**Project Report: Crop Yield Prediction Using Machine Learning**

# Abstract

**This study analyzes factors influencing crop yields in Bangladesh, an area contributing 12-14% to the nation's GDP, focusing on enhancing agricultural productivity through predictive modeling. The project aims to optimize wheat production by analyzing soil urea levels, understanding the correlation between wheat yield and rainfall, and developing a crop recommendation system based on various soil and environmental parameters by employing data preprocessing, regression analysis, and machine learning algorithms. The findings offer insights for optimizing wheat cultivation and provide a decision-support tool for farmers, enabling resource savings and cost-effective farming practices. This approach demonstrates significant potential in improving resource allocation and reducing the use of water, pesticides, and fertilizers, offering health and financial benefits in a cost-sensitive developing country context.**

**Keywords:** Crops production, Prediction, Environmental impact, Machine learning, Linear Regression, Logistic Regression, Random Forest, K Nearest Neighbor, Gradient Boosting Regressor, Xtreme Gradient Boosting Regressor.

# 1. Introduction

## 1.1 Project Overview

Agriculture is a vital sector in Bangladesh, influencing the economy and livelihoods. Optimizing agricultural practices through data analysis can lead to better crop yields, efficient use of resources, and sustainable farming practices. This project aims to develop a machine-learning model to predict crop yields for various types of crops using historical data. Accurate predictions can help farmers and policymakers make informed decisions, optimize resource allocation, and improve crop management practices.

For our project, we utilized several regression models to analyze the dataset, which comprised weather factors, soil characteristics, and the yield of crops such as wheat, jute, Aman rice, Aus rice, Boro rice, potato, and jute rice. The models employed included Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Xtreme Gradient Boosting Regressor (XGBoost), K-Nearest Neighbors (KNN) Regressor, and Decision Tree Regressor. Each model was chosen to explore different aspects of the data and leverage their unique strengths in predicting crop yields based on the given environmental and soil variables. By comparing the performance of these diverse regression techniques, we aimed to identify the most accurate and reliable model for forecasting agricultural productivity under varying conditions.

## 1.2 Project Objectives

* To train and evaluate multiple regression models for predicting crop yields.
* To identify the best-performing model for each crop type based on performance metrics.
* To visualize the results and feature the importance of understanding the factors influencing crop yields.

# 2. Data Preparation

## 2.1 Dataset Collection

The dataset used in this project contains historical data on various crops yielded in Bangladesh in the previous years. The data includes information on major crops like Aus, Aman, and Boro rice, jute, potato, and wheat, along with meteorological data such as rainfall and temperature from 2008 to 2017.

## 2.2 Data Preprocessing

Our dataset, obtained from Kaggle, underwent preprocessing to ensure its suitability for machine learning applications, to handle missing values, normalize the features, and split into training and testing sets.

Our dataset, obtained from Kaggle, underwent preprocessing to ensure its suitability for machine learning applications, particularly by addressing the presence of outliers in the aman rice, wheat, and jute columns. For the wheat production data, we used the Interquartile Range (IQR) method to identify and cap outliers: the IQR, calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1), helped us define a range within which typical data points fall. We established the lower and upper bounds for outliers at 1.5 × IQR 1.5×IQR below Q1 and above Q3, respectively. Values below the lower bound were replaced with the lower bound, and values above the upper bound were replaced with the upper bound, thus mitigating the influence of extreme values and preserving the integrity of the data.

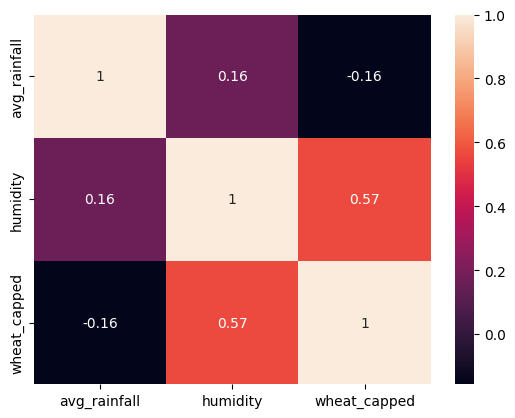
The dataset we used didn’t have any null values in target values. The only place it had missing values was in the name of areas; however, it wasn’t our concern.



**Fig 1: Wheat yield with outliers**



**Fig 2: Wheat yield after capping**



**Figure 3: Wheat yield confusion matrix with outliers**

Our dataset also had columns for different soil types, such as deep brown soil and shallow red brown terrace soil. Different soil types weren’t our concern for this project either; hence, we didn’t change soil type values. The only processing we did was outlier treatment for different types of crops; we examined the relationship between wheat production, average rainfall, and humidity using a heatmap matrix, which revealed a moderate correlation of 0.57 between average rainfall and humidity, suggesting some degree of interdependence between these variables.

To further enhance the dataset’s suitability for machine learning models, we standardized the data using a MinMaxScaler, which scaled all features to a range between 0 and 1, ensuring consistency, improving the model training efficiency, and facilitating better convergence and performance across various algorithms

In short, for the collected dataset, a consistent pre-processing method was applied:

1. Elimination of unnecessary features.

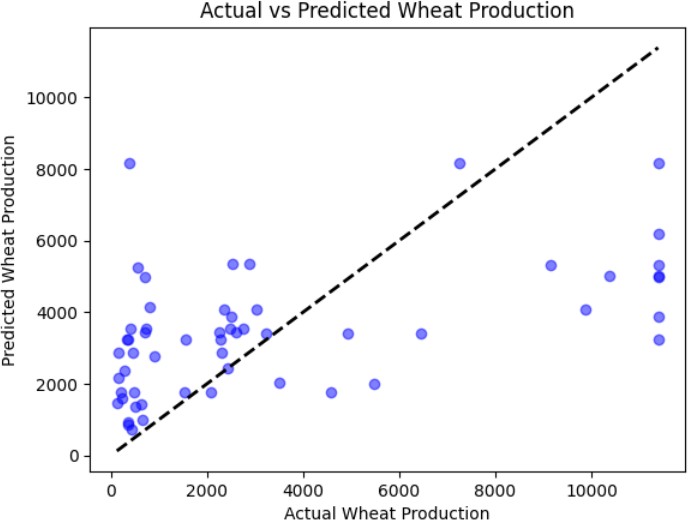
4. Type conversion to ensure uniform float data types.

5. Implementation of min-max scaling for standardized feature scaling.

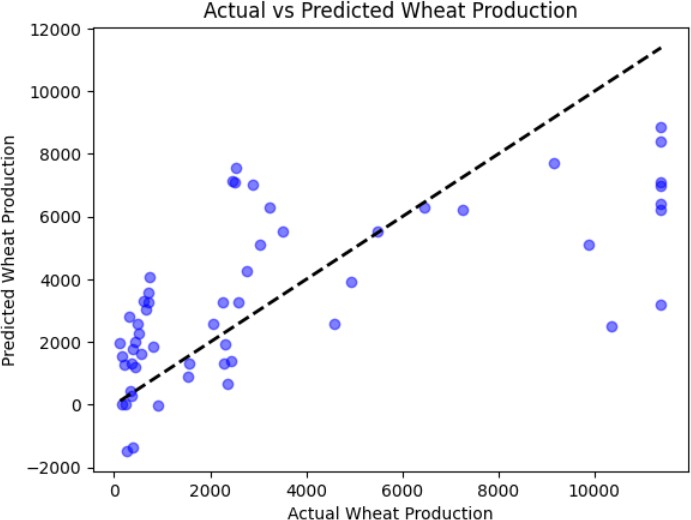
# 3. Methodology

## 3.1 Models Used

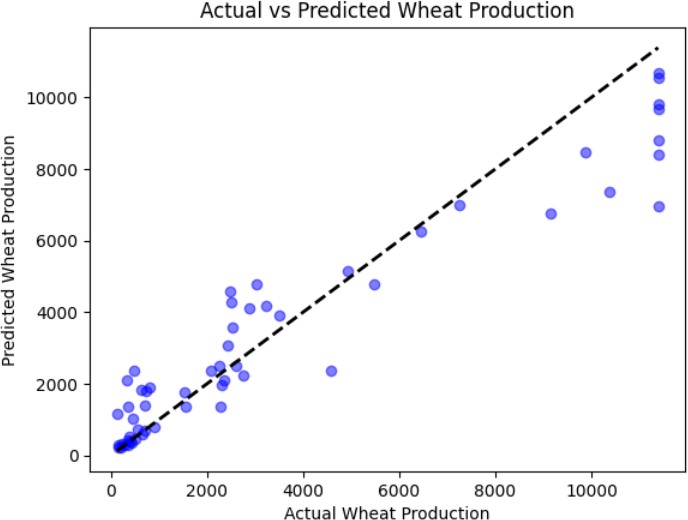
Our project is divided into three parts. In the first part of our project, we employed several machine learning models to predict wheat yield based on average rainfall and humidity. The models used included Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Xtreme Gradient Boosting Regressor (XGBoost), and K-Nearest Neighbors (KNN) Regressor. We trained these models and evaluated their performance using Mean Squared Error (MSE) and R-squared (R2) metrics to ensure accurate predictions.



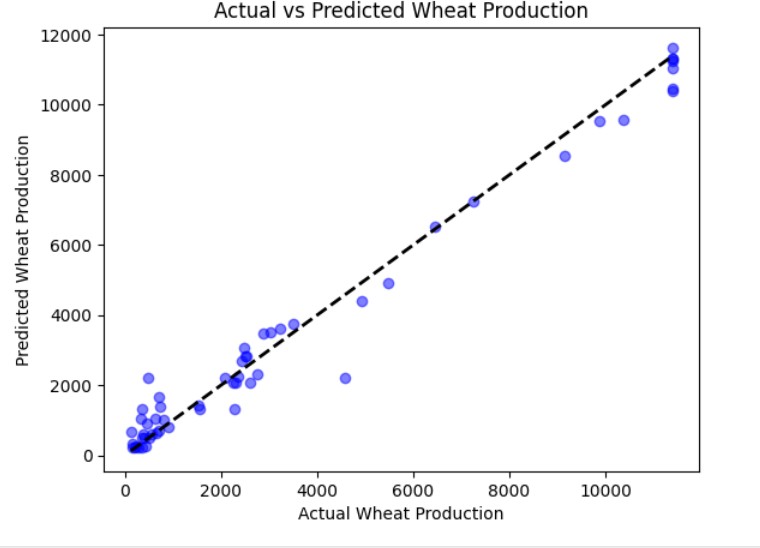
**Fig 4: Predicted wheat production vs actual production using KNN**



**Fig 5: Predicted wheat production vs actual production using Linear Regression**

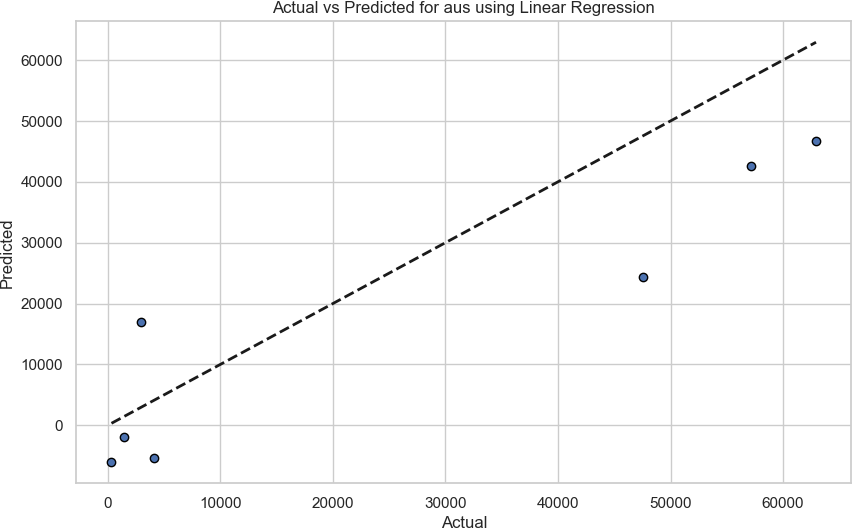


**Fig 6: Predicted wheat production vs actual production using Random Forest Regressor**

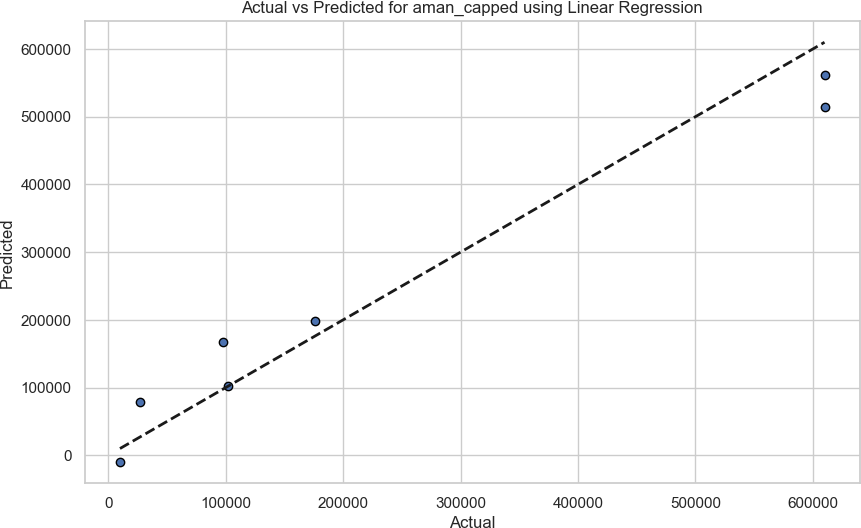


**Fig 7: Predicted wheat production vs actual production using Gradient Boosting Regressor**

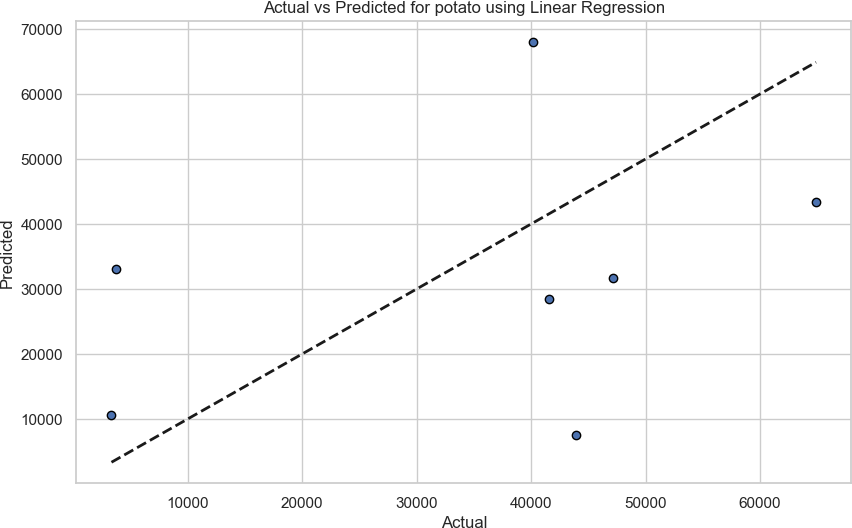
For the second part of the project, we expanded our scope to predict the yields of multiple crops, including aus rice, aman rice, boro rice, jute, wheat, and potato. We utilized Linear Regression, Decision Tree Regressor, Random Forest Regressor, and K-Nearest Neighbors (KNN) Regressor. Similar to the first part, we assessed the performance of these models using MSE and R-squared (R2) values. This comprehensive approach allowed us to compare the effectiveness of different models in predicting agricultural yields under various conditions.



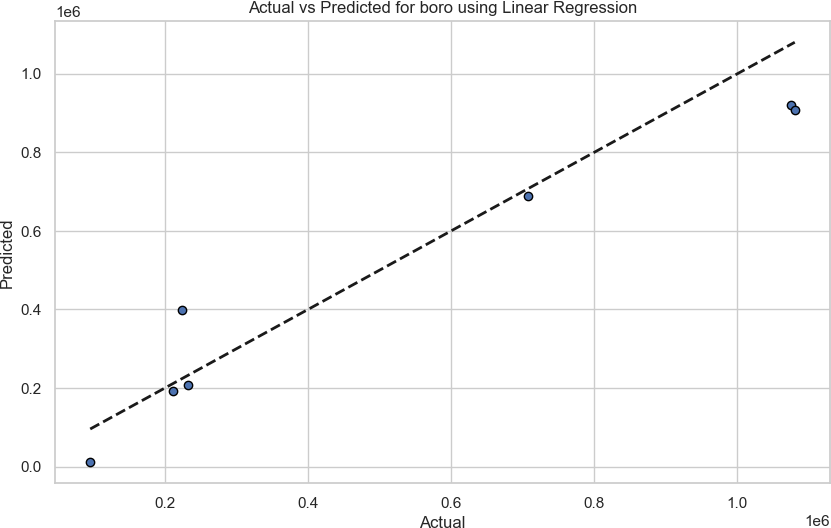
**Fig 8: Actual vs. predicted aus rice production using linear regression**

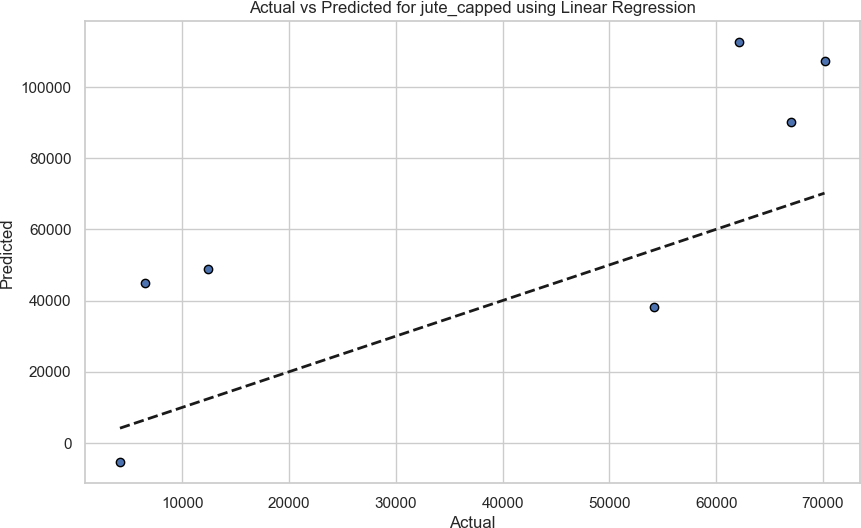


**Fig 9: Actual vs predicted Aman rice production using linear regression**



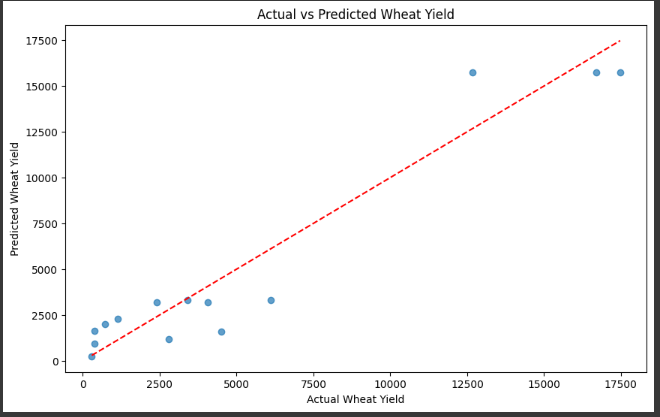
**Fig 10: Actual vs predicted potato production using linear regression**



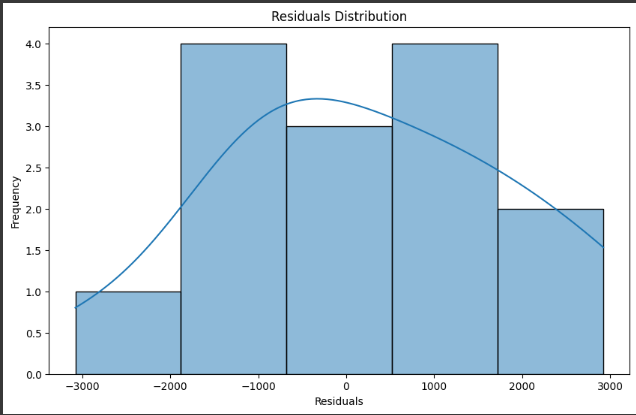
**Fig 11: Actual vs. predicted boro rice production using linear regression**

**Fig 12: Actual vs predicted jute production using linear regression**

The third part of our project analyzes the impact of soil urea levels on wheat production. Again, in this part, the models we trained are Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Xtreme Gradient Boosting Regressor (XGBoost), and K-Nearest Neighbors (KNN) Regressor.



**Fig 13: Actual vs Predicted Wheat Yield**



**Fig 14: Residuals Distribution**

## 3.2 Evaluation Metrics

The models were assessed using the following metrics:

* Mean Squared Error (MSE)
* R-squared (R²)

# 5. Results and Analysis

## 5.1 Evaluation Metrics

For the project's first part, we used average rainfall and humidity values as feature columns and wheat-capped production as the target column. We standardized all values between 0 and 1 using the Min-Max scaler. To split the data into training and testing sets, we selected a test set size of 20 percent and used a random state of 42. We then trained our dataset using several models: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, and K-Nearest Neighbors Regressor. The training metrics used were Mean Squared Error (MSE) and R2. Among these models, the K-Nearest Neighbors Regressor showed the worst performance, with an R2 value of 0.268, while the XGBoost Regressor demonstrated the best performance, with an R2 value of 0.987.

|  |  |  |  |
| --- | --- | --- | --- |
| Target | Model | Metric | Value |
| Aus | Decision Tree  Regression | R2 | 0.9640944275114126 |
| Aman | Decision Tree  Regression | R2 | 0.9716407340668094 |
| Boro | Decision Tree  Regression | R2 | 0.9805250438087878 |
| Jute | Decision Tree  Regression | R2 | 0.9521230822221272 |
| Potato | Decision Tree  Regression | R2 | 0.7810277896236831 |

**Table 1: The best models for each crop type were selected based on the highest R² value in the first part**

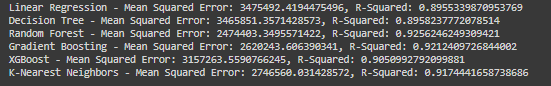
For the second part of the project, we selected rainfall, soil, humidity, and temperature as features, and for the target, we chose aman rice, aus rice, boro rice, jute, and potato. Aman rice and jute had outliers, while potato and boro rice had none. The outliers were capped accordingly. To split the dataset into training and testing sets, we used a test size of 20% and a random state of 42.

The following table summarizes the performance metrics for each model and crop type:

| **Target** | **Model** | **Metric** | **Value** |
| --- | --- | --- | --- |
| aus | Linear Regression | MSE | 192913903.747191 |
| aus | Linear Regression | R2 | 0.7334768954852364 |
| aus | Decision Tree Regression | MSE | 25989057.0 |
| aus | Decision Tree Regression | R2 | 0.9640944275114126 |
| aus | Random Forest Regression | MSE | 249754013.58417144 |
| aus | Random Forest Regression | R2 | 0.6549485870509992 |
| aus | K-Neighbors Regressor | MSE | 151543926.66857144 |
| aus | K-Neighbors Regressor | R2 | 0.7906322093870674 |
| aman\_capped | Linear Regression | MSE | 2811602698.917056 |
| aman\_capped | Linear Regression | R2 | 0.9526012096569408 |
| aman\_capped | Decision Tree Regression | MSE | 1682215686.517857 |
| aman\_capped | Decision Tree Regression | R2 | 0.9716407340668094 |
| aman\_capped | Random Forest Regression | MSE | 2574279990.0776877 |
| aman\_capped | Random Forest Regression | R2 | 0.9566020627377323 |
| aman\_capped | K-Neighbors Regressor | MSE | 7504674826.720001 |
| aman\_capped | K-Neighbors Regressor | R2 | 0.873484077660921 |
| boro | Linear Regression | MSE | 13184261639.528442 |
| boro | Linear Regression | R2 | 0.9162636623079705 |
| boro | Decision Tree Regression | MSE | 3066326100.714286 |
| boro | Decision Tree Regression | R2 | 0.9805250438087878 |
| boro | Random Forest Regression | MSE | 6561082783.070803 |
| boro | Random Forest Regression | R2 | 0.9583290245165198 |
| boro | K-Neighbors Regressor | MSE | 8340350294.354283 |
| boro | K-Neighbors Regressor | R2 | 0.9470284792722873 |
| wheat\_capped | Linear Regression | MSE | 1458581.765076609 |
| wheat\_capped | Linear Regression | R2 | 0.2075268681220972 |
| wheat\_capped | Decision Tree Regression | MSE | 150370.42857142858 |
| wheat\_capped | Decision Tree Regression | R2 | 0.9183011008878452 |
| wheat\_capped | Random Forest Regression | MSE | 791925.9533035716 |
| wheat\_capped | Random Forest Regression | R2 | 0.5697326982578088 |
| wheat\_capped | K-Neighbors Regressor | MSE | 1153232.4914285718 |
| wheat\_capped | K-Neighbors Regressor | R2 | 0.37342850010348516 |
| jute\_capped | Linear Regression | MSE | 1086075428.4148846 |
| jute\_capped | Linear Regression | R2 | -0.38259072857371845 |
| jute\_capped | Decision Tree Regression | MSE | 37609064.571428575 |
| jute\_capped | Decision Tree Regression | R2 | 0.9521230822221272 |
| jute\_capped | Random Forest Regression | MSE | 840876226.2178303 |
| jute\_capped | Random Forest Regression | R2 | -0.07044837202846321 |
| jute\_capped | K-Neighbors Regressor | MSE | 954582244.7742854 |
| jute\_capped | K-Neighbors Regressor | R2 | -0.2151978829060186 |
| potato | Linear Regression | MSE | 554544371.6105651 |
| potato | Linear Regression | R2 | -0.22458306958705987 |
| potato | Decision Tree Regression | MSE | 376705490.5714286 |
| potato | Decision Tree Regression | R2 | 0.16813299423726302 |
| potato | Random Forest Regression | MSE | 260531630.0424857 |
| potato | Random Forest Regression | R2 | 0.42467611326511 |
| potato | K-Neighbors Regressor | MSE | 99160122.18285714 |
| potato | K-Neighbors Regressor | R2 | 0.7810277896236831 |

**Table 2: Metrics for all the models used in the second part of the project**

In the third part of our project, the analysis revealed a positive correlation between soil urea levels and wheat yield, suggesting that optimal urea application can significantly boost wheat production.



**Table 3: Metrics for all the models used in the third part of the project**

## 5.2 Best Model Selection

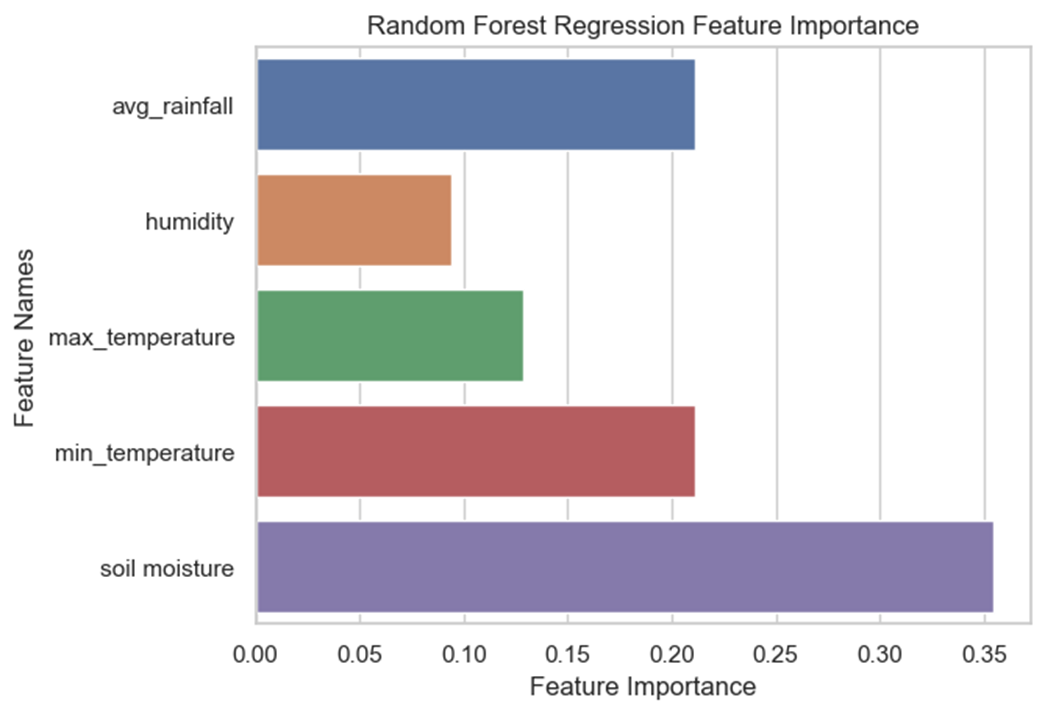
In our second part of the project, the best models for each crop type were selected based on the highest R² value:

|  |  |  |  |
| --- | --- | --- | --- |
| Target | Model | Metric | Value |
| Aus | Decision Tree Regression | R2 | 0.9640944275114126 |
| Aman | Decision Tree Regression | R2 | 0.9716407340668094 |
| Boro | Decision Tree Regression | R2 | 0.9805250438087878 |
| Wheat | Decision Tree Regression | R2 | 0.9183011008878452 |
| Jute | Decision Tree Regression | R2 | 0.9521230822221272 |
| Potato | Decision Tree Regression | R2 | 0.7810277896236831 |

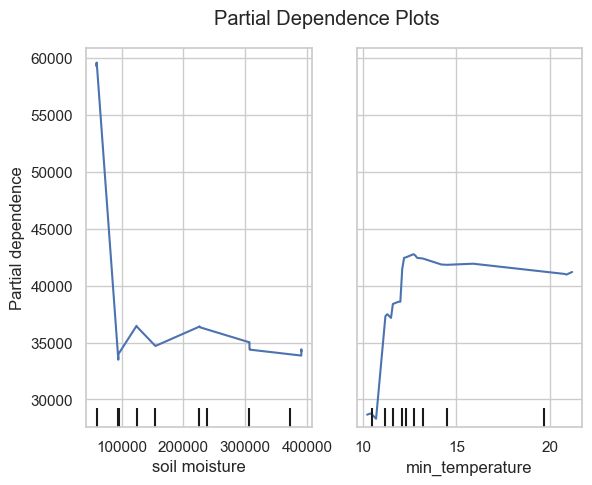
Table 4: Best Model Selection

## 5.3 Visualizations

* Feature Importance:



* Partial Dependence Plots:



# 6. Conclusion

## 6.1 Summary of Findings

## We conducted a study to predict wheat production using various models, such as Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, and K-nearest neighbors Regressor. The XGBoost Regressor outperformed the other models with an R² score of 0.987, indicating its effectiveness in capturing the relationship between features (average rainfall and humidity) and wheat production.

## For the second part of the project, we aimed to predict the production of various crops, including aman rice, aus rice, boro rice, jute, and potato, using features such as rainfall, soil type, humidity, and temperature. The Decision Tree Regressor performed the best among the models tested, providing accurate predictions for all crops mentioned. It demonstrated robustness across different types of crop data, particularly excelling in handling non-linear relationships and interactions within the dataset. However, it did not perform as well for potato production, for which K-Neighbor Regression showed competitive performance.

## 6.2 Future Work

* Incorporate more advanced models and techniques such as Gradient Boosting and Neural Networks.
* Explore hyperparameter tuning to improve model performance further.
* Extend the analysis to additional crops and regions.