WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Yield Gap Analysis Using XG Boost

***Hariharan B[1], Dr. E. K. Girisan[2]***

[1]III-B. Sc-CT, Department of Computer Technology, Sri Krishna Adithya College of Arts and Science, Kovaipudur, Coimbatore.
[2]Assistant Professor, Department of Computer Technology, Sri Krishna Adithya College of Arts and Science, Kovaipudur, Coimbatore.

**ABSTRACT**

The primary aim of this study is to predict crop yields based on factors such as cultivation area, rainfall, and temperature (both maximum and minimum). This approach is designed to assist Indian farmers in estimating crop production under varying environmental conditions. In recent times, machine learning-based methods for crop yield prediction have gained prominence due to their higher accuracy compared to traditional approaches. This research compares the performance of Linear Regression, Support Vector Regression, Decision Tree, and Random Forest models against the XGBoost algorithm. These models are evaluated using metrics such as $R^2$, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The dataset for this analysis was sourced from data.gov.in, covering the years 2000 to 2014. The study focuses on data from four southern Indian states—Andhra Pradesh, Karnataka, Tamil Nadu, and Kerala—chosen for their similar climatic conditions. Results indicate that the proposed XGBoost-based model outperforms the other algorithms, achieving the highest $R^2$ value of 0.9391, demonstrating superior predictive accuracy.

**KEYWORDS:** Crop yield, precision agriculture, Random Forest, Support Vector Regression, XGBoost algorithm, data-driven models, machine learning techniques**.**

## I.INTRODUCTION

Agriculture has been the foundation of modern civilization, as societies began to flourish with its development. Without agriculture, civilization as we know it would not exist today. India ranks as the second-largest producer of agricultural products, following China. Over 50% of India's population relies on agriculture and related industries for their livelihood, highlighting the sector's critical role as the backbone of the nation. Despite advancements in other industries, agriculture continues to employ nearly half of the workforce in India. In 2016-17, the sector contributed 17.32% to the GDP, with a Gross Value Added (GVA) of approximately ₹23.82 lakh crore. Statistically, India accounts for 7.39% of global agricultural output, producing an estimated 275 million tons of grains during 2017-18.

However, despite significant contributions, India is not the leading producer globally. One major reason is the limited adoption of advanced agricultural technologies in many parts of the country. To improve agricultural practices and boost productivity, it is essential to embrace modern tools and techniques.

This article focuses on predicting crop yields and proposes a framework to guide farmers in selecting suitable crops for specific regions, thereby increasing production. Historical climate and harvest data, along with confidential images of crop fields, are analyzed to forecast crop yields and recommend appropriate crops for a given farmland. The proposed model leverages modern techniques such as machine learning, image processing, and data analysis to enhance cultivation methods and achieve better outcomes.

## II.MACHINE LEARNING TECHNIQUES

Machine learning, a subset of advanced computational techniques, allows systems to learn and make predictions or decisions without explicit programming for each specific task. It enables the automation of data analysis and processing of large datasets, thereby reducing manual effort while enhancing accuracy.

The primary goal of machine learning is to create algorithms that improve decision-making by leveraging historical data. Predicting crop yield based on soil characteristics and other environmental factors is often challenging for farmers. Additionally, addressing agricultural pollution and improving land productivity are essential to ensure food security and sustainability.

Optimal utilization of agricultural land is crucial for meeting the country's food demands. This study proposes a crop yield prediction system using the XGBoost algorithm. Machine learning is broadly categorized into three types:

- **Supervised Learning**

- **Unsupervised Learning**
- **Reinforcement Learning**

Below, we describe some of the key algorithms that are instrumental in this research:

**A. Linear Regression**

Linear regression is a method used to model the relationship between input variables (independent variables) and an output variable (dependent variable). This relationship is represented by the following equation:

y= a + bx

Here, a and b are parameters that the model calculates to fit the data. The primary goal is to determine the values of a and b that best represent the underlying relationship in the data.

**B. Decision Tree Regression**

A decision tree is a predictive model that splits the dataset into smaller subsets based on specific features and conditions, forming a tree-like structure. Random forests, a collection of decision trees, build upon this concept by combining multiple trees to improve accuracy.

Decision trees are easy to interpret as they mimic human decision-making by asking a series of questions based on the dataset's features. For example, similar to how a person might ask successive questions to narrow down options, a decision tree processes data until it arrives at the best possible prediction. Increasing the number of trees in a random forest generally improves the model's accuracy by minimizing errors.

Here's the rewritten version of the content:

**C. Support Vector Regression (SVR)**

Support Vector Regression (SVR) is closely related to the Support Vector Machine (SVM) algorithm but is specifically designed for regression tasks, working with continuous data. Unlike traditional regression methods that focus on minimizing the error rate, SVR aims to fit the error within a predefined margin or threshold.

While SVM utilizes two boundary lines to separate classes, SVR operates with a single hyperplane that seeks to minimize deviations within the specified threshold. This approach makes SVR particularly effective for tasks requiring high accuracy in continuous value prediction.

**D. XGBoost Algorithm**

XGBoost, or Extreme Gradient Boosting, is a highly efficient algorithm known for its speed and accuracy. Several factors contribute to its popularity:

- **Execution Speed**: XGBoost is faster than most gradient boosting algorithms. Studies, such as those conducted by Szilard Pafka, indicate that XGBoost outperforms other frameworks like R, Python, Spark, and H2O in terms of speed, accuracy, and memory efficiency.
- **Model Performance**: XGBoost excels in structured or tabular datasets, making it a reliable choice for classification and regression tasks. Its ability to optimize predictive modeling problems is a significant advantage over other algorithms.

**E. Random Forest Algorithm**

The Random Forest algorithm consists of a collection of decision trees built using random subsets of a dataset. It is versatile, supporting both classification and regression tasks. Random Forest is particularly useful because it can handle missing or noisy data effectively.

The process of building a Random Forest involves two main stages:

1. **Creating the Random Forest**:
    - Randomly select a subset of "x" features from the total "z" features, where x<<zx << z.
    - Identify the best split point among the selected features to define a root node.
    - Split the root node into branches (daughter nodes) based on the best division.
    - Repeat this process until the desired number of nodes is created.
    - Construct multiple trees by repeating the steps above "y" times, resulting in a forest of "y" trees.
2. **Making Predictions**:
    - Use the created Random Forest to predict the output for test data by aggregating the results of individual tre

In the initial stage, a set of features is randomly selected, and the best split strategy is applied to identify the root node. This process is repeated to generate branches and leaf nodes. Once several trees are constructed using the same method, they collectively form the Random Forest, which provides robust predictions by averaging or voting the outcomes of the individual trees.

## III.RANDOM FOREST PREDICTION PSEUDOCODE

To perform forecasting using a trained Random Forest algorithm, the following pseudocode is applied:

1. Select the test features and apply the rules of each randomly generated decision tree to predict the outcome and store the resulting predictions.
2. Count the votes for each predicted result.
3. Identify the result with the highest number of votes as the final output from the Random Forest algorithm.

When implementing forecasting with the Random Forest algorithm, the test data is evaluated against the rules of all the randomly constructed decision trees in the forest. For example, if the Random Forest consists of 100 decision trees, each tree will independently predict a result for the given test feature. These predictions may vary between the trees.

Suppose the 100 decision trees produce three different predictions: a, b, and c. The votes for a will correspond to the number of trees predicting a as the result, and similarly for b and c. If a receives the highest number of votes—say, 60 out of 100—then the Random Forest will predict a as the final output.

The voting mechanism ensures that the majority prediction is selected, increasing the reliability of the forecast. The Random Forest algorithm relies on this majority voting principle to make its predictions. Essentially, the Random Forest comprises a collection of individual decision trees working together. Each tree provides a class prediction, and the class with the highest number of votes becomes the final predicted result of the model.

## IV. LITERATURE SURVEY

Shivnath Ghosh et al. [4] utilized the Back Propagation Neural Network (BPN) and Supervised Learning to identify the correlation percentages of factors like organic matter, plant nutrients, and micronutrients that influence crop growth. According to their study, BPN is expected to play a significant role in providing technological support to the agricultural sector in the future. The learning process of BPN is divided into three main stages:

1. **Feed-forward Phase**
2. **Back Propagation of Errors**
3. **Weight Adjustment**

Zhihao et al. [5] applied Support Vector Machine (SVM) and Relevance Vector Machine (RVM) techniques to predict soil moisture levels. For their research, they employed electronic devices such as the “MicaZ mote” and “VH400.” These machine learning techniques required extensive datasets, so they used historical data from Illinois. Their model achieved a 15% error rate and demonstrated a high correlation of 95%.

Vinciya et al. [6] implemented Multiple Linear Regression (MLR) techniques to analyze crop yield predictions. Additionally, they utilized the decision tree algorithm for structured prediction and a supervised learning algorithm for classification tasks. Their research incorporated three primary algorithms:

1. **Decision Tree Algorithm (Structured Prediction)**
2. **Classification (Supervised Learning Algorithm)**
3. **Prediction (Multiple Linear Regression)**

**Decision Tree Algorithm:**

The Decision Tree algorithm is employed for structured prediction or learning, which involves predicting structured outputs rather than scalar discrete or real values. This supervised machine learning method processes data in a hierarchical manner, making it effective for structured predictions. For instance, a decision tree evaluates various conditions to arrive at the most accurate prediction.

$$\hat{y} = \arg\max \{y \in GEN(c)\} (d^1 \Phi(c, y))$$

Updated d,from

$$\hat{y} \text{ to } g: d=d+h(-\Phi(c, \hat{y})+ \Phi(c, g))$$

**CLASSIFICATION (SUPERVISED LEARNING ALGORITHM):**

The risk R(k)R(k) of a function kk is defined as the as: expected loss associated with kk. This risk can be approximated using the training data and is expressed mathematically

$$Remp(k) = 1/N \sum L(yi, k(ci))$$

Where:

- LL is the loss function,
- xx represents the input data,

- yy is the actual output, and
- k(x)k(x) is the predicted output from the model.

This estimation allows for assessing the performance of the function kk using available training data.

**Multiple Linear Regression:**

The values fit in $\hat{y}_i$ is given by the equation

**fo+ f1fi1+ ... + fp Cip**

The value of yi – yˆi is then calculated using the residuals yˆi, which is the difference between the observed and fitted values. When the residuals are calculated, they equal zero (0). The mean-squared error (MSE), commonly referred to as the variance σ2, is calculated by

$$s^2 = \frac{\sum ei2}{n-p-1},$$

The square root of the Mean Squared Error (MSE) serves as an estimate of the standard error, providing a measure of the model's prediction accuracy.

Ying Ding et al. [7] employed Model Predictive Control (MPC) methods, stating that MPC is more effective than the traditional Process Control Model, which originated in industrial applications. MPC is advantageous because it can address non-linear systems and handle large time delays effectively. The development of MPC has been categorized into the following phases:

1. **Classical MPC (1960)**
2. **Improved MPC (2007)**
3. **Latest MPC (2011)**

Each period of MPC development includes several classifications, each with specific advantages and limitations.

Anuja Chandgude et al. [8] utilized Machine Learning, sensors, and Artificial Neural Networks (ANN) to analyze crop growth, predict yields, and identify crop diseases. For data prediction, they employed artificial neural networks, with the following types:

- Perceptron
- Multi-layered Perceptron
- Recurrent Neural Network
- Self-Organizing Maps

The architecture design incorporated components such as:

- Sensors
- Microcontroller
- XBee Module
- IoT Module
- Prediction Modules

Subhadra Mishra et al. [9] examined environmental factors that significantly affect crop production, including weather conditions, soil properties, fertilizer application, and irrigation practices. They used a variety of machine learning techniques, including:

- Artificial Neural Networks
- Information Fuzzy Networks
- Decision Trees
- Regression Analysis
- Clustering
- Principal Component Analysis
- Bayesian Belief Networks
- Time Series Analysis

- Markov Chain Models

Noran S. Ouf et al. [10] explored several processes like crop disease prediction, yield forecasting, weather prediction, and smart crop identification. For crop yield prediction, they utilized Back Propagation Neural Networks and Multiple Linear Regression models. They considered factors such as precipitation, crop biomass, soil evaporation, transpiration, and fertilizer application. Additionally, decision trees, support vector machines, and hidden Markov models were employed in their machine learning framework. They classified machine learning techniques into supervised and unsupervised learning categories.

Fabrizio Balducci et al. [11] and his team divided their research into several phases:

1. **Data Sources:**
    - **CNR (National Research Council Scientific Datasets)**
    - **Istat Statistical Datasets**
    - **IoT Sensors**
2. **Machine Learning Task Design:**
    - Forecasting future datasets (Istat)
    - Comparing machine learning algorithms on incomplete datasets (CNR Scientific Datasets)
    - Reorganizing incomplete data from observation terminals using neural networks, linear regression, and polynomial regression models (IoT sensor datasets).

These approaches highlight the role of advanced technologies in improving agricultural practices and crop yield predictions.

• Reorganizing missing data from observation terminals using Decision Tree, Polynomial Regression, and K-Nearest Neighbors (KNN) algorithms (IoT Sensor Dataset).

• Identifying malfunctioning observation terminals based on sensor values (IoT Sensor Datasets).

## IV. EXPERIMENTAL RESULTS

The primary objective of this research is to predict crop yield under specific climatic conditions. Initially, data was collected, and various algorithms were applied to assess their prediction accuracy. The algorithms tested include Linear Regression, Decision Tree Regression, Random Forest, and XGBoost. The data used for this study was sourced from the official government website **www.data.gov.in**. The datasets included information on maximum and minimum temperatures, seasonal rainfall, cultivated land area, and rice production. This data was arranged district-wise for four states: Tamil Nadu, Andhra Pradesh, Kerala, and Karnataka, covering the years 2000 to 2014.Figure 1 presents the comparison of $R^2$ values for the algorithms applied to the dataset. The $R^2$ value represents the accuracy of each algorithm. For Linear Regression, the $R^2$ value is 0.8885, indicating an accuracy of 88.85%. Similarly, Decision Tree Regression achieved an $R^2$ value of 0.8922, corresponding to an accuracy of 89.22%. For Random Forest and XGBoost algorithms, the $R^2$ values were 0.9314 and 0.9391, translating to accuracy levels of 93.14% and 93.91%, respectively.

The higher the $R^2$ value, the more accurate the algorithm is. Based on the results, the XGBoost algorithm demonstrated the highest $R^2$ value, making it the most effective algorithm for predicting crop yield.

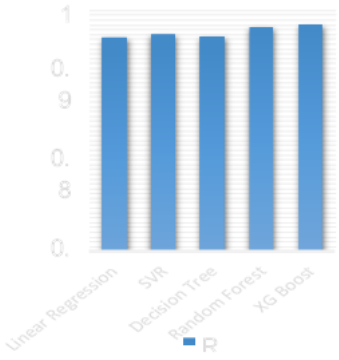


**Figure 1: - Comparison of $R^2$**

Figure 2 presents a comparison of the Mean Square Error (MSE) for four different algorithms. Note that the MSE for Support Vector Regression (SVR) is not included, as the normalized value was not applied. The MSE for Linear Regression is 2,972,466,269.23, while for the Decision Tree Algorithm it

is 2,847,222,194.96. The Random Forest Algorithm shows an MSE of 2,178,239,719.75, and the XGBoost Algorithm has the lowest MSE at 1,999,378,847.49. A lower MSE indicates higher accuracy, making XGBoost the most effective algorithm in this comparison.

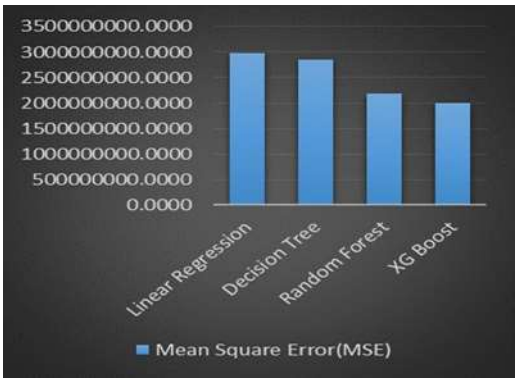


**Figure 2: - Comparison of Mean Square Error (MSE)**

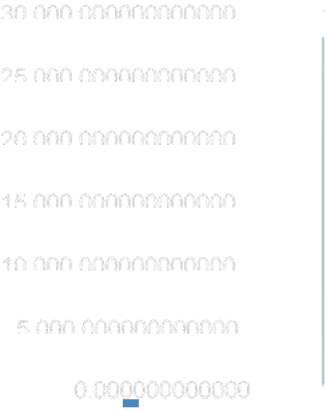

**Figure 3: - Comparison of Mean Absolute Error (MAE)**

Figure 3 illustrates a comparison of the Mean Absolute Error (MAE) for five different algorithms. The MAE for Linear Regression is 27,883.63, for SVR it is 27,402.17, while the Decision Tree Algorithm has an MAE of 24,187.31. The Random Forest Algorithm has an MAE of 21,699.02, and the XGBoost Algorithm has the lowest MAE at 20,613.24. As with MSE, a lower MAE indicates higher accuracy, making XGBoost the most accurate algorithm in this comparison.

**Table 1: - Comparison of $R^2$, MSE, MAE**

| **MODEL** | **R2** | **MSE** | **MAE** |
|---|---|---|---|
| LINEAR REGRESSION | 0.8885 | 2972466269.2335 | 27883.6268 |
| SVR | 0.8991 | NA | 27402.1696 |
| DECISION TREE ALGORITHM | 0.8922 | 2847222174.9642 | 24187.3076 |
| RANDOM FOREST ALGORITHM | 0.9314 | 2178239719.7533 | 21699.0175 |
| XGBOOST ALGORITHM | 0.9391 | 1999378847.4874 | 20613.2361 |

## V.CONCLUSION

Based on the results, it is clear that the XGBoost algorithm outperforms the other algorithms in terms of accuracy, whether measured by R2, MSE, or MAE. The data used to develop the model was sourced from data.gov.in, covering the period from 2000 to 2014. The input variables include area of cultivation, maximum temperature, minimum temperature, and rainfall, while the output variable is production. The proposed XGBoost model demonstrates superior performance when compared to linear regression, SVR, Decision Tree, and Random Forest algorithms. In the future, the plan is to optimize the hyperparameters of the traditional algorithms and evaluate them using new datasets.

**REFERENCE**

Here are the references you provided, formatted in a consistent style:

[1]Omics Online. (n.d.). Agriculture's Role in the Indian Economy. Retrieved from https://www.omicsonline.org/open-access/agriculture-role-on-indian-economy-2151-6219-1000176.php?aid=62176

[2]Jagran Josh. (2018). Sector-wise Contribution in GDP of India. Retrieved from https://m.jagranjosh.com/general-knowledge/what-is-the-sectorwise-contribution-in-gdp-of-india-1519797705-1

[3]Statistics Times. (n.d.). Sector-wise GDP Contribution of India. Retrieved from http://statisticstimes.com/economy/sectorwise-gdp-contribution-of-india.php

[4]Ghosh, S., & Koley, S. (2014). Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks. *IJRITCC, 2*(2), 292-297.

[5]Hong, Z., Kalbarczyk, Z., & Iyer, R. K. (2016). A Data-Driven Approach to Soil Moisture Collection and Prediction. *IEEE Xplore, 2*(2), 292-297.

[6]Vinciya, P., & Valarmathi, A. (2016). Agriculture Analysis for Next Generation High Tech Farming in Data Mining. *IJARCSSE, 6*(5).

[7]Ding, Y., Wang, L., Li, Y., & Li, D. (2018). Model Predictive Control and Its Application in Agriculture: A Review. *Computers and Electronics in Agriculture, June 2018*.

[8]Chandgude, A., Harpale, N., Jadhav, D., Pawar, P., & Patil, S. M. (2018). A Review on Machine Learning Algorithm Used For Crop Monitoring System in Agriculture. *IRJET, April 2018*.

[9]Mishra, S., Mishra, D., & Santra, G. H. (2016). Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. *Indian Journal of Science and Technology, October 2016*.

[10]Ouf, N. S. (2018). A Review on the Relevant Applications of Machine Learning in Agriculture. *IJIREEICE, August 2018*.

[11]Balducci, F., Impedovo, D., & Pirlo, G. (2018). Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. *Machines, 1 September 2018*.

[12]Wani, H. K., & Ashtankar, N. (2017). An Appropriate Model Predicting Pest/Disease of Crops Using Machine Learning Algorithm. In *ICACCS 2017*.

[13]Analytics Vidhya. (2017). Common Machine Learning Algorithms. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?

[14]Dataquest. (n.d.). Top 10 Machine Learning Algorithms for Beginners. Retrieved from https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/

[15]Data Aspirant. (2017). Random Forest Algorithm in Machine Learning. Retrieved from https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/