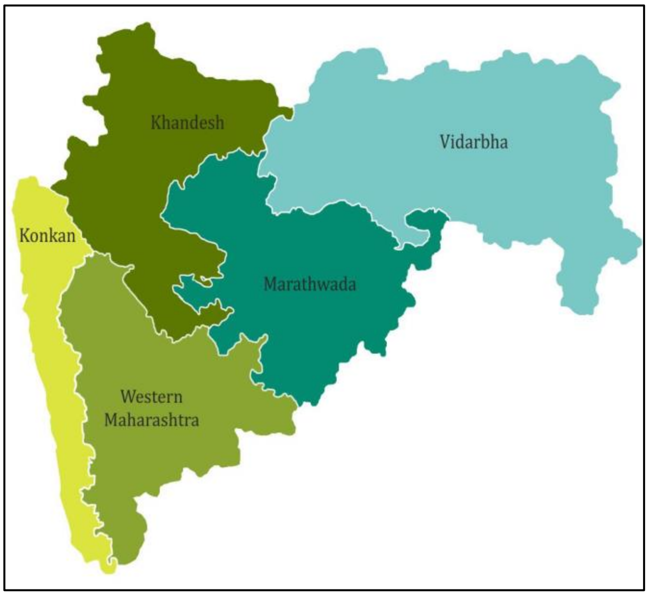

Fig 1 Geographical Location of Study Area

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

Temperature conditions also vary across the state. In many interior areas, summer temperatures can rise beyond 40°C, whereas winters are generally moderate. Such temperature changes influence important stages of crop growth, including germination, flowering, and grain development.

Soil conditions further add to this variation. A large portion of Maharashtra is covered with black cotton soil, which is known for its ability to retain moisture and is well suited for crops like cotton and soybean. Other soil types, such as alluvial and lateritic soils, are also found in different regions and support crops like rice, sugarcane, and pulses.

Farmers grow different crops such as rice, maize, chickpea, cotton, and sugarcane. However, crop yield is not the same every year because it depends on many factors like weather and soil conditions.

Because of these variations, Maharashtra is a suitable region for crop yield prediction using machine learning. The differences in conditions help the model learn patterns and improve prediction accuracy.

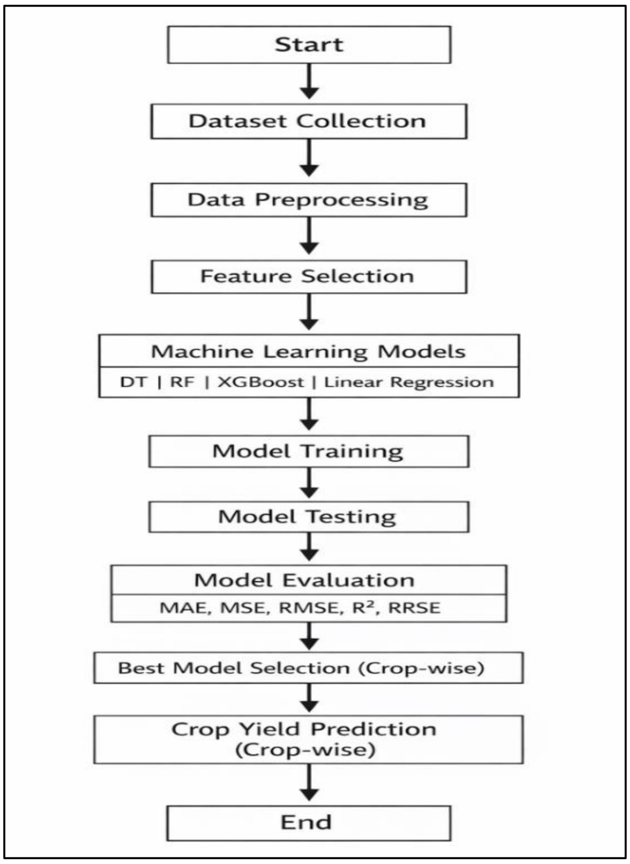

Fig 2 Framework for Crop Yield Prediction Using Machine Learning

➢ *Collection of Data*

The data used in this study was collected from publicly available agricultural and weather sources. Agricultural data was obtained from the ICRISAT dataset available on Kaggle, while weather data was collected from the India Meteorological Department (IMD).

The dataset includes district-wise and year-wise information such as cultivated area, production, and crop yield, along with environmental factors like temperature, rainfall, and humidity. Both datasets were combined to form a single dataset for analysis.

The study focuses on major crops such as rice, maize, chickpea, and cotton. The final dataset is organized in tabular form, where each row represents a specific record and the columns represent input features and crop yield as the target variable.

Python and Jupyter Notebook were used for data analysis and model development, along with libraries such as Pandas, NumPy, Scikit-learn, and XGBoost.

➢ *Data Preprocessing*

Data preprocessing was carried out to improve the quality of the dataset before applying machine learning models. Since the data was collected from different sources, it contained missing values, duplicate records, and inconsistencies.

Initially, duplicate entries and incorrect values were removed to ensure data reliability. Missing values in numerical features were filled using mean or median values, while categorical data was handled using the most frequent value.

Categorical variables such as crop type and district were converted into numerical form using label encoding. Feature scaling techniques such as normalization and standardization were applied to bring all variables to a similar range.

Outliers were identified using statistical methods like Z-score and IQR, and were either removed or adjusted to reduce their impact on the model.

These preprocessing steps helped in improving data quality and ensured better performance of the machine learning models. Proper preprocessing significantly improves model accuracy by reducing noise and inconsistencies in the dataset.

➢ *Feature Selection*

Feature selection is an important step in this study because not all variables are equally useful for predicting crop yield. If too many unnecessary features are included, the model can become complex and may not give accurate results.

In this work, only those features were selected which have a clear effect on crop growth. These include environmental factors such as temperature, rainfall, and humidity, along with soil properties like pH and nutrient levels (nitrogen, phosphorus, and potassium).

To understand the importance of each feature, models like Random Forest and XGBoost were used. These models help in identifying which variables contribute more to the prediction. Based on this, less important features were removed from the dataset.

This process helped in simplifying the data and improving the overall performance of the model. It also reduced the chances of overfitting and made the predictions more reliable.

➢ *Machine Learning Models*

In this study, four machine learning models were used to predict crop yield. These models were selected because they can handle relationships between environmental, soil, and agricultural factors.

- *Decision Tree*

The Decision Tree model works by dividing the data into smaller groups based on different conditions. It makes decisions step by step using input features such as rainfall, temperature, and soil properties.

This model is easy to understand, but it may sometimes fit the training data too closely, which can reduce its performance on new data.

It works by selecting the best feature that minimizes error at each step.

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Where $p_i$is the probability of class i.

- *Random Forest*

Random Forest is an improved version of the Decision Tree model. It builds multiple trees using different parts of the dataset and combines their results to make a final prediction.

This approach reduces errors and gives more stable and reliable results. In this study, Random Forest performed better compared to other models.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Where $N$is the number of trees.

- *XGBoost*

XGBoost is an advanced model that builds trees one after another. Each new tree focuses on correcting the errors of the previous one.

It can handle complex relationships in the data and usually provides high accuracy. However, it requires proper tuning to achieve the best performance.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where $l$is the loss function and Ωis the regularization term.

- *Linear Regression*

Linear Regression is a basic machine learning model that shows the relationship between input variables and crop yield using a straight line.

It works well when the relationship between variables is simple and linear. However, in agricultural data, relationships are often complex, so its performance is usually lower compared to models like Random Forest and XGBoost.

In this study, Linear Regression was mainly used as a baseline model to compare the performance of other machine learning algorithms.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where:

y= predicted crop yield

x= input features (rainfall, temperature, etc.)

$beta$= coefficients

This model works well for simple relationships but may not perform well for complex agricultural data.

➢ *Model Training and Testing*

Training and testing of models is a key part of this study. In this step, the models are first trained using the available dataset and then checked on new data to see how well they perform.

The dataset was split into two parts: 80% of the data was used for training and the remaining 20% was used for testing. The training data helps the models understand how input factors such as temperature, rainfall, humidity, soil pH, and nutrients are related to crop yield.

Once the models are trained, they are tested using the data that was not included in the training process. This helps in evaluating how well the models can predict results for new and unseen data.

To get better results, some important parameters were adjusted. For example, the depth of the tree in Decision Tree, the number of trees in Random Forest, and the learning rate in XGBoost were tuned.

This step helped in improving the model performance and reducing the chances of overfitting. Overall, the training and testing process made the models more accurate and reliable for crop yield prediction.

➢ *Model Evaluation*

To evaluate the performance of machine learning models, different statistical metrics were used. These metrics compare predicted crop yield values with actual values and help in measuring model accuracy.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

In this study, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R^2 score were used, as they are widely applied in regression-based machine learning tasks.

- *Mean Absolute Error (MAE)*

$$MAE = \frac{1}{n}\sum | y_i - \hat{y}_i |$$

MAE measures the average difference between actual and predicted values. Recent work by Aurélien Géron (2022) highlights that MAE is simple and effective for understanding prediction error in regression models.

- *Mean Squared Error (MSE)*

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

MSE calculates squared differences between actual and predicted values, giving more importance to large errors. According to Trevor Hastie *et al.* (2021), MSE is useful when large errors must be penalized more strongly.

- *Root Mean Square Error (RMSE)*

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

RMSE is the square root of MSE and is expressed in the same unit as the output variable.

Recent studies show that RMSE is commonly used to compare machine learning models in prediction tasks (Géron, 2022).

- *R^2 Score (Coefficient of Determination)*

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

R^2 score shows how well the model explains variation in the data. A value close to 1 indicates good performance. As explained by Gareth James *et al.* (2021), R^2 is a standard measure for evaluating regression models.

Using multiple evaluation metrics provides a better understanding of model performance research that combining MAE, RMSE, and R^2 gives more reliable results compared to using a single metric (Géron, 2022; James *et al.*, 2021).

IV. RESULTS AND DISCUSSION

The performance of different machine learning models was evaluated using statistical metrics such as MAE, RMSE, and R^2 score. These metrics helped in comparing the predicted crop yield values with the actual values.

In 4.1 show the histogram to observe how yield values are spread in the dataset.

From the graph, it can be seen that most of the values fall between 400 kg/ha and 1200 kg/ha, which means a large number of records are within this range. As the yield increases beyond 2000 kg/ha, the number of observations becomes smaller.

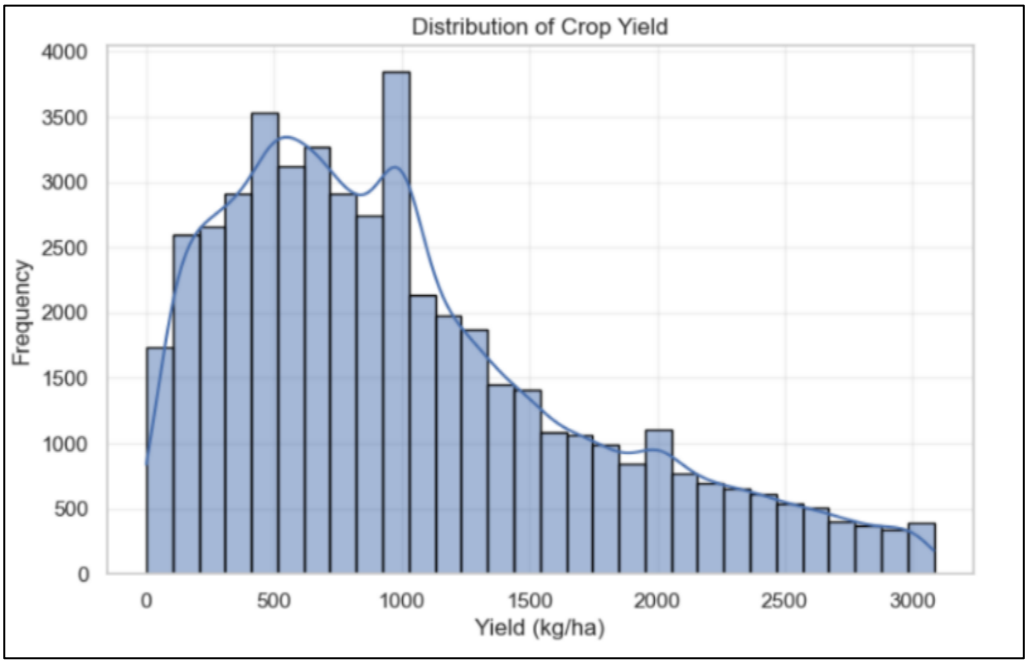

Fig 3 Distribution of Crop Yield (kg/ha)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

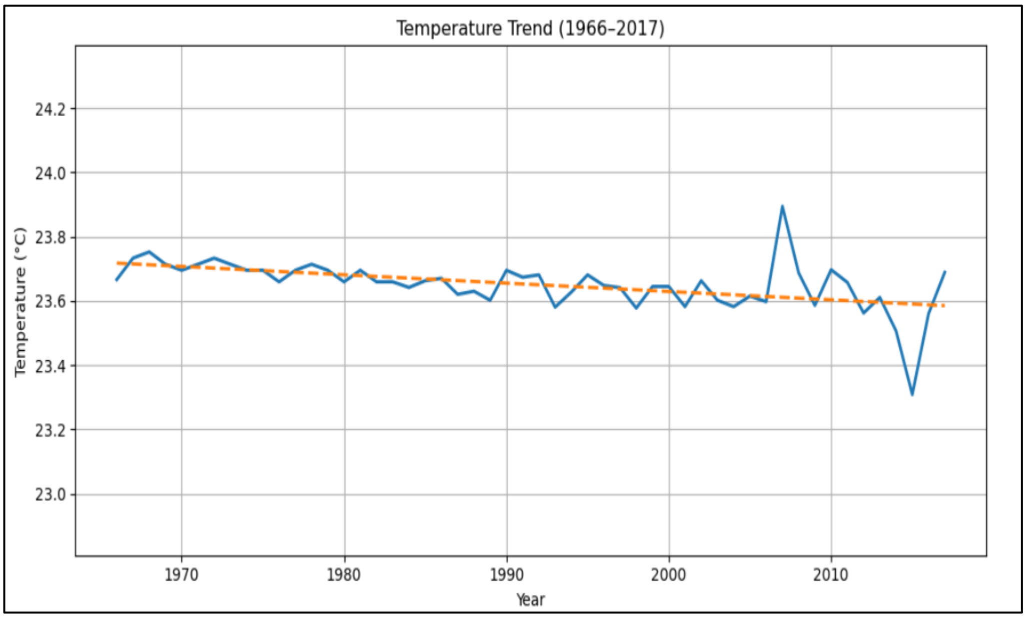

Fig 4 Temperature Trend (1966–2017)

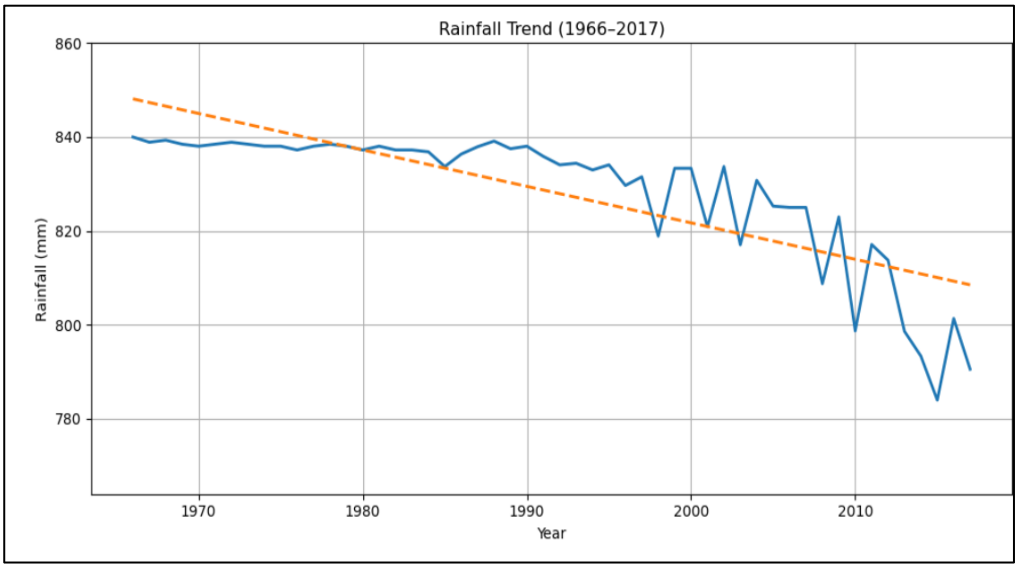

Fig 5 Rainfall Trend (1966–2017)

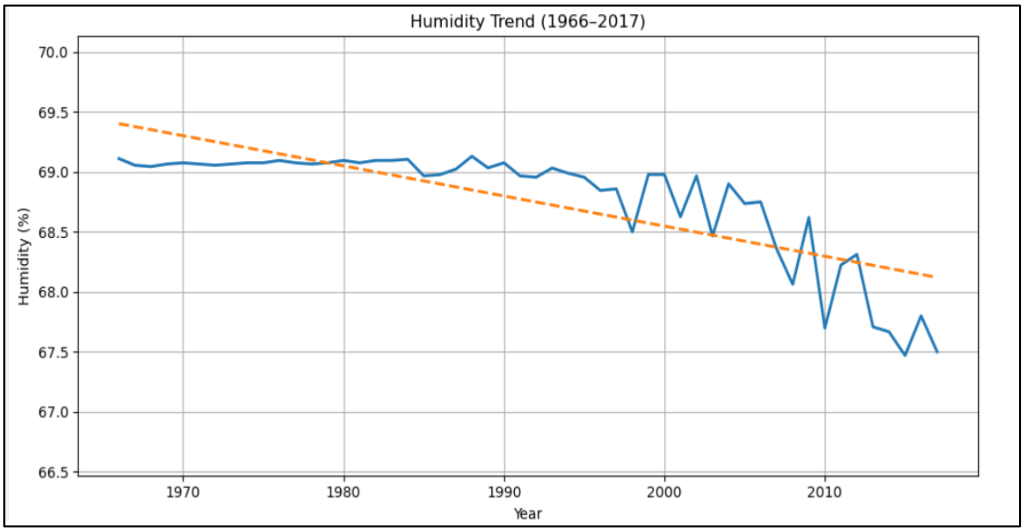

Fig 6 Humidity Trend (1966–2017)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

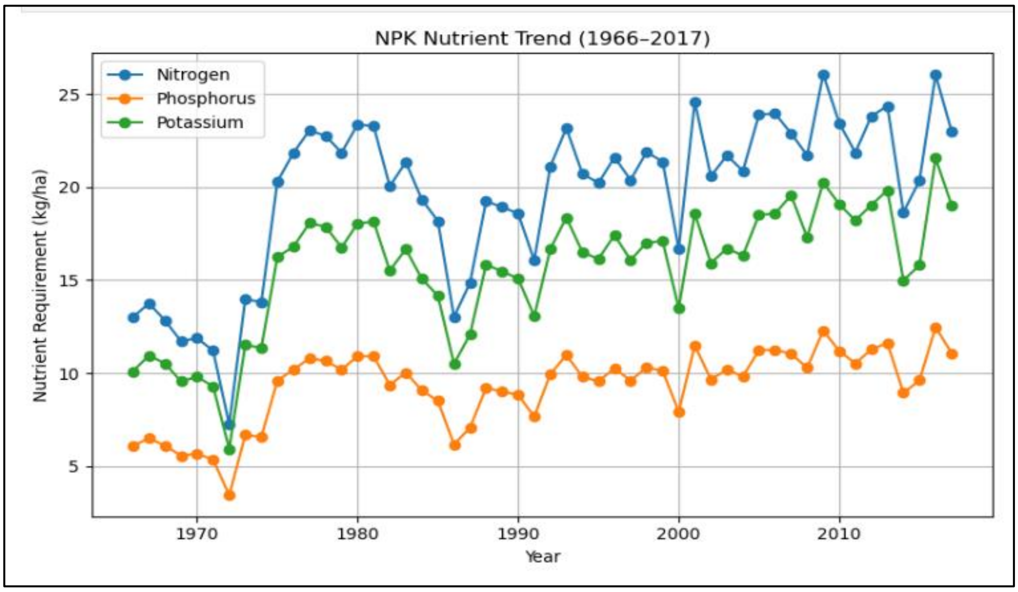

Fig 7 NPK Nutrient Trend (1966–2017)

This pattern shows that medium yield values are more common, while very high yields occur less frequently. The longer tail on the higher side of the graph indicates that there is variation in the data.

Such variation may be due to changes in environmental factors like rainfall, temperature, and soil conditions across different regions and crop yield values are not evenly distributed, and this variation makes the dataset suitable for machine learning-based analysis.

In the fig 4 the temperature trend from 1966 to 2017. It can be observed that temperature values remain relatively stable over the study period, with only minor fluctuations around a consistent range.

In the earlier years, temperature variations are minimal, indicating relatively stable climatic conditions. In the later years, some fluctuations are visible, including occasional increases and decreases. However, these variations are not significant and remain within a narrow range.

The overall trendline suggests a very slight decreasing pattern, although the change is minimal and does not indicate a strong long-term shift in temperature.

This analysis indicates that temperature has remained relatively stable during the study period, with minor year-to-year variations. Even small variations in temperature can influence crop growth stages such as germination, flowering, and grain development, and therefore remain an important factor in crop yield prediction.

In the Fig 5 show the rainfall trend from 1966 to 2017 shows a gradual decline over the years. In the earlier period, rainfall values remain relatively stable with only small variations. However, after the 1990s, the fluctuations become more noticeable, with several years showing a drop in rainfall levels.

The downward slope of the trend line indicates an overall reduction in rainfall across the time period. This decrease suggests that water availability for agriculture may have reduced over time, which can directly impact crop production.

The irregular fluctuations in rainfall also highlight the uncertainty in weather patterns. Such variability makes it difficult to maintain consistent crop yields, as agriculture is highly dependent on rainfall.

In the Fig 6 show the humidity trend from 1966 to 2017 shows a gradual decrease over time. In the earlier years, humidity levels remain relatively stable with only minor variations. However, after the late 1990s, the fluctuations become more noticeable, and a slight downward pattern can be observed.

The overall declining trend suggests that moisture levels in the atmosphere have reduced over the years. Since humidity plays an important role in crop growth and soil moisture retention, such changes may influence agricultural productivity.

The presence of fluctuations indicates that humidity levels are not constant and may vary due to changing climatic conditions. This variability adds uncertainty in predicting crop yield, as consistent environmental conditions are important for stable crop production.

In the Fig 7 show the trend of NPK nutrients from 1966 to 2017 shows noticeable variation over time. Nitrogen levels are consistently higher compared to phosphorus and potassium, indicating its greater requirement in crop production. Phosphorus remains the lowest among the three nutrients, while potassium shows moderate values throughout the period.

Over the years, all three nutrients display fluctuations rather than a constant trend. However, a gradual increase can

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

be observed in nutrient requirements, especially after the 1980s. This suggests a growing dependency on fertilizers to maintain or improve crop yield.

The variations in nutrient levels may be influenced by changes in soil conditions, cropping patterns, and agricultural practices. These fluctuations indicate that nutrient management plays an important role in crop productivity and needs to be considered carefully in yield prediction models.

➢ *Model Performance Analysis*

In this study, the performance of different machine learning models was evaluated to determine their effectiveness in predicting crop yield. Since agricultural data involves complex and non-linear relationships between environmental factors and crop production, it is important to use multiple models and compare their results.

The models used in this analysis include Linear Regression, Decision Tree, Random Forest, and XGBoost. Each model has a different approach to learning patterns from the data. Linear Regression assumes a linear relationship, while Decision Tree and ensemble methods such as Random Forest and XGBoost are capable of capturing more complex and non-linear relationships.

To assess the performance of these models, evaluation metrics such as R^2 score, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Root Relative Squared Error (RRSE) were used. These metrics provide a comprehensive understanding of model accuracy and prediction error.

Table 1 shows the performance comparison of different machine learning models used for crop yield prediction.

Table 1 Performance Comparison of Machine Learning Models

Model	R^2 Score	MAE	MSE	RMSE	RRSE
Linear Regression	0.573	286.07	163830.10	404.75	0.653
Decision Tree	0.755	191.02	94073.75	306.71	0.494
Random Forest	0.852	152.47	56699.46	238.11	0.384
XGBoost	0.869	144.84	50440.87	224.59	0.362

Among all models, XGBoost achieved the highest R^2 score of 0.86 and the lowest error values, indicating superior prediction accuracy. Random Forest also performed well with an R^2 score of 0.85. Decision Tree showed moderate performance, while Linear Regression produced comparatively lower accuracy due to its inability to capture non-linear relationships in the data.

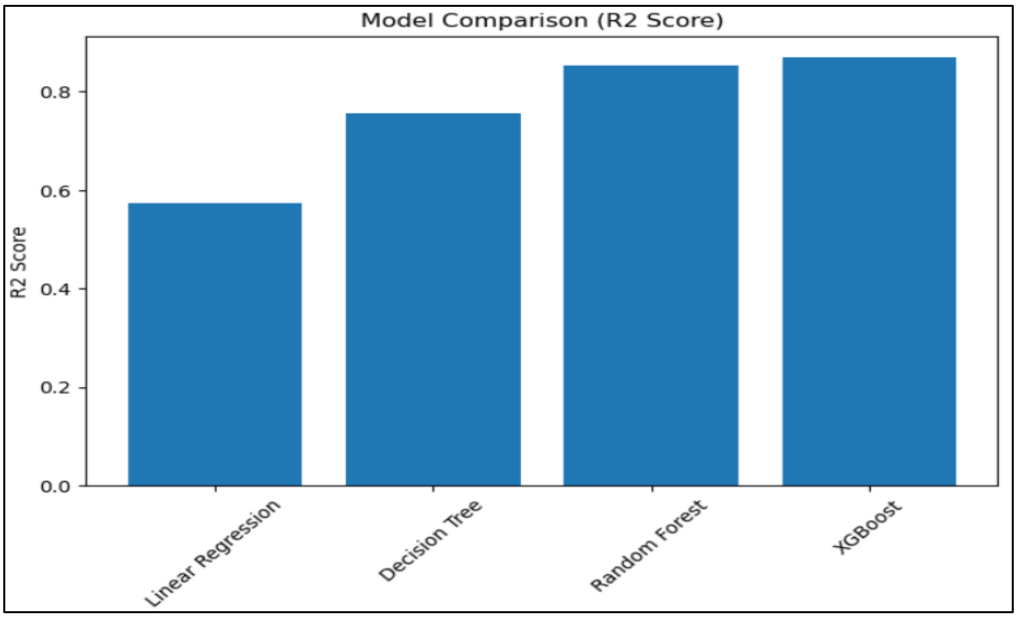

Fig 8 Model Performance Comparison

Figure 8 visually represents the performance of different models. It can be observed that XGBoost and Random Forest outperform other models, indicating their effectiveness in handling complex agricultural data.

These results indicate that ensemble models are more effective for crop yield prediction due to their ability to handle complex patterns in the data.

➢ *Feature Importance Analysis*

Feature importance analysis was carried out using the Random Forest model to identify how different input variables contribute to crop yield prediction. This analysis helps in understanding which factors the model relies on the most while making predictions. Instead of treating all variables equally, the model assigns different levels of importance based on their influence on the output.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

From the results, it can be observed that environmental factors such as rainfall, humidity, and temperature have a strong influence on crop yield. Rainfall plays a key role because it directly affects water availability for crops. In regions or years where rainfall is sufficient and well-distributed, crop growth tends to be better. On the other hand, irregular or low rainfall can lead to reduced yield, which explains its higher importance in the model.

Humidity is another important factor, as it is closely related to moisture conditions in the environment. Proper humidity levels support plant growth and help maintain soil moisture, whereas very low or very high humidity can negatively affect crop development. This makes humidity a significant variable in predicting yield outcomes.

Temperature also contributes to the model, although its impact may vary depending on crop type and growth stage. Temperature influences processes such as germination, flowering, and grain formation. Even small changes in temperature during critical stages can affect the final yield, which is why it is considered an important feature.

The feature importance results indicate that crop yield is not determined by a single factor, but by the combined effect of multiple environmental conditions. This highlights the need to consider several variables together when building machine learning models for agricultural prediction.

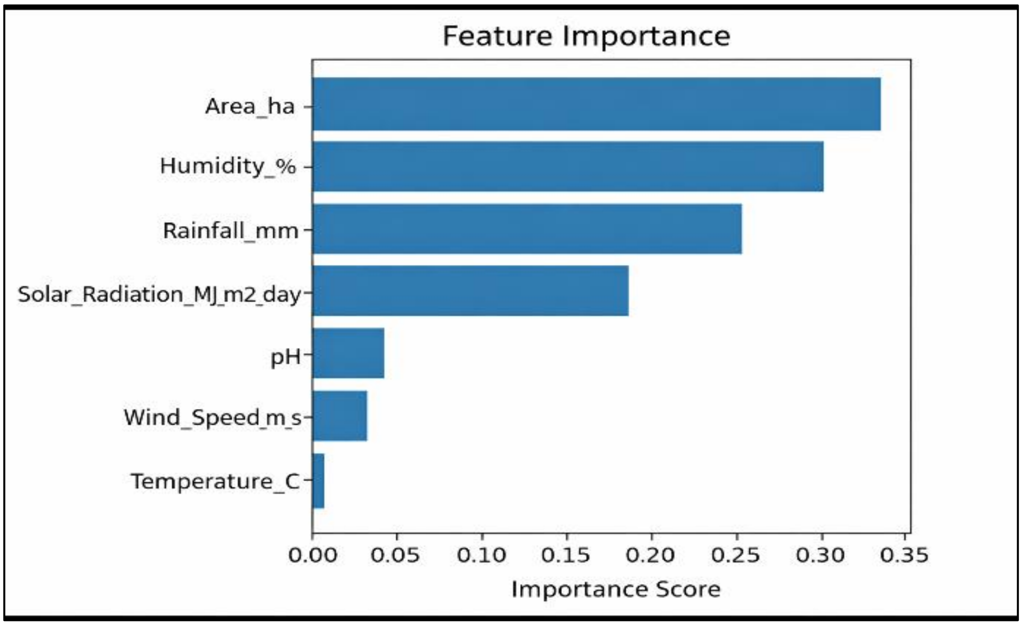

Fig 9 Feature Importance of Input Variables in Crop Yield Prediction

In the fig 9 show the feature importance analysis highlights the relative contribution of different input variables in predicting crop yield. It can be observed that Area (Area_ha) has the highest importance score, indicating that the size of cultivated land plays a major role in determining total crop yield.

Among the environmental factors, humidity and rainfall show high importance, suggesting that moisture availability and water supply are critical for crop growth. These factors directly influence plant development and overall productivity, which explains their strong contribution to the model.

Solar radiation has moderate importance, indicating its role in photosynthesis and energy availability for crops. However, its impact is lower compared to rainfall and humidity.

On the other hand, variables such as pH, wind speed, and temperature show relatively lower importance scores. This suggests that, within the given dataset, these factors have a smaller influence on yield prediction compared to other dominant variables.

Overall, the results indicate that crop yield is primarily influenced by land area and key environmental factors, while other variables play a supporting role. This demonstrates that multiple factors together contribute to accurate prediction rather than a single dominant variable.

- *Prediction Analysis*

In the fig 10 show the scatter plot shows the relationship between actual and predicted crop yield values obtained using the XGBoost model. It can be observed that most of the data points lie close to the diagonal reference line, indicating that the model predictions are reasonably accurate.

The diagonal line represents perfect prediction, where the predicted values are equal to the actual values. Points that are closer to this line indicate better prediction performance.

Some deviations from the line can be observed, especially at higher yield values, which suggests that the model has minor prediction errors in certain cases. These variations are expected due to the complexity and variability present in agricultural data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

The XGBoost model is able to capture the general pattern of crop yield effectively, demonstrating good predictive capability.

This confirms that the model performs well in estimating crop yield, although slight variations remain due to environmental uncertainties.

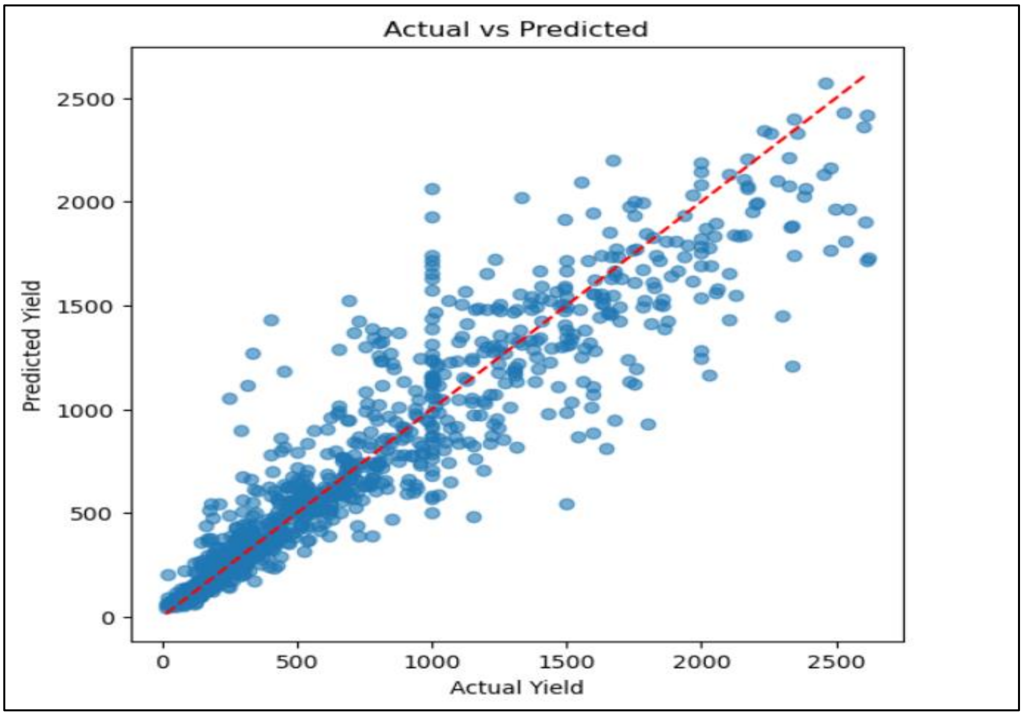

Fig 10 Actual vs Predicted Crop Yield

➢ *Crop-Wise Analysis*

The crop-wise analysis shows that model performance varies across different crops. It can be observed that crops such as chickpea achieve higher prediction accuracy, indicating more stable patterns in the data. In contrast, crops like rice and maize show relatively lower performance due to higher variability in environmental conditions.

This variation suggests that crop yield prediction is not uniform across all crops, and different crops respond differently to environmental and soil factors. Therefore, crop-specific modeling approaches may further improve prediction accuracy.

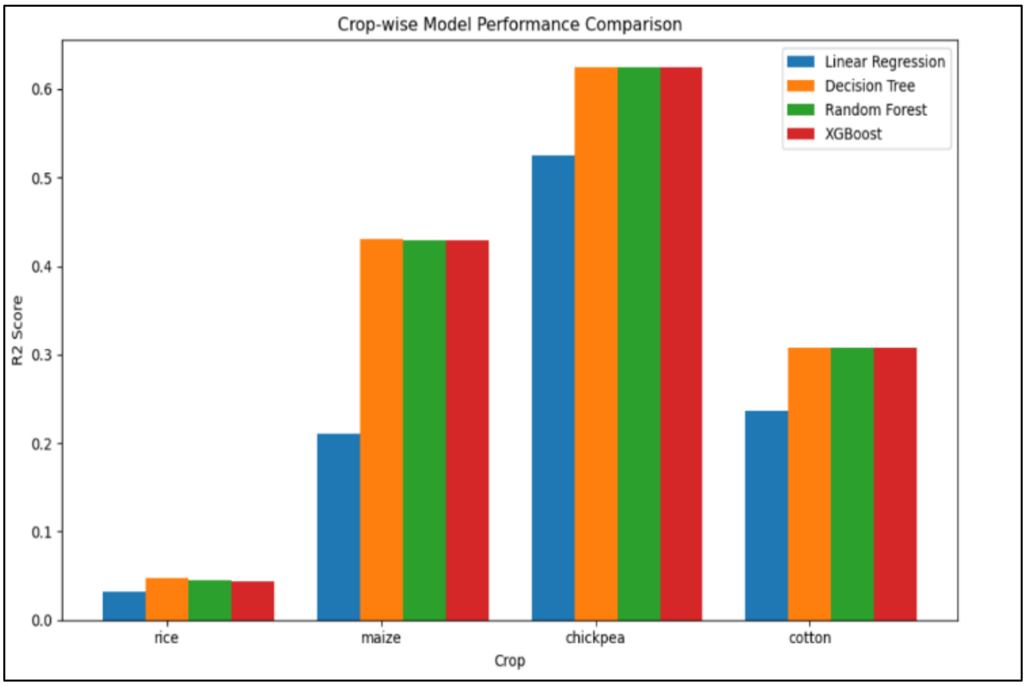

Fig 11 Crop-wise Model Performance Comparison using R² Score

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

In the fig 11. x show the crop-wise model performance comparison shows how different machine learning models perform across various crops based on the R^2 score. It can be observed that model performance varies significantly depending on the crop type.

Among all crops, chickpea shows the highest prediction accuracy across all models, with R^2 values exceeding 0.6. This indicates that chickpea yield follows more stable and predictable patterns, making it easier for machine learning models to learn and predict accurately.

For maize, the performance is moderate, with Decision Tree, Random Forest, and XGBoost showing similar results, while Linear Regression performs comparatively lower. This suggests that non-linear models are better suited for capturing patterns in maize yield data.

In the case of cotton, the performance is relatively lower than chickpea and maize, but still acceptable. Ensemble models such as Random Forest and XGBoost perform better than Linear Regression, indicating the presence of complex relationships in the data.

Rice shows the lowest prediction accuracy among all crops, with very low R^2 values. This suggests that rice yield is highly variable and influenced by multiple external factors, making it difficult for the models to capture consistent patterns.

Overall, the results indicate that ensemble models (Random Forest and XGBoost) consistently perform better across most crops, while Linear Regression shows the weakest performance. This highlights the importance of using advanced models for handling complex agricultural datasets.

These findings confirm that crop-specific variability plays a crucial role in prediction performance, and selecting appropriate models is essential for accurate crop yield forecasting.

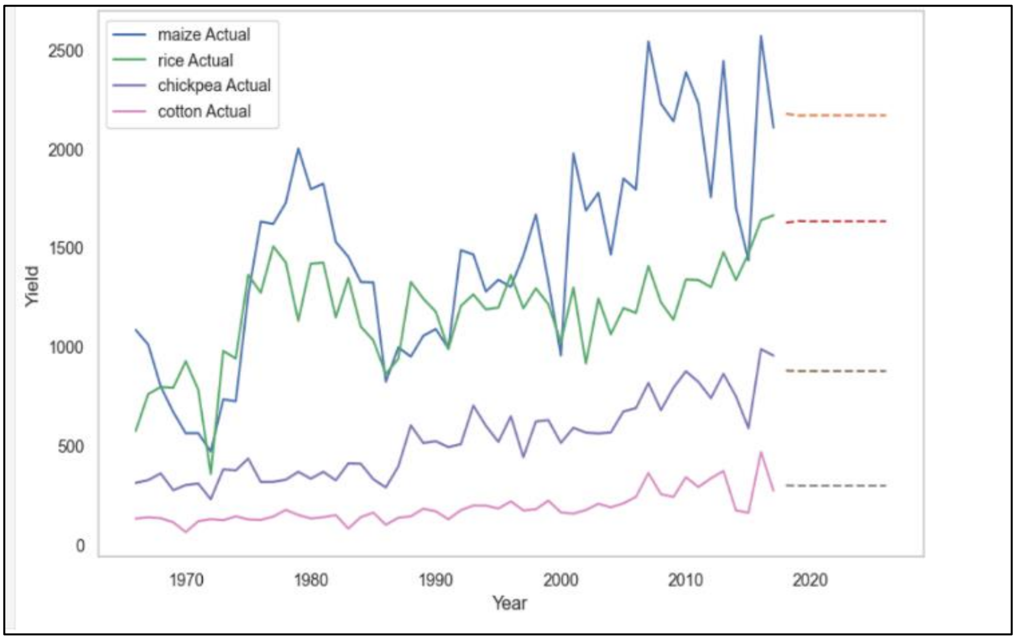

Fig 12 Historical and Predicted Crop Yield Trends for Major Crops (1966–2026)

In fig 12 graph shows the historical and predicted crop yield trends for major crops. The solid lines represent the actual yield values, while the dashed lines indicate future predictions up to the year 2026.

It can be observed that crops such as maize and rice show relatively higher yield values compared to chickpea and cotton. Over the years, an increasing trend is visible in most crops, although fluctuations are also present due to environmental variability.

The predicted values suggest that yield may remain stable or show gradual changes in the coming years. However, these predictions are based on historical patterns and may vary depending on future climatic and environmental conditions.

The graph provides a clear understanding of past trends and expected future behavior of crop yield, which can be useful for agricultural planning and decision-making.

V. CONCLUSIONS

This study focuses on applying machine learning techniques to predict crop yield using environmental and soil-related factors. Different models such as Linear Regression, Decision Tree, Random Forest, and XGBoost were used to analyze how well they can learn patterns from agricultural data.

Among these models, XGBoost showed the best overall performance, with Random Forest also giving strong results. This indicates that ensemble-based methods are more effective in handling complex and non-linear relationships

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep024

compared to simpler models. On the other hand, Linear Regression performed relatively lower, which suggests that linear models may not be sufficient for capturing real-world variations in crop yield.

The study also shows that environmental factors like rainfall, humidity, and temperature have a strong impact on crop production. The feature importance analysis highlights that these variables contribute more significantly compared to other factors included in the dataset.

When analyzing results crop-wise, it is clear that prediction accuracy differs across crops. Chickpea shows better prediction results, which may be due to more consistent patterns in the data. In contrast, crops such as rice and maize show lower accuracy, indicating higher variability and influence of external conditions.

The comparison between actual and predicted values suggests that the models are able to capture overall trends effectively, although some differences are observed in certain cases. These variations are expected because agricultural systems are influenced by many dynamic and unpredictable factors.

In addition, future predictions indicate that yield trends may change differently for each crop. These predictions should be considered as estimates based on past data and not exact outcomes, as real-world conditions can vary over time.

Results of this study show that machine learning, especially ensemble methods, can be effectively used for crop yield prediction. The findings can help in better planning and decision-making in agriculture, particularly in managing resources and improving productivity.

REFERENCES

[1]. A. Venugopal, et al., “Crop prediction and yield estimation using machine learning techniques,” International Journal of Agricultural Research, 2023.

[2]. T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Computers and Electronics in Agriculture, vol. 177, 2020.

[3]. S. Agarwal, et al., “Hybrid machine learning and deep learning approach for crop yield prediction,” Journal of Agricultural Informatics, 2024.

[4]. S. Addu, et al., “Crop yield prediction using machine learning techniques with environmental parameters,” Computers, vol. 13, no. 137, 2024.

[5]. P. Shalini, et al., “Crop yield forecasting using LSTM, CatBoost and XGBoost models,” Journal of Smart Agriculture Systems, 2025.

[6]. M. Ramesh and S. Kumaresan, “Stacked ensemble model for crop yield prediction,” Agricultural Data Science Journal, 2025.

[7]. J. H. Jeong, J. P. Resop, N. D. Mueller, et al., “Random forests for global and regional crop yield prediction,” PLOS ONE, vol. 11, no. 6, 2016.

[8]. L. Zhang, Z. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on CNN-based crop prediction,” IEEE Geoscience and Remote Sensing Magazine, 2018.

[9]. H. Patel and M. Ramesh, “Weather-based crop yield prediction using XGBoost algorithm,” International Journal of Agricultural Technology, 2020.

[10]. J. Ma, Y. Zhang, and L. Chen, “Time-series crop yield prediction using LSTM networks,” IEEE Access, vol. 8, 2020.

[11]. C. Cedric, et al., “Country-level crop yield prediction using machine learning techniques,” Agricultural Systems, 2019.

[12]. B. Batoolet al., “Comparison of Aqua Crop and machine learning models for tea yield prediction,” Agricultural Water Management, 2021.

[13]. M. Oikonomidis, et al., “Hybrid deep learning models for crop yield prediction,” Computers and Electronics in Agriculture, 2022.

[14]. S. Kamath, et al., “Application of Random Forest in agricultural data analysis,” International Journal of Data Science in Agriculture, 2020.

[15]. M. Hasan, et al., “Hybrid KNN and Random Forest model for crop yield prediction,” Journal of Agricultural Engineering, 2021.

[16]. R. Boppudi, “Optimized deep ensemble model for crop yield prediction,” Expert Systems with Applications, 2022.

[17]. M. Shahhosseini, et al., “Ensemble machine learning for crop yield prediction using weather data,” Agronomy Journal, 2020.

[18]. Y. Lu, et al., “Stacking ensemble with attention mechanism for crop yield prediction,” IEEE Transactions on Neural Networks, 2021.

[19]. S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” Nature Scientific Reports, 2019.

[20]. A. Mahdizadeh Gharakhanlou and L. Perez, “Crop yield prediction using ensemble learning algorithms,” Remote Sensing, 2020.