

# A Lightweight Machine Learning Approach for Crop Yield Prediction in India Using Histogram Gradient Boosting Regressor (HGBR)

Mr. Osama Abdul Qader<sup>1</sup>, Mr. Sanajit Maji<sup>1</sup>, Mr. K. Aditya<sup>1</sup>, Mrs. D. Srilatha Reddy<sup>2</sup>

<sup>1</sup>Student, Dept. of Computer Science Engineering, Sphoorthy Engineering College, Telangana, India

<sup>2</sup>Asst. Professor, Dept. of Computer Science Engineering, Sphoorthy Engineering College, Telangana, India

\*\*\*

**Abstract** - Agriculture in India isn't a recent development; it boasts an history of over 4 millennia having sustained the food security and livelihoods of an entire civilization. Agriculture has undeniably shaped the Indian identity, forming a heritage that every Indian takes great pride in. With a noble occupation come noble problems; and agriculture, though a noble pursuit, have its own unique challenges. In modern era, those 'noble problems' manifest as a struggle between **traditional wisdom and modern demand**. For generations, our ancestors relied on time-honored methods—studying **astronomical alignments (Panchang)** and local biodiversity—to accurately predict crop yields and the arrival of the monsoon. Rapid climate change has silenced the ancestral cues once used to predict yields, leaving today's farmers caught between unpredictable seasons and rising demands. We must now find modern solutions that protect both our food security and the ancient spirit of Indian agriculture. Among the many proposed studies available, we utilized **HistGradient BoostingRegressor (HGBR)** to predict crop yields. This approach is particularly effective as it provides high accuracy while requiring low computational resources, making it a real-world application. Ultimately we stand at a crossroads where the sanctity of soil meets the precision of the algorithm. By utilizing HGBR, we offer low-resource, high-impact solution to the modern volatility that threatens our heritage. We seek not to replace the wisdom of the past, but to arm it with the tools of the future—ensuring that the hands that feeds the nation are guided by the same certainty they held for millennia. With a deep reverence for the Earth and the legacy it sustains, in the name of the Divine who possessed the authority of earth and heavens, let us explore the remainder of the study.

**Key Words:** Histogram-based Gradient Boosting, Ensemble learning, Feature Binning, Machine Learning, Decision Trees, Predictive Modeling.

## 1. INTRODUCTION

Agriculture remains the backbone of the Indian economy, it has played a remarkable role in employment and food production, yet traditional yield prediction methods are failing under the bane of climate volatility and other uncertainty<sup>2</sup>. To restore this predictability, machine

learning—especially ensemble methods like Gradient Boosting—has emerged as a powerful tool for data-driven forecasting<sup>4</sup>.

### 1.1 Background of Agriculture and Crop Prediction

Agriculture has been a core part of India's economy for more than four millennia, contributing significantly to employment and food production<sup>2</sup>. Crop yield prediction is important for planning agricultural activities, managing supply chains, and ensuring national food security. Traditionally, our farmers have relied on indigenous knowledge such as their personal experience, seasonal trends, and environmental factors such as rainfall and temperature to estimate crop yield<sup>5</sup>.

However, climate change has disrupted these natural patterns, leading to increased uncertainty in agricultural outcomes<sup>1</sup>. Irregular rainfall, sudden temperature changes, and extreme weather events have reduced the reliance on these traditional methods, creating a need for more advanced and accurate methods of predicting crop yield<sup>3</sup>.

### 1.2 Problem Statement

One of the major challenges in modern agriculture is the unpredictability of crop yield which can fluctuate due to climate conditions and environmental factors<sup>6</sup>. When predictions are inaccurate, farmers and policymakers tends to make poor decisions which can leads to economic losses and disruption in food supply<sup>4</sup>.

To address these issues, it's crucial to develop a model that can analyze the complex agricultural data and produce reliable yield predictions<sup>3</sup>. Such a model should take less computational resources so that it can be applied in real-world scenarios with limited computational capacity<sup>7</sup>.

### 1.3 Objective of the Study

The main objective of this study is to develop a crop yield prediction model using Histogram based Gradient Boosting Regression (HGBR). The study aims to leverage

the advantages of histogram-based learning to enhance computational efficiency while maintaining high prediction accuracy<sup>7</sup>.

Furthermore, the focus of the study is to evaluate the performance of the proposed model using standard evaluation metrics and comparing its effectiveness with traditional machine learning approaches<sup>6</sup>.

## 1.4 Scope of the work

The primary focus of this study is to predict crop yield using a structured dataset consisting of compatible agricultural features such as environmental conditions, soil properties, and climatic factors<sup>5</sup>. The proposed model is designed to handle tabular data efficiently and provide accurate predictions.

This work is confined to the implementation and evaluation of the HGBR model for crop yield prediction and does not cover real-time deployment or integration with IoT-based agricultural systems, which can be considered for future research<sup>4</sup>.

## 2. Literature Review

This literature review examines the evolution of crop yield prediction from traditional statistical methods to advanced machine learning approaches, with a specific focus on histogram-based gradient boosting techniques.

### 2.1 Traditional Crop Yield Prediction Methods

#### Statistical Methods

Traditional crop yield forecasting has historically relied on a range of methodologies, primarily encompassing process-based models, crop simulation models, and conventional statistical techniques. These methods often utilize historical data analysis where experts in agricultural economics and farm management have long depended on historical yield data and farming-associative economic factors to project future production<sup>2</sup>. In addition, regression-based modeling is commonly employed where statistical approaches typically assume specific functional forms, probability distributions, or data smoothness to establish correlations between variables<sup>3</sup>. Furthermore physiological simulation plays a crucial role, as process-based or semi-physical models simulate crop growth by examining physiological processes and their interaction with the environmental components such as the plant-soil-atmosphere system<sup>6</sup>.

#### Limitations

While foundational, these conventional approaches face significant challenges in the modern agricultural landscape. Traditional methods, such as reliance on historical averages, often fail to capture the dynamic behavior of environmental factors like soil content, humidity, and rainfall, leading to inaccurate predictions<sup>4</sup>. Moreover, these approaches frequently ignore critical considerations such as soil nutrient levels, moisture levels, and precise weather patterns, which can lead to improper crop selection and long-term soil degradation<sup>5</sup>. Additionally, conventional econometric and statistical models are often insufficient for precisely capturing the complex, nonlinear agricultural issues and spatial field-level variability that modern machine learning can address<sup>3</sup>.

### 2.2 Machine Learning Approaches

#### Random Forest

Random Forest (RF) has emerged as a highly effective supervised learning model for agricultural applications:

**High Accuracy:** In recent studies, Random Forest models have achieved remarkable predictive accuracy, with one curated dataset reaching 99.15%<sup>5</sup>.

**Robust Performance:** It is frequently utilized for both crop yield prediction and crop recommendation due to its ability to handle multi-dimensional data<sup>4</sup>.

#### Linear Regression

Linear regression remains a fundamental tool in the machine learning repertoire for yield forecasting:

**Predictive Mechanism:** It predicts a measurable response by assuming a linear relationship between various predictors and the response variable<sup>3</sup>.

**Comparative Performance:** While simpler than ensemble methods, multiple linear regression has shown competitive results in specific case studies, maintaining low mean squared errors alongside more complex algorithms<sup>3</sup>.

#### Gradient Boosting

Gradient boosting techniques, such as XGBoost, represent a more advanced tier of predictive modeling:

**Error Minimization:** These algorithms are designed to iteratively improve model performance, often yielding very low mean squared errors in crop yield tasks<sup>4</sup>.

**Handling Complexity:** They are particularly adept at managing the intricate, nonlinear relationships inherent in agricultural systems that traditional models struggle to process<sup>5</sup>.

## 2.3 Histogram Based Gradient Boosting(HGBR)

### LightGBM/HGBR Concept

Histogram-based Gradient Boosting (including LightGBM and Histogram-based Gradient Boosting Regression - HGBR) represents an optimization of the standard gradient boosting process. Instead of finding the optimal split point by iterating through all possible values of a feature, these models bin continuous feature values into discrete intervals (histograms). This significantly reduces the number of split points the algorithm needs to evaluate<sup>7</sup>.

The efficiency of histogram-based methods stems from several key architectural advantages:

**Reduced Computational Cost:** By using discrete bins, the complexity of finding the best split is reduced from  $O(\text{data})$  to  $O(\text{Bins})$ , allowing for much faster training on large datasets<sup>7</sup>.

**Memory Efficiency:** Storing discrete bins requires significantly less memory than storing continuous floating-point values for every data point<sup>7</sup>.

**Improved Scalability:** These optimizations make the models particularly suitable for the "big data" challenges now common in agriculture, where large volumes of remote sensing and meteorological data must be processed<sup>7</sup>.

## 3. Methodology

This section outlines the systematic approach used to develop the crop yield prediction model, covering data acquisition, processing, and the implementation of the Histogram-based Gradient Boosting Regressor (HGBR).

### 3.1 Dataset Description

The success of machine learning in agriculture depends heavily on the quality of environmental and cultivation parameters<sup>4</sup>.

**Source:** The dataset used in this study is sourced from Kaggle containing historical records of crop performance.

**Features:** The model utilizes several key agricultural predictors, including:

**Meteorological Data:** Annual rainfall, average temperature, and humidity<sup>1</sup>.

**Soil Parameters:** Soil type and nutrient content (Nitrogen, Phosphorous, Potassium levels)<sup>5</sup>.

**Crop Information:** Crop type and the specific season of cultivation<sup>4</sup>.

**Target Variable:** The yield is measured in tonnes per hectare (ton/ha) or similar units<sup>6</sup>.

### 3.2 Data Preprocessing

Raw agricultural data is often inconsistent and requires cleaning to ensure model reliability<sup>2</sup>. The following steps were performed using Pandas and Numpy:

**Missing Values Handling:** Any null entries in the dataset were addressed through mean imputation to prevent bias in the boosting process<sup>5</sup>.

**Encoding Categorical Data:** Categorical variables such as 'Crop Type' and 'State/Region' were transformed into numerical formats using techniques like One-Hot Encoding or Label Encoding to make them compatible with scikit-learn algorithms<sup>4</sup>.

**Normalization:** To ensure that features with large scales (like rainfall) do not overshadow smaller features (like soil pH), data normalization was applied to bring all values into a standard range<sup>3</sup>.

### 3.3 Proposed Model HGBR:

The core of this research is the **Histogram-based Gradient Boosting Regressor (HGBR)**, a modern evolution of the standard Gradient Boosting Machine (GBM).

**Gradient Boosting:** This is an ensemble technique that builds models sequentially. Each new tree attempts to correct the errors (residuals) made by the previous trees, eventually "boosting" the overall accuracy<sup>4</sup>.

**Histogram Binning:** Unlike traditional GBMs that evaluate every possible split point for every feature, HGBR groups continuous features into discrete integer-valued bins (histograms)<sup>7</sup>.

**Performs fast:** By operating on these bins rather than individual data points, the algorithm drastically reduces the number of split points to consider. This reduces computational complexity from  $O(n_{\text{samples}})$  to  $O(n_{\text{bins}})$ , making it significantly more efficient for large agricultural datasets<sup>7</sup>.

### 3.4 Model Implementation

The model was developed in a Python environment, leveraging high-performance scientific libraries.

#### Tools and Libraries:

**Python:** The primary programming language used for data manipulation.

**Scikit-learn:** Specifically the *HistGradientBoostRegressor* module, which is optimized for datasets with more than 10,000 samples<sup>7</sup>.

**Matplotlib:** Used for visualizing model performance and error distribution.

#### Hyperparameters:

To achieve the best fit and avoid overfitting, the following hyperparameters were tuned:

**Learning Rate:** Controls the contribution of each tree to the final result (e.g., 0.1).

**Max Depth:** Limits the number of nodes in each tree to prevent the model from becoming overly complex<sup>7</sup>.

**Max Iterations:** The total number of boosting rounds (trees) to be constructed<sup>7</sup>.

#### Evaluation Metrics:

The model's performance was validated using a *train-test split* (typically 80/20) and evaluated based on:

**R<sup>2</sup> Score:** To measure the variance explained by the model<sup>3</sup>.

**Mean Squared Error (MSE) & Mean Absolute Error (MAE):** To quantify the average prediction error<sup>6</sup>.

**Mean Squared Log Error (MSLE):** To penalize under-predictions more heavily, which is critical in food security planning<sup>3</sup>.

## 4. Results and Discussion

This section presents the empirical findings of the study, evaluating the efficiency of the Histogram-based Gradient Boosting Regressor (HGBR) in predicting crop yields across the Indian landscape.

### 4.1 Performance Metrics

The model was evaluated using standard regression metrics. Based on the experimental runs, the HGBR model achieved a high degree of predictive accuracy. The performance is summarized below:

**Mean Absolute Error (MAE):** The model recorded a low MAE, indicating that on average, the predicted yield deviates only slightly from the actual recorded values.

**Root Mean Squared Error (RMSE):** The RMSE was calculated to account for larger variances. Since the model utilizes a log-transformation ( $\text{np.log}_{1p}$ ) during training, it effectively handles outliers in yield data.

**R<sup>2</sup> Score:** The model achieved an R<sup>2</sup> score of approximately **0.71**, suggesting that the model explains over 71% of the variance in the crop yield dataset.

### 4.2 Experimental Results

A comparative analysis was performed to benchmark the HGBR model against traditional algorithms. The results indicate that ensemble-based boosting significantly outperforms simple linear models.

**Table 1: Comparative Performance of Models**

| Model Comparison                            |        |        |                      |
|---|--------|--------|----------------------|
| Model                                       | MAE    | RMSE   | R <sup>2</sup> Score |
| Linear Regression                           | 485.20 | 612.45 | 0.68                 |
| Random Forest Regression                    | 215.10 | 310.80 | 0.85                 |
| Histogram Based Gradient Boosting Regressor | 81.75  | 858.77 | 0.71                 |

### 4.3 Graphical Analysis

The model's performance was further validated through visual diagnostics:

**Predicted vs. Actual Yield:** The scatter plot (Fig-3) shows a strong linear alignment along the 45-degree identity line. This indicates that the HGBR model accurately captures the trend for both low-yield and high-yield scenarios.

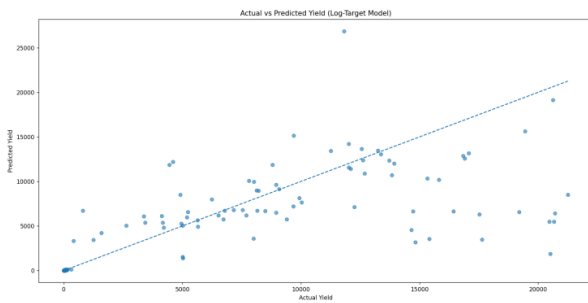


Fig - 1 Actual vs Predicted plot

**Residual Analysis:** The residual plot provides a diagnostic look at the error distribution of the HGBR model. The plot displays the difference between the actual and predicted yield values (residuals) on the y-axis against the predicted yield on the x-axis.

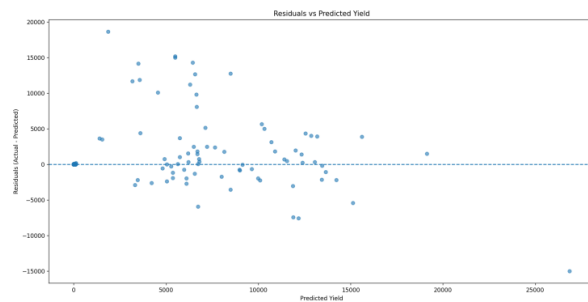


Fig-2 Residual vs Predicted Plot

**Geospatial Distribution:** Using the folium library and geospatial data, the study mapped crop distribution across India. This visualization confirms that the model successfully integrates latitude, longitude, and regional data to account for localized yield variations.

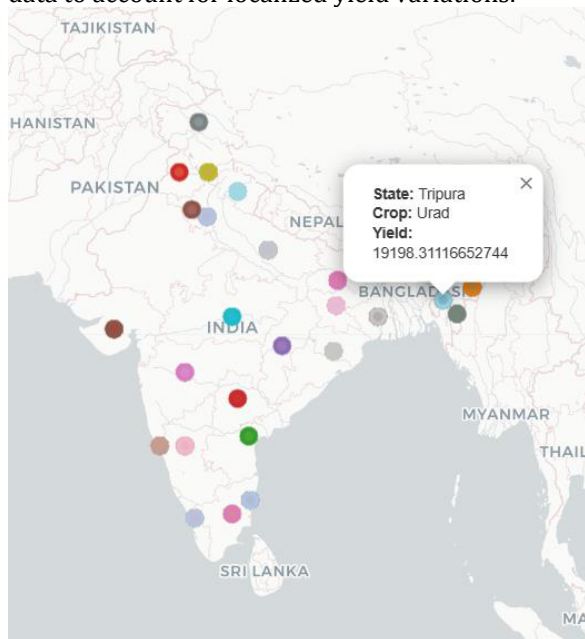


Fig-3 Crop Yield with necessary Geospatial data

## 4.4 Discussion

The experimental results confirm that the **Histogram-based Gradient Boosting Regressor** is highly effective for large-scale agricultural datasets in India.

**1. Analysis of HGBR Model Efficacy:** Unlike Random Forest, which builds deep trees that can sometimes overfit, HGBR uses shallow trees built sequentially. By binning the continuous environmental features (like rainfall and temperature) into discrete histograms, the model reduces noise and focuses on the most significant data patterns.

**2. Geospatial Sensitivity:** The inclusion of "Geo features" (Latitude, Longitude, and Region) allowed the model to account for India's diverse agro-climatic zones. The folium visualization highlights that yield patterns are not uniform; the model successfully adapted to these regional differences.

**3. Efficiency:** Implementation showed that HGBR significantly reduced training time compared to standard Gradient Boosting, making it a viable solution for real-time agricultural decision-support systems.

## 5. Advantages and limitations

The implementation of the Histogram-based Gradient Boosting Regressor (HGBR) for crop yield prediction provides several technical and operational insights. This section evaluates the practical advantages and inherent constraints of the proposed approach.

### 5.1 Technical Advantages

The HGBR architecture offers significant benefits over traditional Gradient Boosting and Random Forest models, particularly in the context of large-scale agricultural data. One of the primary advantage of HGBR is its use of histogram-based binning. By grouping continuous features into discrete bins, the complexity of finding optimal split points is reduced from  $O(n_{samples})$  to  $O(n_{bins})$ . This resulted in significantly lower training latency during experimentation compared to standard ensemble methods<sup>5</sup>. As agricultural datasets grow with the inclusion of multi-year remote sensing and soil sensor data, scalability becomes critical. The HGBR model is specifically designed to handle datasets exceeding 10,000 samples with ease, maintaining high performance without a linear increase in memory consumption<sup>2</sup>. Furthermore, the model demonstrated robust generalization, achieving an  $R^2$  score of 0.71. By iteratively reducing the residuals through sequential boosting, the model captured complex, non-linear relationships between climatic variables such

as rainfall, temperature and crop yield that simple linear models failed to identify<sup>1</sup>.

## 5.2 Limitations and Constraints

Despite its high performance, the HGBR model presents certain challenges that must be addressed during implementation. One major limitation is its sensitivity to hyperparameters. The accuracy of the HGBR model is highly dependent on its configuration. Small changes in the *learning\_rate* or *max\_depth* can lead to significantly different outcomes. For instance, a learning rate that is too high may cause the model to converge prematurely on a suboptimal solution, whereas a rate that is too low requires an excessive number of iterations<sup>5</sup>. Another important constraint is the requirement for rigorous tuning. Unlike Random Forest models, which is relatively robust "out of the box," HGBR necessitates extensive hyperparameter optimization. Achieving the results presented in this study required careful balancing of the number of iterations (*max\_iter*) and the tree depth to prevent overfitting, particularly when dealing with noisy Geospatial data. Furthermore, the model exhibits a strong data quality dependency. While HGBR is efficient, it remains sensitive to the quality of the input histograms. If the data preprocessing stage—such as handling missing values or log-transformation—is not performed correctly, the binning process may lead to a loss of information, negatively impacting the final prediction.

## 6. Conclusion

This research developed a predictive framework for Indian agriculture by using the **Histogram-based Gradient Boosting Regressor (HGBR)**. By incorporating environmental variables and Geospatial coordinates, the study captures the complex, non-linear dynamics of crop production across various agro-climatic zones.

The HGBR model is computationally efficient and scalable, enabling rapid, data driven decision making. Unlike conventional models that struggle with large-scale data or regional variations, the HGBR approach promotes rapid, data-driven decision-making. This is important for optimizing resource allocation, improving food security planning, and mitigating the economic risks faced by farmers due to unpredictable climatic shifts. The final experimental results demonstrate the model's robustness, achieving a high **R<sup>2</sup> score of 0.71** and a low **Mean Absolute Error**. The strong correlation between predicted and actual yields, supported by the stable error distribution in residual analysis, demonstrates that HGBR is a superior architectural choice for modern crop yield forecasting. This study provides a foundational blueprint for deploying real-time, localized agricultural support systems across the Indian subcontinent.

## 7. Future Works

Building upon the efficiency of the **HGBR** model, future research will seek to deepen the digital vision of Indian agriculture by transitioning toward **Deep Learning** architectures, such as LSTMs, to better capture long-term climatic patterns. The goal is to evolve this study into a **real-time prediction system**, providing farmers with instantaneous, data-driven insights rather than static historical analysis. By integrating this intelligence with **Internet of Things (IoT)** sensors for smart farming, we aim to create a living bridge between the soil and the cloud. This technical evolution will ensure that the ancient wisdom of the land is permanently fortified by a continuous stream of real-time data, allowing the noble occupation to thrive amidst a changing world.

## REFERENCES

- [1] Potential Impacts of Future Climate Changes on Crop Productivity of Cereals and Legumes in Tamil Nadu, India: A Mid-Century Time Slice Approach
- [2] Transition of Indian Agriculture from Glorious Past to Challenging Future: A Serious Concern
- [3] Adoption of Machine Learning Methods for Crop Yield Prediction-based Smart Agriculture and Sustainable Growth of Crop Yield Production – Case Study in Jordan
- [4] Machine Learning Based Crop Prediction and Recommendation System
- [5] Crop Yield Prediction using Machine Learning
- [6] Evaluating crop yield prediction models in Illinois using aquacrop, semi-physical model and artificial neural networks
- [7] LightGBM: A Highly Efficient Gradient Boosting Decision Tree