# Machine Learning Approaches for Crop Yield Prediction in Punjab, India: A Comparative Analysis

Dr. Manpreet Kaur
*Assistant Professor, Faculty of Computing,*
*Guru Kashi University, Talwandi Sabo, PB, India*
*(Email: apmanpreetkaur@gmail.com)*

***Abstract:*** *This study investigates the implementation of machine learning models on extensive crop data to predict crop yield in Punjab state, India. The primary objective of this research was to determine which machine learning model demonstrates superior performance in providing accurate predictions. Two machine learning models (decision tree and random forest regression) were implemented, and gradient boosting regression was utilized for optimization. The results indicate that gradient boosting regression reduces the yield prediction error by 5%. Furthermore, for the given dataset, random forest regression exhibited superior performance compared to the other models.*

***Keywords:*** *Machine Learning, Crop Yield Prediction, Punjab, Decision Tree, Regression*

## 1. Introduction

Agriculture is a vast field, and crop prediction is crucial for ensuring food security. With climate change and a growing population, accurate forecasting can assist farmers in making informed decisions regarding planting time, crop selection, and resource management. Utilizing data on weather patterns, soil health, and fertilizer application can enhance yield predictions. Technologies such as machine learning and remote sensing are increasingly employed to improve these forecasts, allowing for more precise and timely information. Addressing the challenges of crop yield forecasting is essential for adapting to the changing environmental conditions and optimizing agricultural productivity. Crop yield prediction involves the estimation of the amount of produce that can be harvested from a specific area under certain conditions. The following are some important aspects:

- Accurate prediction helps ensure a sufficient food supply in the face of population growth. Resource Management: Helps in optimizing the use of water, fertilizers, and pesticides. Economic Planning: Farmers and governments can make informed financial decisions based on expected yields.
- Factors Influencing Crop Yield Temperature, rainfall, and sunlight directly affect crop growth. Nutrient content, pH, and soil moisture are crucial for crop health. Crop rotation and pest management also affect crop production.
- Challenges in Yield Prediction Climate variability complicate weather forecasts. Data quality must be accurate and sufficient; otherwise, it can lead to unreliable predictions. Unexpected pest invasions or diseases can drastically affect crop yield.

The primary aim of this research is to examine data and utilize factors such as soil composition, temperature, fertilizer application, land characteristics, and irrigation practices to create a novel approach for predicting crop yields. The remainder of this study is structured as follows. A comprehensive analysis of the literature review is outlined in Section 2, the formulation of the proposed work is presented in Section 3, and the results and discussion are presented in Section 4. In Section 5, we offer our conclusions and recommendations for further research.

In Punjab state, the total geographical land is 5.0362 million hectares, and the land used for agriculture is 4.2 million ha. Fig. 1.1 represents the location of the Punjab state in India and the region of study. Punjab, a state of India, is located in the northwestern part of the country. As of 2023, the gross state domestic product of Punjab was estimated to be approximately Rs. 6.5 lakh crore (Approximately $78 billion). The state has a diverse economy with significant contributions from agriculture, manufacturing, and services. The key sector includes agriculture (particularly wheat and rice production).

## 2. Literature Review

Given limited resources and environmental constraints, farmers face challenges in maintaining optimal crop productivity. Crop yield estimation can be enhanced without compromising quality through the application of machine-learning algorithms. A machine-learning model analyzes factors affecting crop yield to provide more accurate predictions. However, previous research has predominantly focused on developing suitable agricultural environments [1]. Current research primarily emphasizes analytical techniques that provide limited crop information. The extracted data may be insufficient for accurate crop yield prediction. Agricultural production is influenced by climate and weather variations. Favorable weather conditions contribute to high crop yields. High-quality seeds result in increased crop productivity (however, predicting the yield genotype and phenotype of the crop requires examination when using high-quality seeds) [2]. Additional factors, such as water availability, soil nutrient content, and weed presence, can impact crop productivity [3].

Ortiz-Bobea et al. [4] presented a model of weather effects on total productivity factors in agriculture at the global level, with the final model indicating that anthropogenic climate change reduced total productivity factors by 21%. The research in [5] conducted a comprehensive study of climate, water, and crop yield models to identify the impact of climate on crops. The authors in [6] utilized meteorological data to introduce a weather forecasting model for crop yields in Europe. Furthermore, Bornn and Zidek [7] introduced a study based on precision agriculture, employing a statistical model and incorporating spatial dependence for the Canadian Prairies. Suitable conditions for crop growth and yield using AVHRR in Poland were discussed in [8].

To predict maize crop yield, researchers considered remote sensing data of the leaf area index and soil moisture, proposing a model that utilized sequential data integration [9]. A study was conducted using four vegetation indices—SAVI, PVI, NDVI, and GVI—and a neural network-based crop yield prediction model to predict crops [10]. The researchers in [11] applied an ensemble Kalman filter to integrate soil moisture estimation, aiming to reduce errors resulting from ambiguity in the temporal rainfall distribution-based crop

model. Chlorophyll content in the leaf area index also contributes to crop yield prediction [12]. The effect of extreme weather conditions on Mediterranean crops is a significant issue that should be incorporated into crop models for improved forecasting [13].

Precision agriculture employs advanced tools and technologies to optimize soil conditions and crop management for maximum productivity. In precision agriculture, real-time data on agricultural environments and meteorological conditions are collected using sensors deployed on farms, and crop yield predictions are generated to assist farmers in making informed agricultural decisions [14]. The data collected by these sensors are substantial in volume and can, therefore, be processed using big data analytics. The outcomes of such analyses can provide benefits to farmers as well as contribute to national economic development [15]. Big data analytics and machine learning algorithms have the potential to significantly increase crop yield. To implement machine learning algorithms, the current study utilized rice and wheat crops and applied decision tree, random forest, and gradient boosting regression techniques. The objectives of the current study are 1) to implement machine learning techniques to predict crop yield for future years and 2) to validate the results using MAE, MSE, and $R^2$ validation metrics. The majority of previous studies have focused on image-processing techniques and statistical models for prediction. The proposed method employs machine learning, which enhances computational efficiency and predictive accuracy compared to traditional statistical models.


**Figure 1: Punjab state location in India**

## 3. Proposed Work

### 3.1 Data Acquisition

The data for this study were acquired from the agricultural department of the government of Punjab for the years 2000 to 2021 from ten major crop-producing districts, namely

Fatehgarh, Sangrur, Ludhiana, Moga, Ferozepur, Bathinda, Jalandhar, Faridkot, Patiala, and Mukatsar of state Punjab, (Figure 1). The final data focus on two crops including rice and wheat, along with the area (hectare), Production (Tonnes), yield (Tonnes/Hectare) and rainfall in the past 21 years. The data were pre-processed before applying the machine-learning algorithms.

### 3.2 Methods

Initially, data were obtained from multiple government agencies, and the raw data were pre-processed to eliminate irrelevant and unnecessary data. This step also included the conversion of categorical data to numeric data. Furthermore, missing values were identified and filled with the appropriate mean values required. The data were divided into features and labels, which were further divided into training and testing datasets, respectively. Figure 2 depicts the framework of this study.

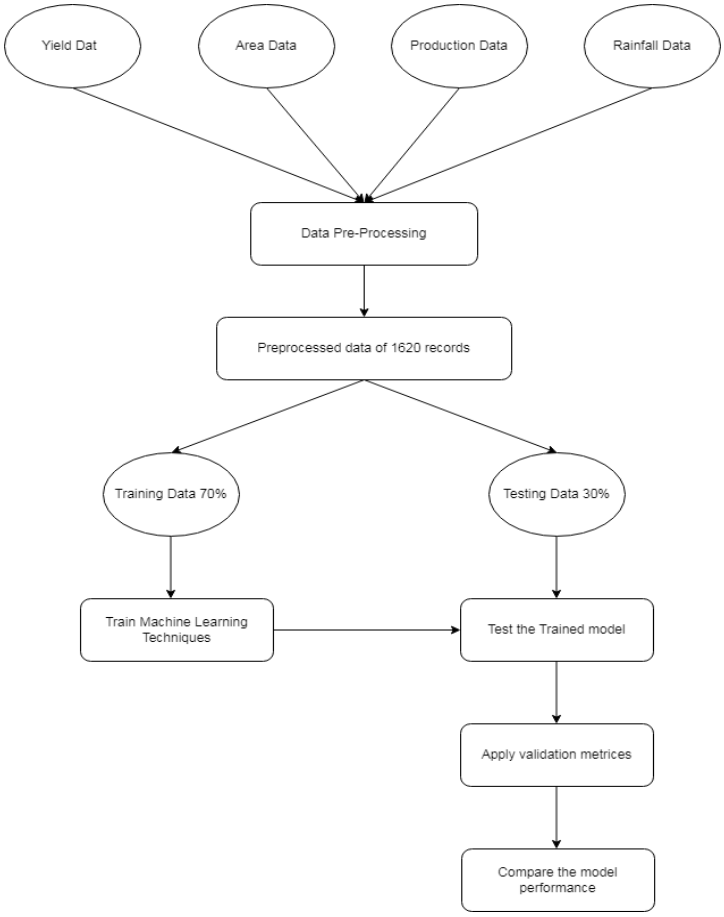


**Figure 2: Framework of the study**

#### *3.2.1 Decision Tree*

The decision tree regression algorithm is in the subdomain of supervised machine learning and can also be utilized for regression/value prediction analysis and classification tasks. For the current study, the objective is to predict the value of the target variable, that is, crop production, by the directives that train the model by providing training data. It

employs a tree to represent the model or potentially solve the problem. In decision trees, attribute selection is the most crucial task, which can be performed through two mechanisms: information gain and Gini impurity. It utilizes various country/state parameters, such as area, area under irrigation, crop year, and crop season (Kharif, Rabi, or Whole Year).

#### *3.2.2 Random Forest Regression*

The Random Forest algorithm is a supervised machine learning algorithm. It creates a forest with numerous trees in a non-deterministic manner. Generally, in a Random Forest Regression, the accuracy increases with a higher number of trees. Random forests effectively manage the challenges posed by missing values and mitigate over fitting when a substantial number of trees are present in the forest. This algorithm primarily consists of two stages: the initial stage involves the creation of a random forest, while the subsequent stage entails extracting predictions from the regression developed in the first stage. It randomly selects a subset of rows from the dataset to create a stump tree and attempts to identify the maximum number of trees for a condition to determine the prediction. The model utilized annη_estimator value of ten and random states 101.

#### *3.2.3 Gradient Boosting Regression*

Gradient boosting combines weak prediction models to generate ensemble models. The gradient boosting algorithm is applicable for regression and classification tasks and fits models that predict continuous values. It employs multiple fixed-size decision trees, selected by the η_estimator parameter, to construct an additive model. The model-fitting process is initialized with a constant value, which is the mean value of the target. In subsequent stages, negative gradients are predicted to fit the estimator. The model employed an η_estimator of 100, random_state of 42, and max_depth of 4. The learning rate was utilized to sequentially add new trees to reduce residual errors in the predictions.

### 3.3 Tables and Figures

#### 3.3.1 Analysis of production over the years for multiple categories

Figure 3 depicts the crop yield (in Tonnes) and land area (in Hectares) for Punjab from 2002 to 2022. In Punjab, the peak crop yield reaches 1,105.54 lakh tonnes across 4.2 million Hectares, while the average yield stands at 102,605.7 tonnes over 97,763.05 hectares. Figure 4 presents an analysis of production for various sample crops, notably Rice and Wheat. Rice occasionally yields surprising results despite the considerable area allocated, whereas wheat consistently exceeds expectations. Figure 5 examines the average production of crops across ten major districts. The graphic reveals that rice and wheat production is substantial in nearly all districts. Figure 6 illustrates the production trends for the same area (Ludhiana) from 2002 to 2022. Shifting patterns prompt us to explore the underlying reasons for these changes, and the factors influencing crop production are not always distinct from one another.

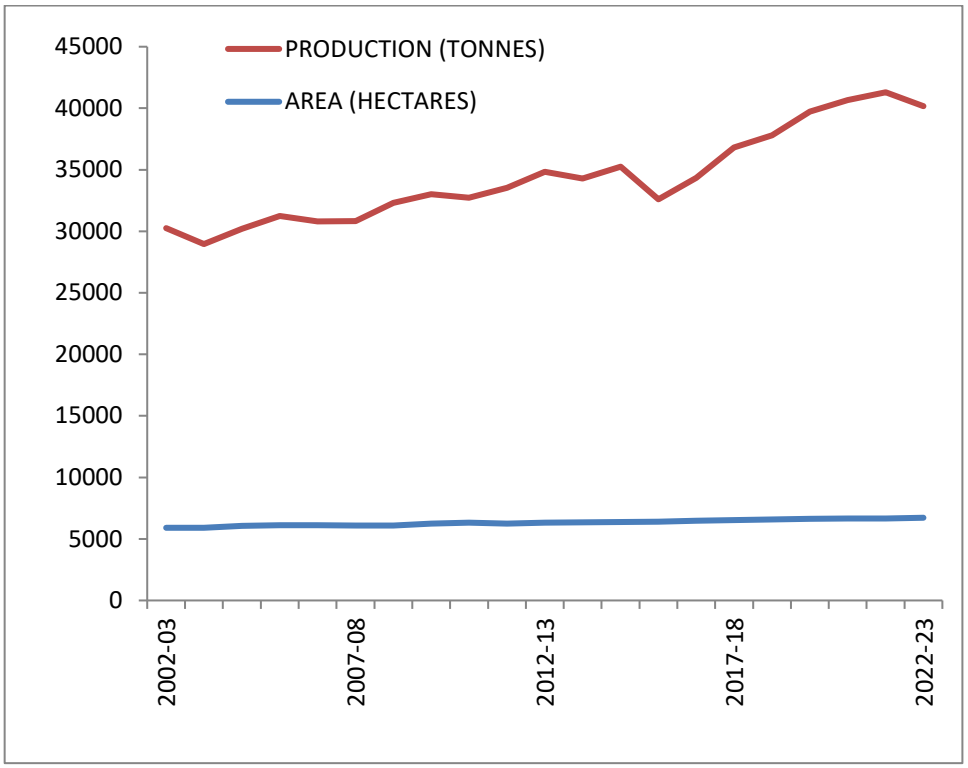


**Figure 3: Analysis of crop production and area over the years**

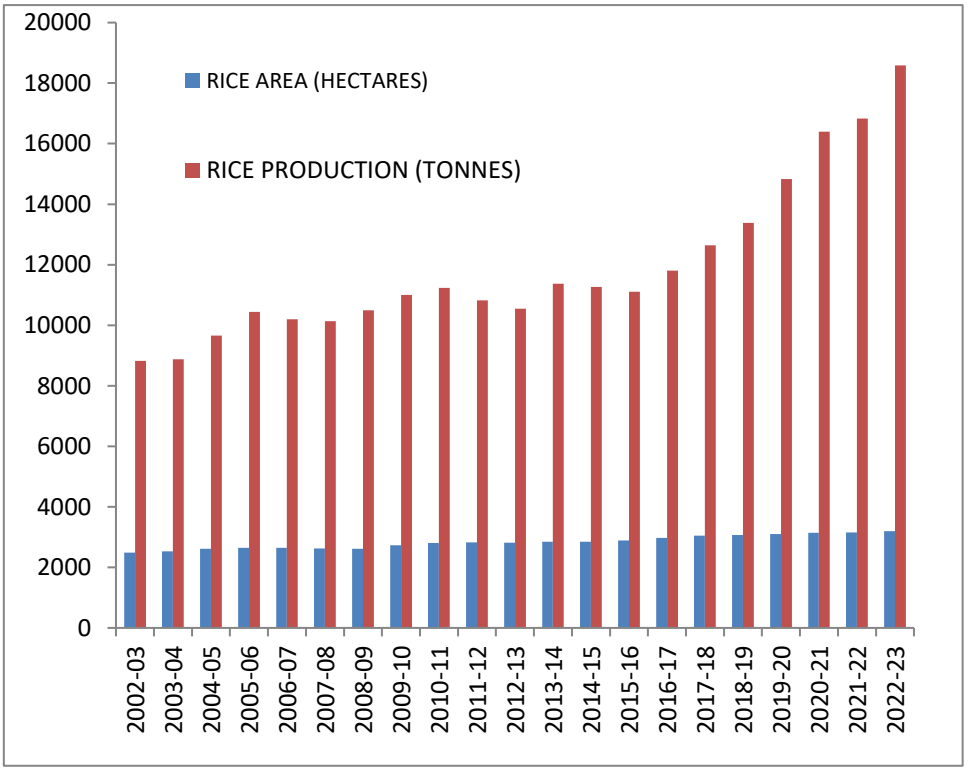


**Figure 4: Analysis of production and area for Rice crop over the years**

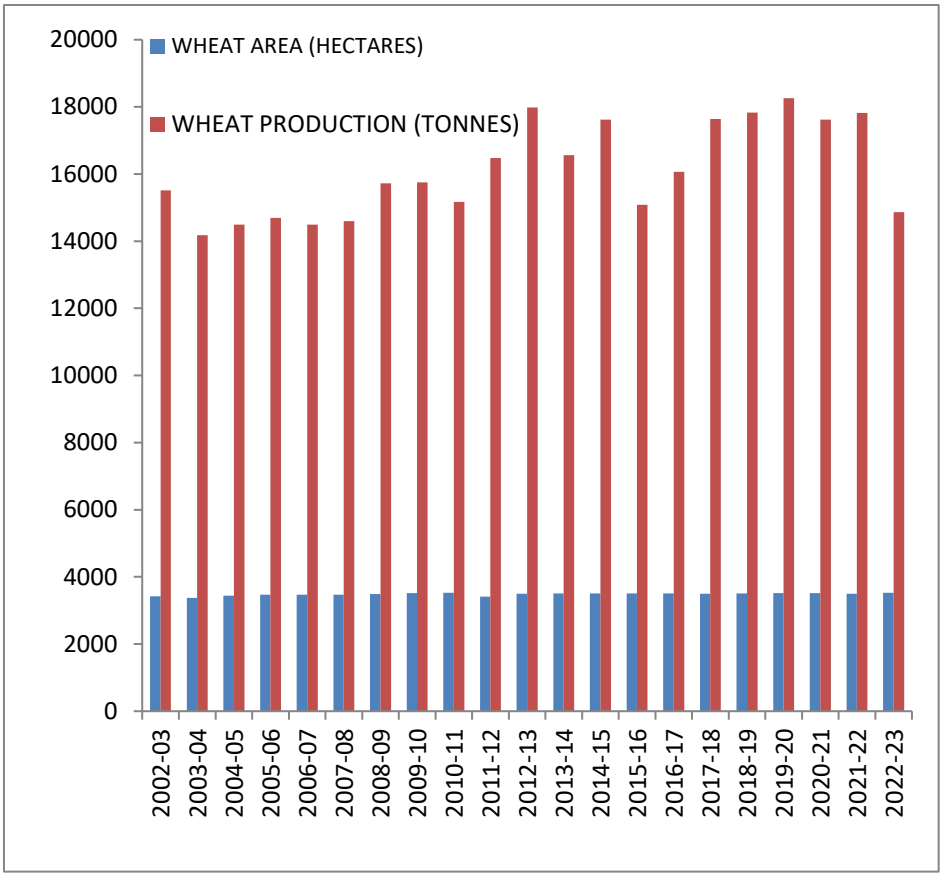


**Figure 5: Analysis of production and area for Wheat crop over the years**

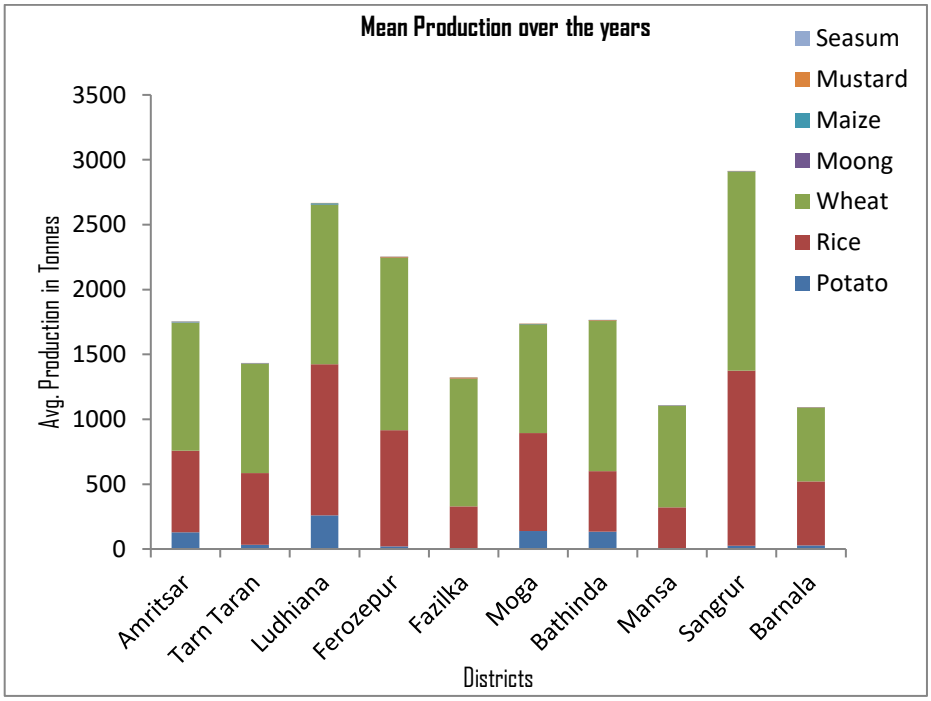


**Figure 6: Crop wise mean production over the years for the selected districts of Punjab**

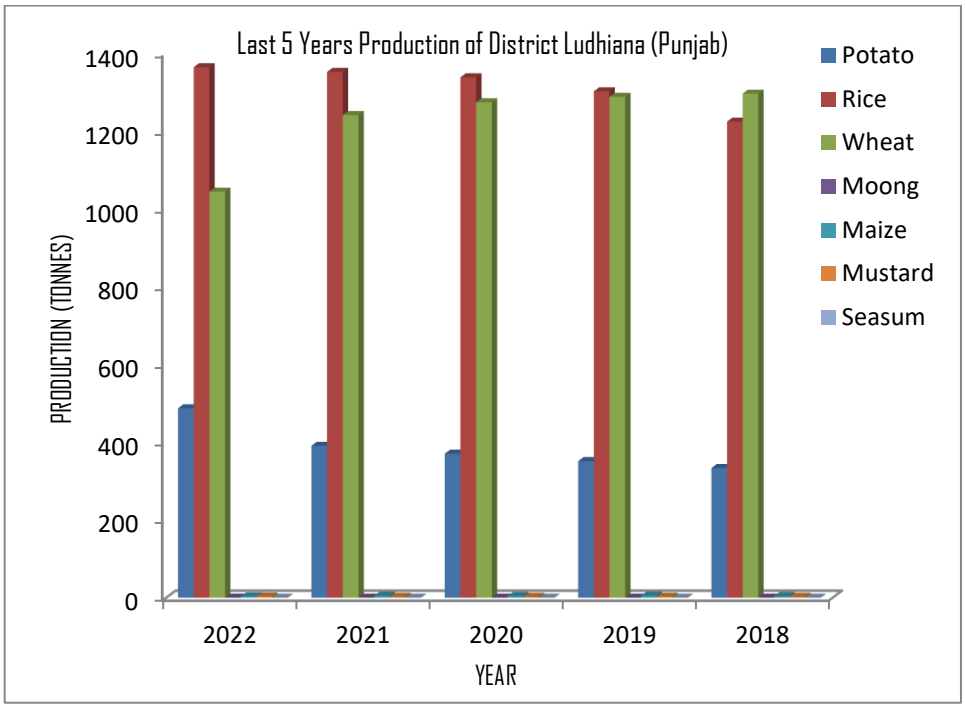


**Figure 7: Analysis of the last five years production for selected crops in Ludhiana District.**

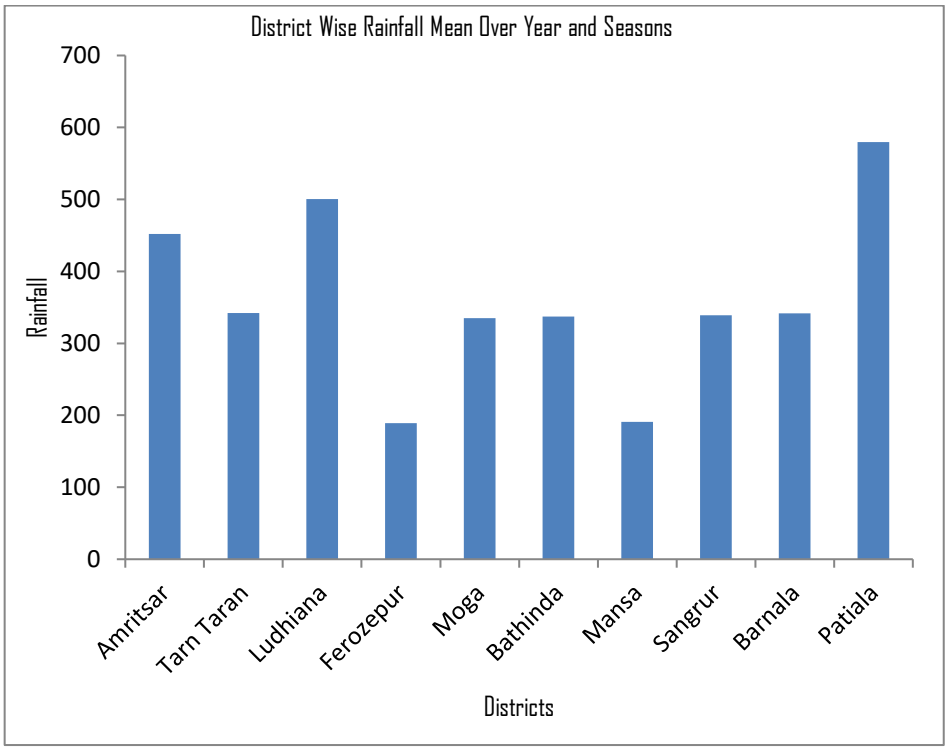


**Figure 8: Analysis of rainfall in selected districts**

### 3.4 Validation Matrix

Model accuracy is assessed with the help of validation matrices, such as mean absolute error (MAE), mean squared error (MSE), and Correlation coefficient ($R^2$). They are demarcated in equation 1, 2 and 3 below:

$$\text{MAE} = \frac{1}{k}\sum_{j=1}^{k}\left|X\frac{p}{j} - X_j\right| \quad (1)$$

$$\text{MSE} = \frac{1}{k}\sum_{j=1}^{k}(X_j - X_j^P)^2 \quad (2)$$

$$R^2 = 1 - \frac{Sum\ of\ Squares\ of\ Residues}{Total\ sum\ of\ squares} \quad (3)$$

MAE and MSE serve as metrics for quantifying the discrepancy between predicted and actual values. The Mean Absolute Error (MAE) quantifies this difference by calculating the average of the absolute deviations across the dataset. In contrast, the Mean Squared Error (MSE) assesses the variance between the actual and predicted values by squaring the average of the differences within the dataset. The coefficient of determination ($R^2$) indicates the proportion of variance in the dependent variable that can be explained by the independent variables; a higher $R^2$ value suggests a more effective model.

## 4. Results And Discussion

This study employed three methodologies—decision tree, random forest, and gradient boosting regression for predicting crop yields, utilizing the Anaconda platform. The outcomes of these methods were compared against those of linear regression, lasso regression, and ridge regression. Table 1 presents a summary of the comparative analysis of the models based on mean squared error and accuracy score.

**Table 1: Accuracy score and mean squared error of Punjab's Agricultural Data**

| Model | Accuracy Score | Mean Squared Error |
|---|---|---|
| **Decision Tree Regression** | 0.9108 | 26431.7033 |
| **Gradient Boosting Regression** | 0.9377 | 22581.3408 |
| **Lasso Regression** | 0.7537 | 57344.8808 |
| **Linear Regression** | 0.7243 | 57348.2639 |
| **Random Forest** | 0.9271 | 23335.6236 |
| **Ridge Regression** | 0.7643 | 57296.2971 |

The table above illustrates that gradient-boosting regression exceeds the performance of all other methods with an accuracy of 93.7%. The regression techniques do not surpass the decision tree and random forest models, which led us to choose the most effective models for our predictions. To implement the machine learning model, the dataset was split into testing and training sets with a ratio of 3:7, meaning 30% of the total data was allocated for testing the model while 70% was designated for training. Consequently, all models were trained using data from 2002 to 2022 to estimate crop yield. The decision tree yielded significant results for this study, achieving an

accuracy of 90.64%, with a Mean Absolute Error (MAE) of 27.30 and a Mean Squared Error (MSE) of 22.63. Decision trees enhance the clarity of predictions, showcasing how each factor influences the outcome. Table 2 summaries the accuracy MAE, MSE, and $R^2$ values for all three methodologies.

**Table 2: Performance of Random Forest, Decision Tree, and Gradient Boosting Regression**

| Model | Accuracy | MAE | MSE | $R^2$ |
|---|---|---|---|---|
| **Decision Tree** | 90.64 | 27.30 | 22.63 | 90.64 |
| **Random Forest** | 91.71 | 23.33 | 16.14 | 91.71 |
| **Gradient Boosting Regression** | 93.71 | 22.68 | 16..01 | 93.77 |

In this research, gradient boosting regression achieved an accuracy score of 93.77% in predicting the crop outcomes for the current dataset. The findings indicate that while the production of certain crops has declined, their market prices continue to rise. This trend suggests a decrease in crop production coupled with sustained demand, as evidenced by the positive slope observed in the data. Furthermore, the analysis highlights that wheat and rice are the predominant crops cultivated across the ten selected districts in Punjab; however, there are numerous other crops that warrant attention for enhanced profitability. The disparity between the production increase of these alternative crops and their demand suggests that they could yield greater economic returns. Table 3 presents a comprehensive list of crops alongside their price fluctuations.

**Table 3: Crops with a slow increase in production but a high increase in prices**

| Crop | Production Variance | Price Variance |
|---|---|---|
| **Sugarcane** | 5337.582915 | 270 |
| **Cotton** | 1858.245163 | 200 |
| **Bajra** | 3758.323689 | 235 |
| **Jowar** | -44011.15643 | 126 |
| **Maize** | 841.898058 | 309 |
| **Sesamum** | -5511.47722 | 224 |
| **Turmeric** | 1595.301173 | 280 |
| **Barley** | 2770.5447 | 283 |
| **Millet** | -2021.2613 | 123 |
| **Pulses** | -154.09443 | 250 |

## 5. Conclusion

The yield of crops is influenced by a multitude of factors, and research in this area is highly beneficial for agricultural practitioners. This study aimed to determine the most effective machine learning methodologies for predicting crop yields in the Punjab region. The research focused on ten selected districts within Punjab, utilizing data spanning from 2002 to 2022. Among the various machine learning algorithms tested, ridge regression, lasso regression, and linear regression did not yield satisfactory outcomes. In contrast,

decision trees and random forests demonstrated significant effectiveness, with the gradient boosting regression achieving the most favorable results for the dataset in question. The predictive outcomes generated by the different techniques were assessed using validation metrics, revealing an $R^2$ value of 93.77 for the gradient boosting regression, the highest among the methods, while the decision tree recorded the lowest at 90.64. This study underscores the advantages of employing machine learning algorithms for crop yield forecasting. Future research could expand to encompass other regions across the country. Identifying crops that exhibit notable trends over time, such as declining yields, may facilitate a deeper understanding of the underlying causes. The findings could assist farmers in making informed decisions regarding crop selection to maximize profits while minimizing risks. Furthermore, enhanced predictive capabilities could enable government agencies to better prepare for irregularities through improved resource allocation, including insurance, logistics, and other essential resources.

## REFERENCES


[1] *A. Morshed, R. Dutta, and J. Aryal, Recommending Environmental Knowledge As Linked Open Data Cloud Using Semantic Machine Learning. 2013.*

[2] *B. Parent and F. Tardieu, "Can current crop models be used in the phenotyping era for predicting the genetic variability of yield of plants subjected to drought or high temperature?," J. Exp. Bot., vol. 65, no. 21, pp. 6179–6189, Nov. 2014.*

[3] *J. Han et al., "Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China," Remote Sens., vol. 12, no. 2, Art. no. 2, Jan. 2020.*

[4] *A. Ortiz-Bobea, T. R. Ault, C. M. Carrillo, R. G. Chambers, and D. B. Lobell, "Anthropogenic climate change has slowed global agricultural productivity growth," Nat. Clim. Change, vol. 11, no. 4, pp. 306–312, Apr. 2021.*

[5] *Y. Kang, S. Khan, and X. Ma, "Climate change impacts on crop yield, crop water productivity and food security – A review," Prog. Nat. Sci., vol. 19, no. 12, pp. 1665–1674, Dec. 2009.*

[6] *P. Cantelaube and J.-M. Terres, "Seasonal weather forecasts for crop yield modelling in Europe," TellusDyn. Meteorol.Oceanogr., vol. 57, no. 3, pp. 476–487, Jan. 2005.*

[7] *L. Bornn and J. Zidek, "Efficient stabilization of crop yield prediction in the Canadian Prairies," Agric. For. Meteorol., vol. 152, pp. 223–232, Jan. 2012.*

[8] *K. Dabrowska-Zielinska, F. Kogan, A. Ciolkosz, M. Gruszczynska, and W. Kowalik, "Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices," 2002.*

[9] *A. V. M. Ines, N. N. Das, J. W. Hansen, and E. G. Njoku, "Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction," Remote Sens. Environ., vol. 138, pp. 149–164, Nov. 2013.*

[10] *S. S. Panda, D. P. Ames, and S. Panigrahi, "Application of vegetation indices for agricultural crop yield prediction using neural network techniques.," Remote Sens., vol. 2, no. 3, pp. 673–696, 2010.*

[11] *A. J. W. de Wit and C. A. van Diepen, "Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts," Agric. For. Meteorol., vol. 146, no. 1, pp. 38–56, Sep. 2007.*

[12] *D. Haboudane, J. R. Miller, N. Tremblay, P. J. Zarco-Tejada, and L. Dextraze, Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture," Remote Sens. Environ., vol. 81, no. 2–3, pp. 416–426, Aug. 2002.*

[13] *M. Moriondo, C. Giannakopoulos, and M. Bindi, "Climate change impact assessment: The role of climate extremes in crop yield simulation," Clim. Change, vol. 104, pp. 679–701, Feb. 2011.*

[14] *A. McBratney, B. Whelan, T. Ancev, and J. Bouma, "Future Directions of Precision Agriculture," Precis. Agric., vol. 6, no. 1, pp. 7–23, Feb. 2005.*

[15] *D. Howe et al., "The future of biocuration," Nature, vol. 455, no. 7209, pp. 47–50, Sep. 2008.*