SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I1P118

Volume 12 Issue 1, 236-244, January 2025

Original Article

Explainable Model for Agricultural Crop Yield Prediction in Indian Conditions with SHAP Analysis

Yogita Dubey[1], Aniket Sakhare[2], Atharva Tasare[3], Santosh Kakad[4], Roshan Umate[5]

[1,2,3,4]*Department of Electronics and Telecommunication Engineering, Yeshwantaro Chavan College of Engineering, Maharashtra, India.*
[5]*Department of Research and Development, Datta Meghe Institute of Higher Education and Research, Maharashtra, India.*

[1]*Corresponding Author : yogeetakdubey@yahoo.co.in*

Received: 22 November 2024 Revised: 28 December 2024 Accepted: 14 January 2025 Published: 30 January 2025

Abstract *- This paper presents an ensemble-based approach to agricultural crop yield forecasting, focusing on the Indian context. The study integrates different forecasting models to improve forecasting accuracy for agricultural complexity data analysis. SHAP (SHapley Additive exPlanations) is used to give a clear idea of the contribution of each factor in the prediction model to improve model interpretation. The dataset used for this study contains yields of 55 crops over 6 seasons in 30 countries in 23 years (1997) -2020 available). Besides demonstrating the method's effectiveness, it emphasizes the need for explicit modeling that can provide valuable insights for better agricultural practices and ultimately contribute to higher yields that will be sustainable in Indian agriculture.*

Keywords *- Ensemble learning, Explainable AI, Yield prediction, Quantitative assessment, SHAP analysis.*

1. Introduction

Agricultural productivity is a cornerstone of economic stability and food security, particularly in agrarian economies like India. Accurate crop yield prediction is crucial for planning and decision-making, impacting everything from farmer income to national food supply. Adopting Machine Learning (ML) for agricultural crop yield prediction has gained momentum, particularly in regions like India, where agriculture plays a pivotal role in the economy. ML techniques offer advanced tools to analyse complex agricultural data, improving yield predictions and aiding in better decision-making for farmers and policymakers. Various ML techniques have been explored for their effectiveness in predicting crop yields. Due to the diverse agroclimatic conditions in India, region-specific studies have also been conducted to tailor machine learning models to local conditions [1-4].

Various Machine Learning techniques have been explored for their effectiveness in predicting crop yields. These techniques range from traditional models like linear regression to advanced neural networks and ensemble methods [5-7]. Effective crop yield prediction relies heavily on the quality and relevance of data. Studies have emphasized the importance of integrating various data sources and selecting key features to improve prediction accuracy [8, 9]. Several studies have applied ensemble methods to predict crop yields under various Indian agricultural conditions, demonstrating their effectiveness in handling the complexity of agricultural data [7, 10-13]. Ensemble-driven techniques and explainable AI, like SHapley Additive exPlanations (SHAP), have shown promise in enhancing prediction accuracy and model interpretability. SHAP values are a game-changer in the realm of explainable AI, providing clear insights into the contribution of each feature in a prediction model. This is particularly crucial in agriculture, where understanding the impact of various factors can drive better decision-making [14-18]. From the above literature it is observed that ensemble-driven approaches present a robust and effective solution for agricultural crop yield prediction in Indian conditions. These methods improve predictive accuracy and offer valuable insights for optimizing agricultural practices, ultimately contributing to enhanced productivity and sustainability in Indian agriculture. Despite their effectiveness, many advanced ML models are often perceived as "black boxes" due to their lack of interpretability. Farmers and policymakers need transparent models to understand the predictions and make informed decisions. In this paper, an ensemble-driven approach is used for the prediction of yield in Indian Conditions. SHAP analysis is performed for model explanation and interpretability.

2. Feature Analysis

This section describes the detailed analysis of the parameters or factors contributing to yield prediction along with methodology adapted for yield prediction using machine learning.

2.1. Parameters

All the major factors contributing towards the yield of various crops, along with the topographical conditions, are thoroughly analysed in this section. By analysing these factors and parameters, researchers can prepare a proper report depicting the strategies that should be used for the utmost utilization of resources along with the prevention and detection of any unwanted events. The dataset contains data on the yield of 55 crops over 6 different seasons for 30 states over a tenure of 23 years (1997 - 2020). It also contains other topographical features such as rainfall, area, pesticides, fertilizers, etc. The dataset was collected from Kaggle for reference study [19]. A total of 10 parameters are used for the analysis of crop yield. Out of these 10, 7 parameters have numerical values, while the other 3 have categorical values. The parameters having categorical data are crop, season and state. The parameters having numerical values are crop year, area, production, annual rainfall, fertilizers, pesticide and yield. Their minimum, maximum and mean values are given in Table 1. Categorical features are shown in Table 2. They include crops (55), seasons (6) and states (30).

Table 1. Factors contributing yield prediction with min, max and mean values

SN	Parameters	Min. Value	Max. Value	Mean Value
1	Crop Year	1997	2020	2009
2	Area	0.50	5.080810e+07	1.7996e+05
3	Production	0.00	6.326000e+09	1.6439e+07
4	Annual Rainfall	301.30	6552	1437
5	Fertilizer	54.17	4.835407e+09	2.4103e+07
6	Pesticide	0.09	1.575051e+07	4.8848e+04
7	Yield	0.00	21105	7.9954e+01

Table 2. Factors contributing to yield prediction with categorical attributes

SN	Feature	Categories
1	**Crops**	"Arecanut, Arhar/Tur, Urad, Horse-gram, Gram, Moong, Masoor, Bajra, Jowar, Ragi, Maize, Banana, Barley, Black Cardamom, Cashew nut, Castor seed, Turmeric, Pepper, Coconut, Coriander, Cotton, Cowpea, Dry chillies, Garlic, Ginger, Groundnut, Guar seed, Jute, Khesari, Linseed, Mesta, Cereals, Moth, Niger seed, Oilseeds, Onion, Rabi pulses, Kharif pulses, Summer Pulses, Peas & beans, Potato, Rapeseed, Mustard, Rice, Safflower, Sunflower, Sannhamp, Sesamum, Small millets, Soyabean, Sugarcane, Sweet potato, Tapioca, Tobacco, Wheat, oilseeds"
2	**States**	"Jammu and Kashmir, Haryana, Uttarakhand, West Bengal, Uttar Pradesh, Andhra Pradesh, Arunachal Pradesh, Assam, Bihar Delhi, Gujarat, Himachal Pradesh, Jharkhand Karnataka, Kerala, Madhya Pradesh, Chhattisgarh, Maharashtra, Goa, Manipur, Odisha, Meghalaya, Mizoram, Punjab, Nagaland, Sikkim, Tripura, Tamil Nadu, Telangana, Puducherry"
3	**Seasons**	"Autumn, Kharif, Rabi, Summer, Whole Year, Winter"

The analysis for these factors is described below.

2.1.1. Year

The dataset contains the yield prediction data for the period of 1977 to 2020. The yield prediction over these time periods is shown in Figure 1. It can be observed from Figure 1 that the yield has increased over the year, but after 2014, it is showing a declining trend. Reasons can be climate change, decreased soil fertility, and continuous changes in rainfall patterns. There has been a shift towards more frequent and intense rainfall events, leading to both droughts and floods, which is one of the contributing reasons for the decline in the yield after 2014. Additionally, the increased frequency of extreme weather events, such as heat waves and pest outbreaks, has also contributed to the decline in crop yields after 2014. Furthermore, the shift towards more intensive and unsustainable agricultural practices, such as excessive use of fertilizers and pesticides, has led to soil degradation and reduced crop yields over time.

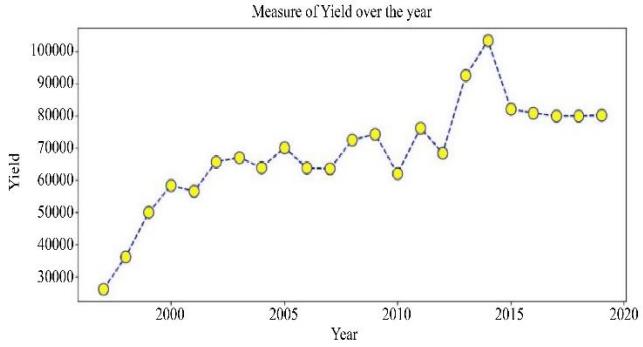

Fig. 1 Graph displaying year-wise yield

2.1.2. Area under Cultivation

The area under cultivation has a great impact on the yield of crops. The yield is also expected to increase as the area increases, but it is not always necessary. This is because factors such as soil quality, climate, and irrigation can affect the yield more than the area alone. This shows that while a

larger area can provide higher yields, other factors, such as soil quality and climate, also play an important role in determining the yield. Figure 2 shows the area under cultivation spanning from 1977 to 2020.

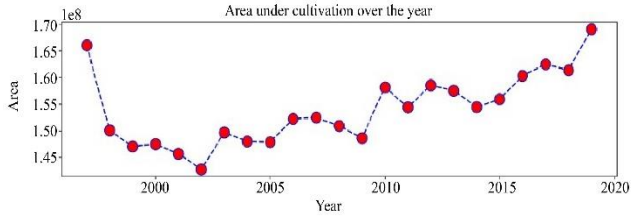

Fig. 2 Year-wise distribution of cultivable land

It can be observed in Figure 2 that the area under cultivation is increasing continuously, which in return tends to give us more yield. Many agriculturists have been able to turn unused and degradable land into cultivable land by using fertilizers, pesticides and other techniques.

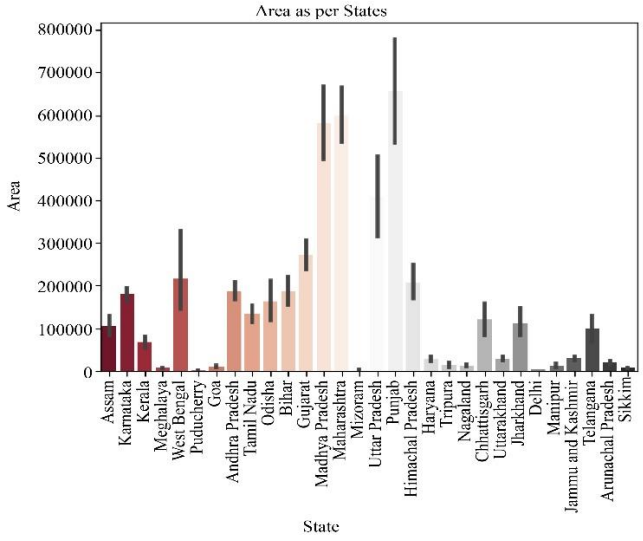

Fig. 3 Graph displaying cultivable land area present in individual states

The overall distribution of cultivable area per state is shown in Figure 3. More cultivable land generates more yield along with more employment opportunities for rural people. It also portrays economic development as agriculture has been the backbone of our country's economy.

2.1.3. Production

The production of crops is also one of the most important contributing factors that affect yield. There is a direct relation between production and yield; more production means more yield. More production means more resources being used and more favourable conditions for various crops, such as more beneficial use of fertilizers, pesticides, soil quality, rainfall and more. Figure 4 shows the total crop production per state, where it can be seen that the South Indian states have more production levels.

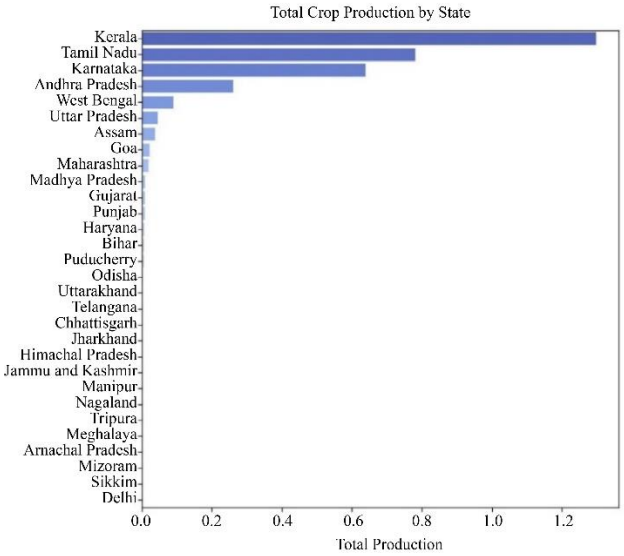

Fig. 4 Total production per state

2.1.4. Annual Rainfall

Annual rainfall has a significant impact on the yield of crops. Proper rainfall is essential for the good yield of crops.

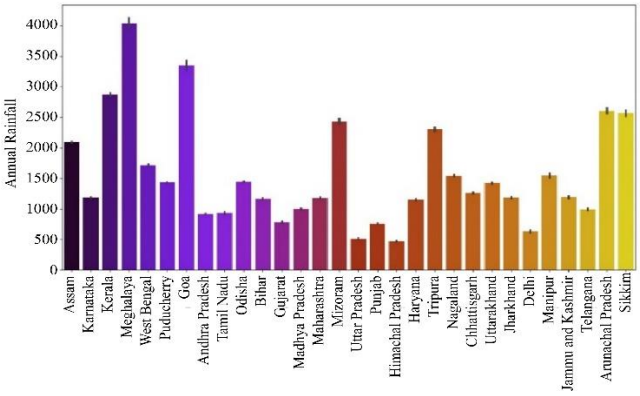

Fig. 5 Annual rainfall per state

Low rainfall levels result in scarcity of water; however, higher rainfall levels result in drought situations, and both these situations are very unfavourable for crop growth and yield. Higher rainfall may also lead to the spreading of crop diseases. Proper rainfall level indicates higher yield and good crop health. In these ways, we can observe how annual rainfall affects the yield of crops. Annual rainfall per state is depicted in Figure 5.

2.1.5. Fertilizers and Pesticides

Proper use of fertilizers can result in higher yield by providing crops with all the necessary nutrients required for good crop health and growth, but overuse of fertilizers may result in degradation of the quality of the soil by turning it into uncultivable land. Figure 6 indicates how farmers are involving the use of fertilizers in their conventional method of irrigation to increase their production and yield.

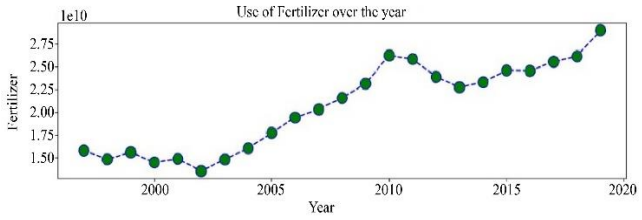

Fig. 6 Graph of fertilizer being used per year

The use of pesticides affects the growth of crops, and yields vary considerably. Proper use of pesticides indicates that the pests and diseases are controlled, which increases the crop yield by reducing the possible losses. Pesticides protect the crops from damage, and ensure proper unharmed growth of the crops. Farmers have adopted the use of pesticides in their traditional method of irrigation to boost crop yields, which can also be seen in Figure 7.

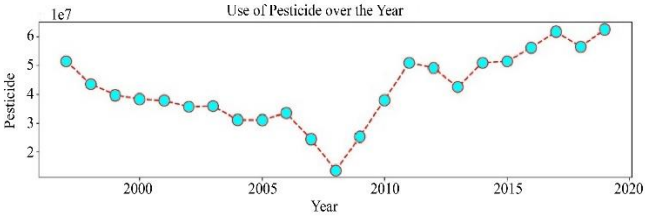

Fig. 7 Graph displaying pesticides used per year

However, it should be noted that the overuse of pesticides and fertilizers results in the degradation of the soil quality along with very poor yield quality, which has harmful health effects on food consumers. Sustainable irrigation practices never advise to overuse of pesticides and fertilizers but to use them in an optimized way.

2.1.6. Season

The seasons have a big impact on how well crops grow. Crops like rice, sugarcane, and small grains grow better when it rains a lot in the summer because they need lots of water and warm weather to grow well. On the other hand, crops like wheat, barley, and beans grow better in the cool and dry Rabi season.

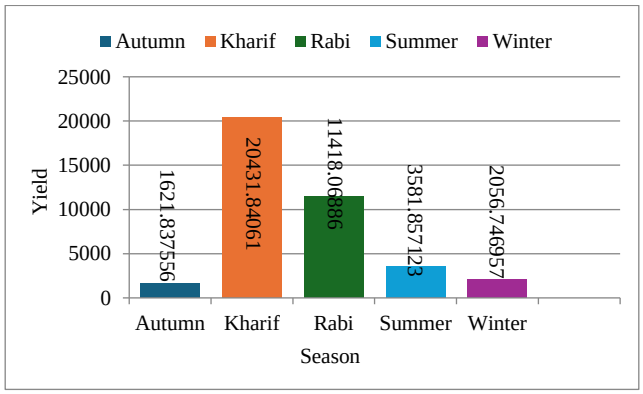

Fig. 8 Bar plot displaying season-wise yield

Farmers need to adjust their farming methods to match the specific needs of each crop during different seasons to get the best results. Figure 8 shows the season-wise yield in 30 different states in India.

3. Methodology

Machine learning methods can be divided into widely different approaches. It depends on the purpose and use, such as decision trees, random forests, and gradient growth. Extreme gradient optimization Adaptive improvements Histogram based optimization and the SHapley Annotation (SHAP) and other interpretive frameworks. Each method provides an effective and efficient approach to solving complex problems, such as agricultural yield forecasting.

3.1. Decision Tree (DT)

DT is a supervised learning model widely used for classification and regression tasks. This is shown as a flowchart-like tree structure, in which each node corresponds to an experiment or position in the feature, branches represent the results of these experiments, and leaf nodes represent the final prediction. Each method from the root to the leaf node can be interpreted as a specific rule.

Algorithm

- Start with the entire dataset. For each feature X_j, consider all possible splits S.
- Calculate the reduction in the sum of squared errors $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$ for each split.
- Choose the split that maximizes $\Delta SSE = SSE_{parent} - (SSE_{left} + SSE_{right})$ SSE_{parent} - sum of squared errors before the split SSE_{left} - sum of square error of left child node
  SSE_{right} - the sum of the square error of the right child node
- Apply the same process recursively to the left and right child nodes. Continue splitting until a stopping rule is met. Commonly used stopping rules are:
  - Minimum Error Reduction: Stop if the reduction in error from a split is less than a predefined threshold.
  - Maximum Tree Depth: Stop if the depth of the tree exceeds a predefined limit.
  - Impurity Measure Threshold: Stop if the impurity measure (e.g., variance or SSE) is below a certain threshold.
- For a defined input, the travel process in this traversal is traversed in the tree from the root to a leaf node. Now, it will return the mean target value of complete observations at the leaf node as a prediction. $\hat{y}_{leaf} = \frac{1}{N}\sum_{i=1}^{N} y_i$
  N – Number of observations in left node
  y_i - target values

3.2. Random Forest (RF)

RF is an ensemble-based kind of machine learning where the training phase considers multiple decision trees as individual classifiers to interpret a single data point. - In the case of regression, the final derived value is the average of predictions made by the individual trees. - The algorithm has incorporated the bagging and random selection of features to reduce over fitting and improve predictive power. Each of those trees was constructed on a bootstrap sample and determined for the best split at each node by randomly selecting a subset of features. - All trees then averaged their predictions to yield the final value, making random forests not overly sensitive to noise in the data and robust against over fitting.

For each bootstrap sample Db, initialize tree Tb, and at each node of Tb , randomly select m features. Determine the best split based on the selected m features. Split the node and repeat the process for child nodes until the stopping criteria are met. For a new observation x, predict $\hat{y}_b$ from each tree Tb and then aggregate predictions by averaging using $\hat{y} = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_b$

RF exploit simultaneously the advantages of many DTs through the technique of bagging and the random selection of features, whose end result is better prediction and robustness.

3.3. Gradient Boosting (GB)

GB redesigns the boosting problem using gradient descent on a numeric objective, which is the loss of the model can be optimized by adding more weak learners and minimizing that loss. The steps involved are,

- Initialize the model with a constant prediction, usually the mean of the target values
  $\hat{y}_i^{(0)} = \frac{1}{N}\sum_{i=1}^{N} y_i$.
- Compute the negative gradient for each observation using
  $r_i^{(m)} = -\left.\frac{\partial L\left(y_i, \hat{y}_i^{(m-1)}\right)}{\partial \hat{y}_i^{(m-1)}}\right|_{\hat{y}_i = \hat{y}_i^{(m-1)}} = y_i = \hat{y}_i^{(m-1)}$,
  L - loss function (Mean Squared Error)
  $L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$
  y_i - actual output, $\hat{y}_i$ - predicted output
- $h_m(x)$ to the residuals $r_i^{(m)}$ using
  $h_m = \arg\min_h \sum_{i=1}^{N}\left(r_i^{(m)} - h(x_i)\right)^2$
- $\hat{y}_i^{(m)} = \hat{y}^{(m-1)} + \eta h_m(x_i)$
- Predict the final model output by taking the sum of all the weak learners as $\hat{y}_i = \hat{y}^{(0)} + \sum_{m=1}^{M}\eta h_m(x_i)$

GB combines the power of multiple weak models to make a strong predictive model; by iteratively fitting weak learners to the negative gradient of the loss function and updating the model, Gradient Boosting effectively minimizes the prediction error.

3.4. Extreme Gradient Boosting (XGBoost)

XGBoost extends gradient boosting by incorporating regularization techniques and optimization enhancements [23]. The objective function to be minimized in XGBoost for regression is given by

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

$l(y_i, \hat{y}_i)$ - loss function, y_i - the actual output,

$\hat{y}_i$ - predicted output, $\Omega(f_k)$ - Regularization Term,

T- number of trees (iterations).

XGBoost uses regularization techniques like L1 and L2 regularization on the weights of the trees, and tree pruning to control model complexity.

3.5. Adaptive Boosting (AdaBoost)

The AdaBoost algorithm represents an ensemble learning technique with popularity in machine learning applications. It combines multiple weak learners to form a much stronger, more accurate model. It iterates through training weak classifiers by decision stumps (simply decision trees) to adjust their weights according to performance.

All training data points are initially given an equal weight. It then concentrates more on the misclassified data points after training the weak learner, increasing their weights and decreasing the weights of the correctly classified points, which is given by

$$L = \sum_{i=1}^{n} e^{-y_i F(x_i)}$$

y_i- actual output, $F(X_i)$ is the predicted value by the current ensemble of weak learners for the i^{th} instance. The objective is to find the ensemble F(X) that minimizes this exponential loss. Each training sample (x_i, y_i) is associated with a weight w_i updated after each iteration. The algorithm flow is described below

- Initially, all weights are set $w_i = \frac{1}{n}$ for $i = 1, 2, ...n$.

- For each iteration i, fit the weak learner $h_t(x)$ to the training set using weights $\{w_i\}$. Compute the weighted error rate e_t of h_t using $e_t = \sum_{i=1}^{n} w_i . I(y_i \neq h_t(x_i))$ Here, I is the indicator function.
- Calculate the learner's weight $\alpha_t = \frac{1}{2}\log\left(\frac{1-e_t}{e_t}\right)$.
- Update the weights using $w_i \leftarrow w_i . \exp(-\alpha_t . y_i . h_t(x_i))$ and normalize it.
- Final prediction is obtained by **c**ombining the weak learners into a strong learner using $F(x) = \sum_{t=1}^{T} \alpha_t . h_t(x)$

In the context of regression, AdaBoost adjusts the weights of training instances to sequentially fit multiple weak learners, each of which attempts to minimize the exponential loss. The final prediction F(X) is a weighted combination of these learners, where each learner contributes proportionally to its accuracy in predicting the residual errors from previous iterations [24].

3.6. Histogram Boosting (HB)

Histogram-based learning is a gradient-boosting algorithm that learns decision trees sequentially. With each iteration, it learns a tree to minimize some loss function and adds new trees to the ensemble. These histograms allow efficient computation of gradient statistics for every histogram bin at the time of tree construction. The algorithm for yield prediction using histogram based gradient boosting is given below.

- Initialize the prediction with a constant value, typically the mean of the target variable. For each iteration t, Compute the pseudo-residuals $r_i^{(t)} = -\frac{\partial L\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}}$.
- Create histograms for each feature by aggregating the pseudo-residuals and, if applicable, the second-order gradients into bins. Choosing the split that maximizes the reduction in the loss function.
- Fit a decision tree to the binned data using the chosen split points. Update the model using $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$

  Where η is the learning rate and f_t is the newly fitted tree.

3.7. Interpretable Model Using SHapley Additive exPlanations (SHAP)

SHAP is the methodology that is used for interpreting the output of machine learning models. The core concept behind the SHAP values is related to cooperative game theory as well as Shapley values. The other methodologies do not explain how each feature contributed to the predictions.

This might lead to fairness but enables ease of understanding for one and all. SHAP is useful because it tells us what each feature is important in making predictions. The model's prediction $f(x)$ is seen as the payout from a cooperative game where features collaborate to produce the output. The SHAP value for the feature i in the context of a specific input x is given by

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}\left[f_{S\cup\{i\}}(x) - f_S(x)\right]$$

Where $N = \{1, 2, ...n\}$ is the total dataset, $|S|$ is the cardinality of the S, $f_S(x)$ is denotes the model output using only the features in S ensuring fair and consistent feature importance values based on the cooperative game theory concept of Shapley v0alues. In this paper, model interpretation is carried out using SHAP analysis to provide an explanation of obtained results on yield prediction [26].

4. Results and Discussion

Six machine learning algorithms are applied to agricultural datasets for yield prediction. The quantitative assessment of the yield prediction is carried out using the following metrics [27, 28]. Consider N is the total number of samples in the dataset, y_i is the actual value of the yield, $\hat{y}_i$ is the predicted value of the yield after applying ML algorithms, $\bar{y}_i$ is the average value of yield prediction.

A high R^2 score indicates that the model accurately captures the variability in crop yields based on input features. It helps understand how well the model can generalize and how much of the yield variability is explained by the model. The quantitative assessment of six ML algorithms for yield prediction using MAE, MSE and RMSE is shown in Table 3.

The best-performing model is GB with a very low MAE of 0.0151 and the lowest. A higher R2 score of 0.9340 shows that it perfectly fits the data and has high predictive accuracy, while random forest with the lowest MAE of 0.0128 (less than the linear regression model) and efficient R2 score of 0.9041 discloses its exceptional predictive ability.

$$\text{Mean Absolute Error (MAE)} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|$$

$$\text{Mean Square Error (MSE)} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}$$

$$R^2 \text{ Score} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}$$

Table 3. Quantitative assessment of six machine learning algorithms for yield prediction using MAE, MSE and RMSE

Model	MAE	MSE	RMSE	R2_score
Histogram Gradient Boosting	0.0245	0.0985	0.3138	0.9052
Decision Tree	0.0156	0.1444	0.3800	0.8610
Extreme Gradient Boosting	0.0173	0.1361	0.3689	0.8690
Random Forest	0.0128	0.0996	0.3157	0.9041
Adaptive Boosting	0.0213	0.1165	0.3413	0.8879
Gradient Boosting	0.0151	0.0685	0.2618	0.9340

Other models such as DT, XGBoost, AdaBoost, and HB exhibit different levels of performance, each showing potential in specific metrics but generally followed by GB and RF a little in overall predictive power and accuracy. The scatter plot in Figure 9 depicts the performance of HB, DT, XGBoost, RF, AdaBoost, and GB. Each model's predictions are represented by different colored dots, as indicated in the legend. The green diagonal line represents perfect predictions (y = x). Points closer to this line indicate higher accuracy of the model's predictions. With the highest R^2 score of 0.9340, Gradient Boosting's predictions are most closely aligned with the real values, demonstrating fine overall performance. Histogram Gradient Boosting and Random Forest, both models display robust predictive performance with R^2 ratings of 0.9052 and 0.9041, respectively.

Their predictions are also intently clustered around the perfect line. XGBoost and Decision Tree have R^2 rankings of 0.8690 and 0.8610, indicating appropriate, however slightly much less correct predictions in comparison to the pinnacle performers. AdaBoost, although performing moderately nicely with an R^2 score of 0.8879, its predictions are slightly greater scattered compared to the other ensemble models. Figure 10 shows SHAP values for feature-contributing yield prediction assigns a cost to each feature inside the version, which indicates how lots that function contributed to the model's prediction.

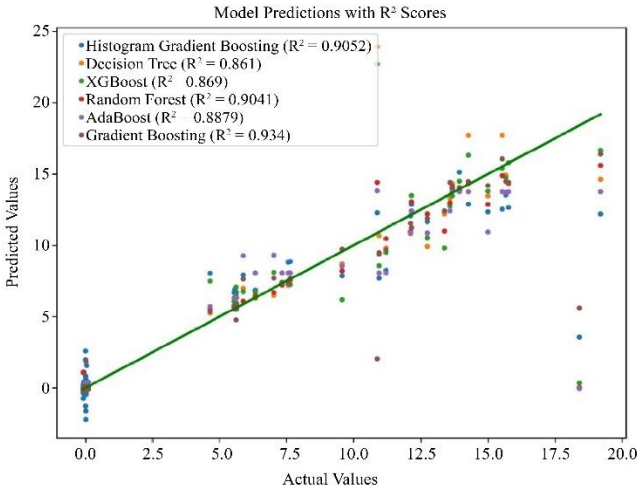

Fig. 9 Model prediction using R^2 score

In this case, the capabilities are elements that might affect crop yield, including production, area, pesticide use, state, crop kind, annual rainfall, crop year, and season. The x-axis of the graph suggests the common impact of the model output importance, which is essentially how much the characteristic changed the model's prediction. Positive values imply that the function elevated the model's crop yield prediction, even as bad values suggest that the characteristic decreased the prediction.

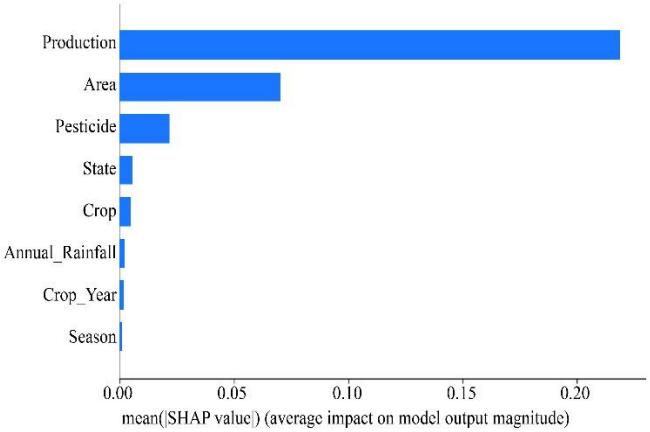

Fig. 10 SHAP values for feature contributing yield prediction

The y-axis of the graph shows the functions themselves. The height of the bar for every feature indicates the average magnitude of the feature's effect on the model's predictions. For instance, the bar for "Production" is the highest, meaning that production location had the biggest effect on the version's crop yield prediction.

The SHAP summary plot shown in Figure 11 helps us to understand which factors are most important for predicting crop yields and how these factors affect the model's predictions. The SHAP summary shows two things; the first is feature importance, which is represented by the y-axis of the factors used by the model to make its predictions. Items are listed in order of importance, with the most important item at the top. In this graph, crop production is most affected by production, followed by area, pesticide state and crop characteristics. The second is feature impact on the model, represented by the x-axis and the dot color. The x-axis shows the SHAP value, which indicates how much a factor contributed to a particular predictor. Positive values indicate that the trait increased the prediction of the crop yield model, while negative values indicate that the trait decreased the prediction.

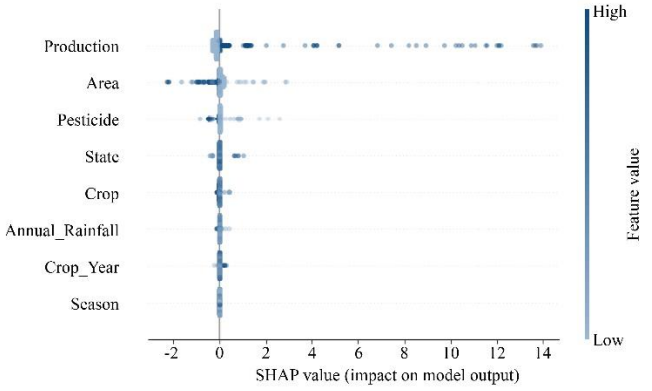

Fig. 11 Impact of SHAP values on predicting yield

5. Conclusion

In this comparative study of various machine learning models for crop prediction, including HB, DT, XGBoost, RF, AdaBoost, and GB, the performance metrics highlight the superiority of ensemble methods. GB emerges as the best-fit model with the highest R^2 score of 0.9340, the lowest MSE of 0.0685, and a RMSE of 0.2618. RF also performs exceptionally well, with a high R^2 score of 0.9041, a low MAE of 0.0128, and an RMSE of 0.3157. AdaBoost shows strong performance with an R^2 score of 0.8879 and an RMSE of 0.3413, making it reliable but slightly less effective than GB and RF. XGboost and DT show similar high performance, with XGBoost having an R^2 score of 0.8690 and an RMSE of 0.3689 and DT having an R^2 score of 0.8610 and an RMSE of 0.3800. HB performs well with an R^2 score of 0.9052 and an RMSE of 0.3138 but is slightly less efficient than the top models. These results underscore the effectiveness of ensemble methods, particularly GB and RF, in inaccurate crop prediction.

References

[1] Andrew Crane-Droesch, "Machine Learning Methods for Crop Yield Prediction and Climate Change Impact Assessment in Agriculture," *Environmental Research Letters*, vol. 13, no. 11, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan, "Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review," *Computers and Electronics in Agriculture*, vol. 151, pp. 61-69, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Aruvansh Nigam et al., "Crop Yield Prediction Using Machine Learning Algorithms," *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Devdatta A. Bondre, and Santosh Mahagaonkar, "Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms," *International Journal of Engineering Applied Sciences and Technology*, vol. 4, no. 5, pp. 371-376, 2019. [Google Scholar] [Publisher Link]

[5] P.S. Maya Gopal, and R. Bhargavi. "A Novel Approach for Efficient Crop Yield Prediction," *Computers and Electronics in Agriculture*, vol. 165, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Dhivya Elavarasan, and P.M. Durairaj Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," *IEEE Access*, vol. 8, pp. 86886-86901, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Farhat Abbas et al., "Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms," *Agronomy*, vol. 10, no. 7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Potnuru Sai Nishant et al., "Crop Yield Prediction Based on Indian Agriculture Using Machine Learning," *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Mamunur Rashid et al., "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction," *IEEE Access*, vol. 9, pp. 63406-63439, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Dilli Paudel et al., "Machine Learning for Large-Scale Crop Yield Forecasting," *Agricultural Systems*, vol. 187, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Mummaleti Keerthana, Priyanka Kunapuli, and Durga Bhavani Banavath, "An Ensemble Algorithm for Crop Yield Prediction," *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Tirunelveli, India, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mohsen Shahhosseini, Guiping Hu, and Sotirios V. Archontoulis, "Forecasting Corn Yield with Machine Learning Ensembles," *Frontiers in Plant Science*, vol. 11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Masahiro Ryo, "Explainable Artificial Intelligence and Interpretable Machine Learning for Agricultural Data Analysis," *Artificial Intelligence in Agriculture*, vol. 6, pp. 257-265, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Tongxi Hu et al., "Crop Yield Prediction via Explainable AI and Interpretable Machine Learning: Dangers of Black Box Models for Evaluating Climate Change Impacts on Crop Yield," *Agricultural and Forest Meteorology*, vol. 336, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Alejandro Morales, and Francisco J. Villalobos, "Using Machine Learning for Crop Yield Prediction in the Past or the Future," *Frontiers in Plant Science*, vol. 14, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] S. Iniyan, V. Akhil Varma, and Ch Teja Naidu, "Crop Yield Prediction Using Machine Learning Techniques," *Advances in Engineering Software*, vol. 175, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ersin Elbasi et al., "Crop Prediction Model Using Machine Learning Algorithms," *Applied Sciences*, vol. 13, no. 16, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Akshat Gupta, Akshat Saini, and Abhay Dewedi, Agricultural Crop Yield in Indian States Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset

[20] Leo Breiman et al., *Classification and Regression Trees*, 1st ed., Chapman and Hall/CRC, pp. 1-368, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Andy Liaw, and Matthew Wiener, "Classification and Regression by Random Forest," *R News*, vol. 2, no. 3, pp. 18-22, 2002. [Google Scholar]

[22] Jerome H. Friedman, "Stochastic Gradient Boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367-378, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[23] Tianqi Chen, and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 1st ed., Springer New York, NY, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[25] Fabian Pedregosa et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. [Google Scholar]

[26] Scott M. Lundberg, and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach California USA, pp. 4768-4777, 2017. [Google Scholar] [Publisher Link]

[27] Aurelien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed., O'Reilly Media, 2019. [Publisher Link]

[28] Sebastian Raschka, and Vahid Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2*, Packt Publishing, pp. 1-772, 2019. [Google Scholar] [Publisher Link]