Crop Yield Prediction Using Machine Learning

Capstone DATA 2206 : Professor Sam Plati

Group Name: *Team Gigabyte*

Group Members:

Prashant Verma (100967364),

Nana Esi Hinson (100957828),

Sheetalben Mukeshbhai Jadav (100951636),

Sayali Kumbhar (100950732)

Date: December 06, 2024

Background

Hook: Let's say a farmer finds it difficult to forecast how much crop they will gather each season.
Every error in judgement has a price: excessive fertiliser use, water waste, and a precarious reliance on erratic weather. Imagine being able to precisely forecast agricultural yields, which would allow farmers to make prudent resource allocations and prepare for climate threats before they become more severe.

Problem Statement: Predicting crop yields is a significant problem in the agriculture industry. Predictability is hampered by the unpredictability of the interactions among weather patterns, soil characteristics, and agricultural techniques, which affects sustainability, resource allocation, and global food security.

Objective: Create a machine learning model that can reliably forecast crop yields by analyzing agricultural, environmental, and soil data. This will allow for data-driven decisions to be made in order to maximize resource allocation, improve sustainability, and increase global food security.

Significance: To improve food security and increase agricultural production, farmers and agribusinesses can implement sustainable practices, minimize climatic risks, and maximize resource use with the help of accurate crop yield forecasts.

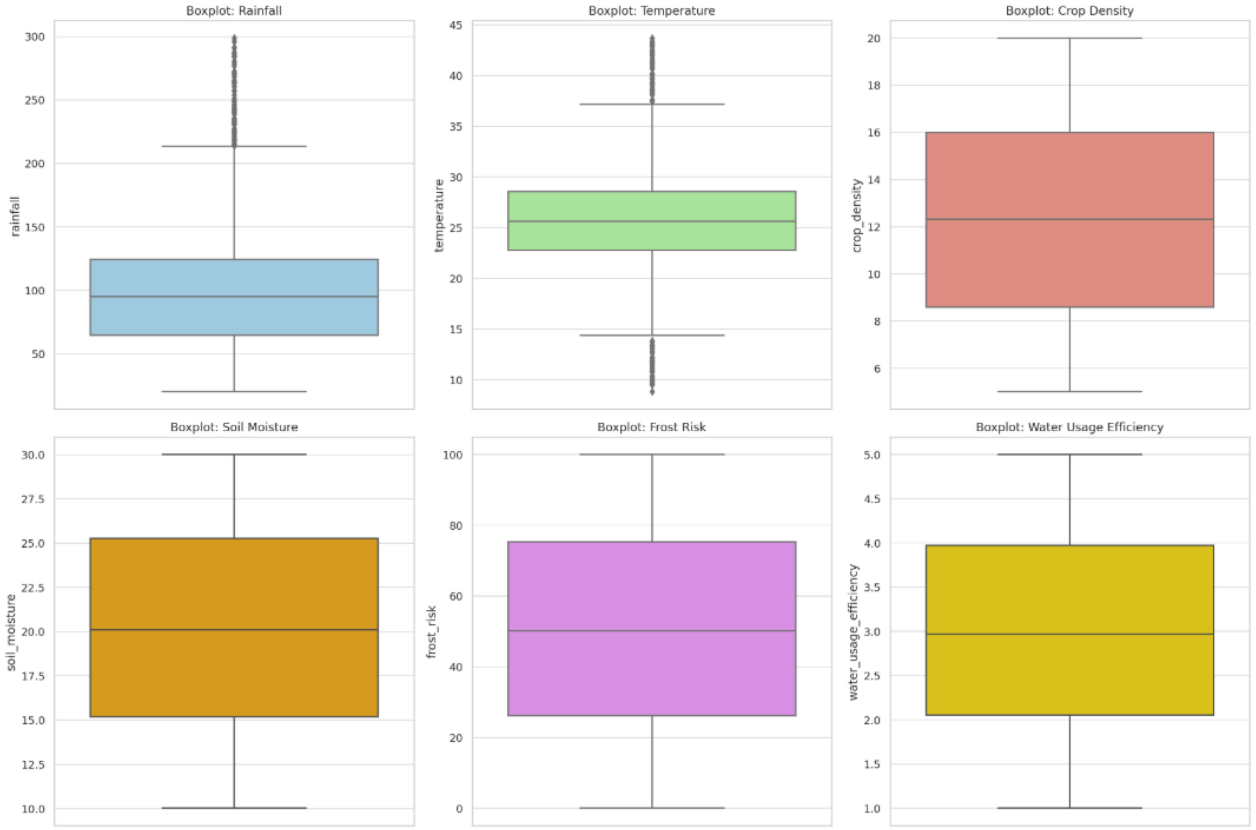
Data Overview

Data Source: Agricultural data from a dataset of 2,200 entries with 23 attributes, including soil characteristics, farming methods, environmental factors, and contextual variables including crop density, rainfall, fertilizer use, nitrogen levels, and proximity to cities.

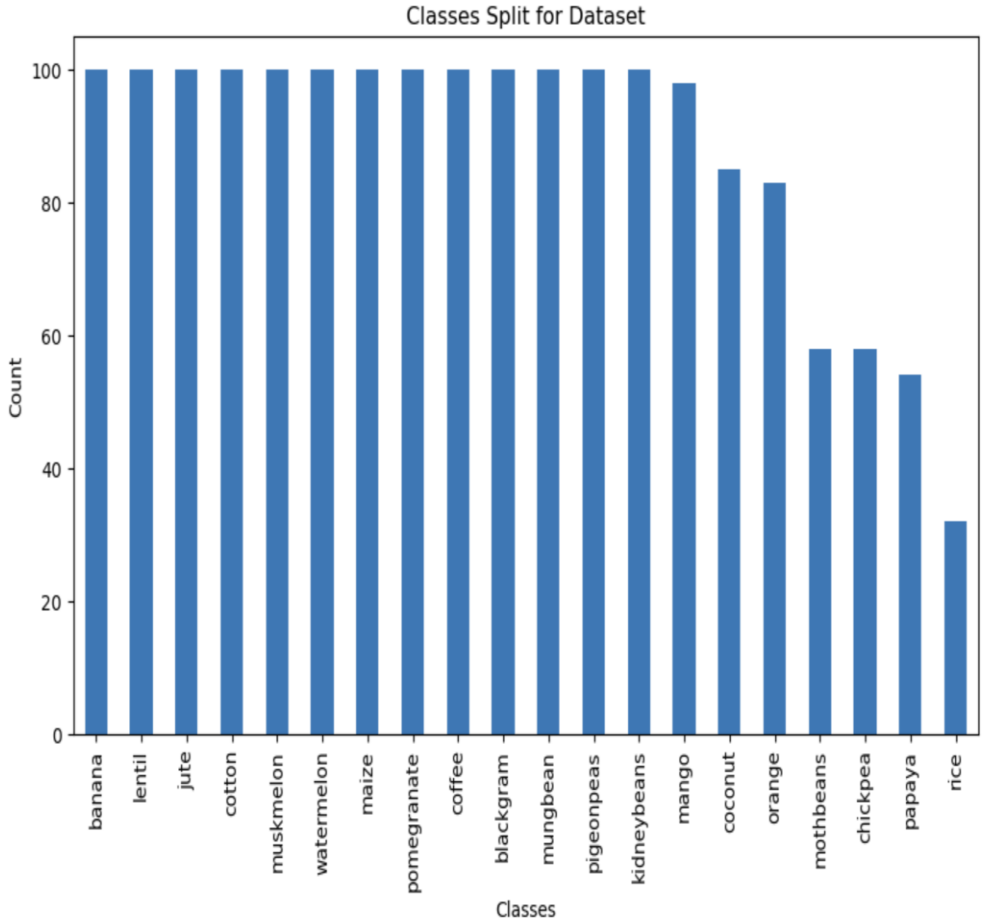
Data Type : Regression-based supervised learning issue with the goal of forecasting continuous crop yields based on different soil, environmental, and agricultural input characteristics.

Key Variables: Nitrogen levels, phosphorus, potassium, soil pH, organic matter, temperature, humidity, rainfall, soil moisture, sunlight exposure, fertilizer usage, irrigation frequency, pest pressure, crop density, and frost risk.

Exploratory Data Analysis (EDA)

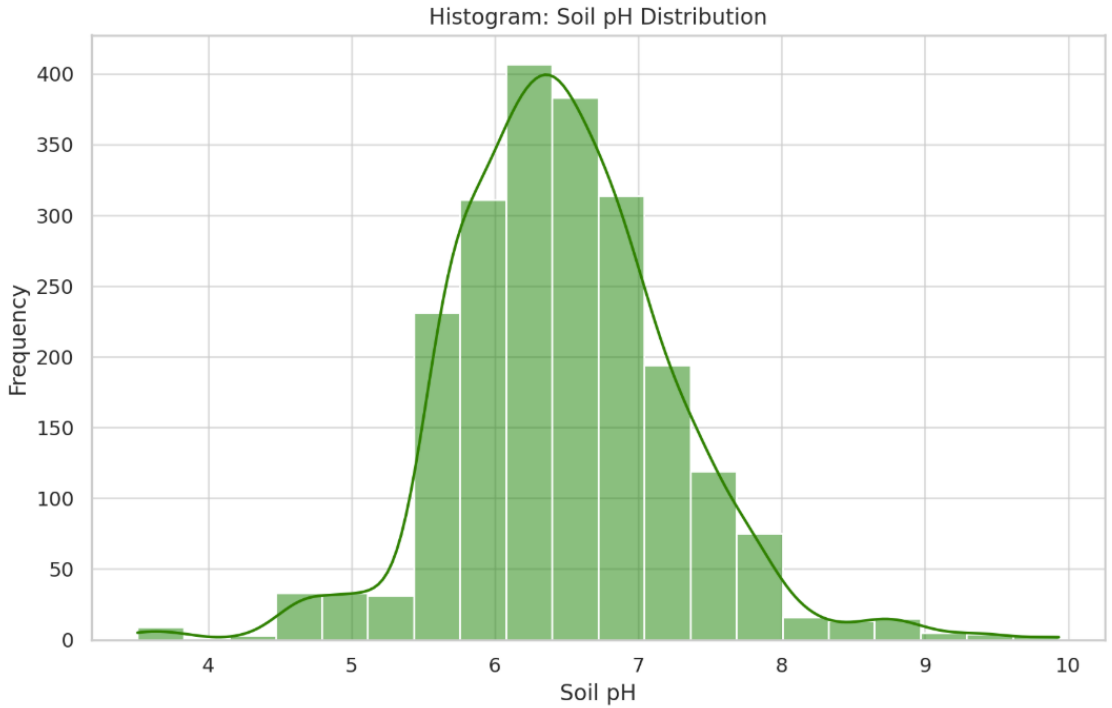


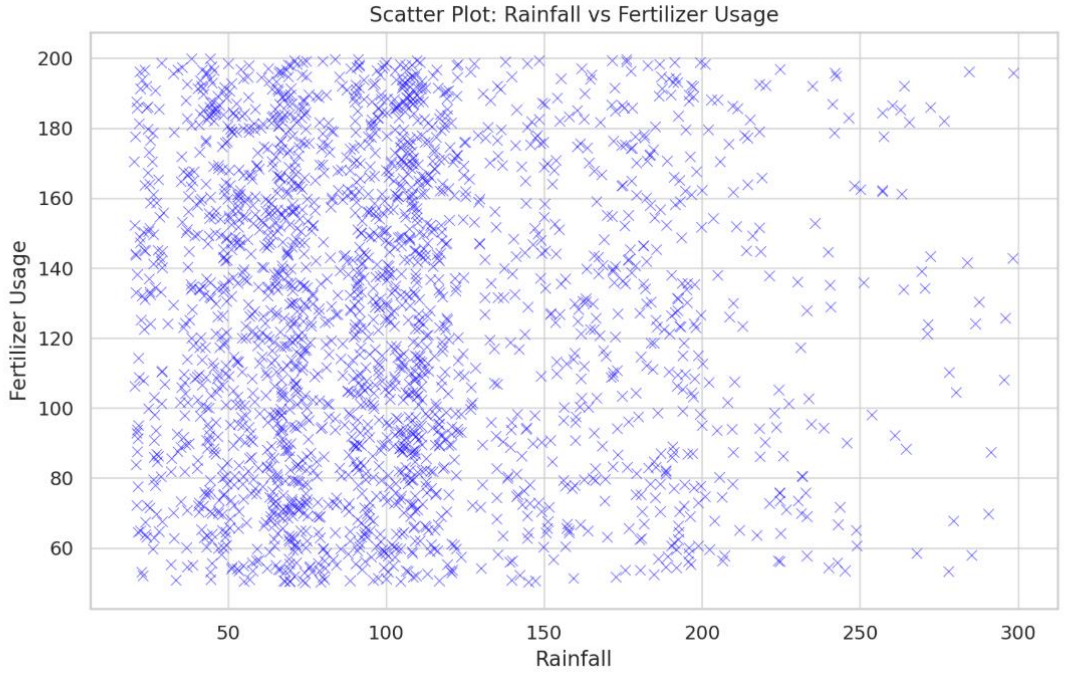
Outliers: Rainfall, temperature, crop density, soil moisture, frost risk, and water usage efficiency were addressed during preprocessing.



Tukey Method : Above is the cleaned dataset after the outlier detection and elimination.

Exploratory Data Analysis (EDA)





Histogram: Soil pH distribution shows a typical spread, confirming its stability.

Scatter Plots: Rainfall and fertilizer usage show clear trends and interdependencies.

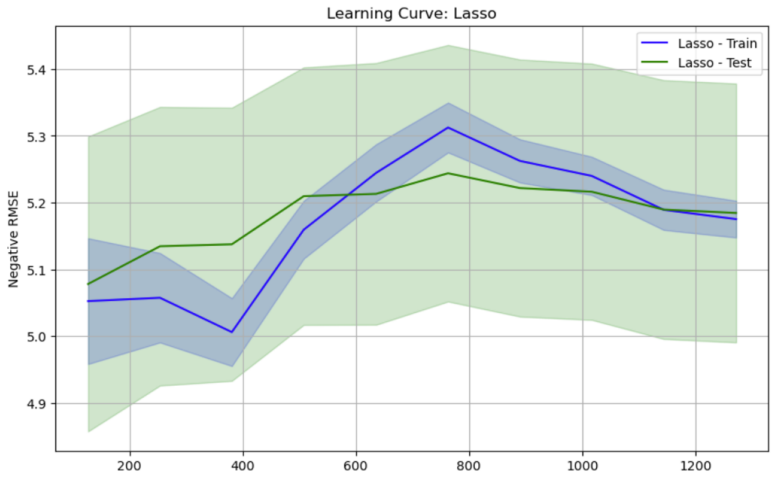
Modeling Approach

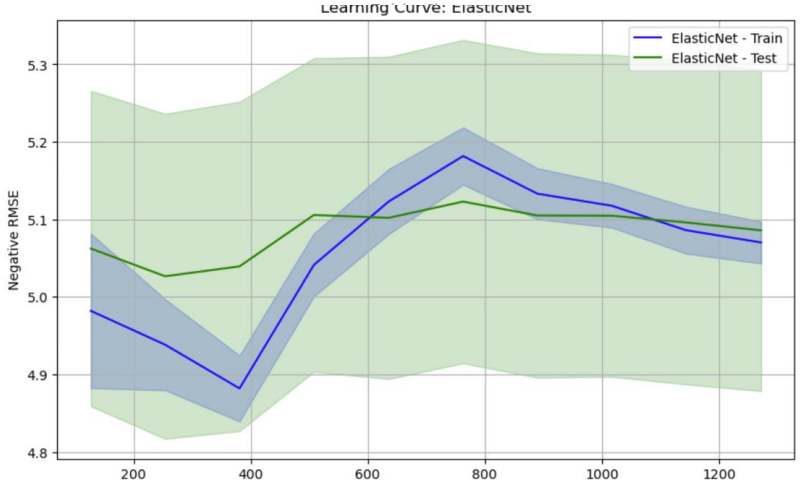
Methods Overview: We applied regression models, including Linear Regression, Random Forest Regression, and Gradient Boosting Regression, to predict crop yields and evaluated their performance using Root Mean Squared Error (RMSE).

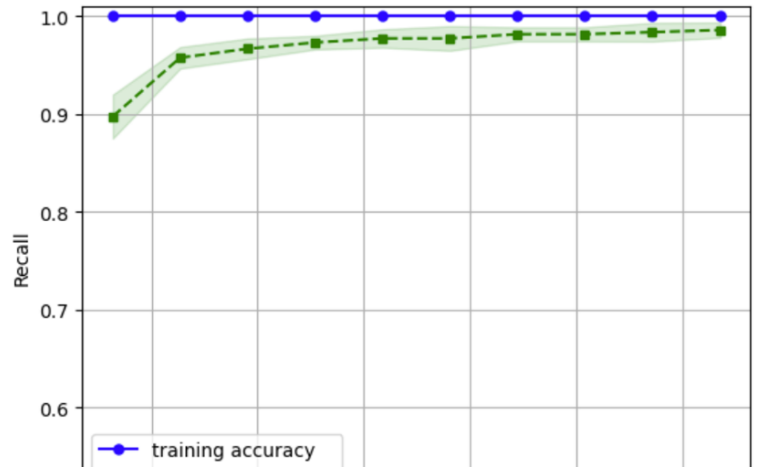
Rationale for Selection:

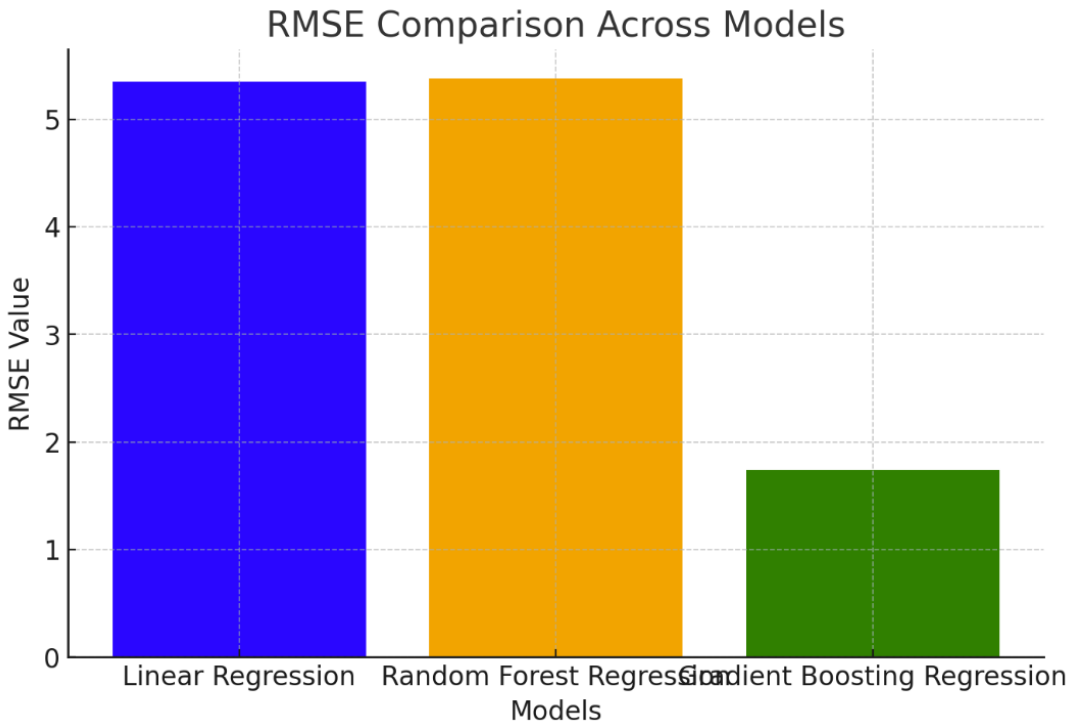
Linear Regression: For establishing a baseline and understanding linear relationships between variables. We also conducted comparative analysis between Lasso, Ridge and Elastic Net. Where, **Lasso and ElasticNet** were the top performers.

Random Forest and Gradient Boosting Regression: To capture complex, non-linear interactions and improve prediction accuracy.









Model Outputs

1.Gradient Boosting Regression:

1. **Test RMSE:** 1.7384
2. **Cross-validation RMSE:** -1.7892
3. **Optimal Hyperparameters:**
    1. *Learning rate: 0.1*
    2. *Max depth: 7*
    3. *Number of estimators: 50*

2.Random Forest Regression:

1. **Test Accuracy:** 99%
2. **Key Predictors:** Rainfall (20%), Humidity (18%), Soil nutrients (N: 12%, P: 13%, K: 14%)

The performance of Gradient Boosting Regression, Random Forest Regression, and Linear Regression is shown in this RMSE comparison graph. The Gradient Boosting Regression model performs better at forecasting crop yields, as evidenced by the lowest RMSE.

Model Evaluation

Key Metrics:
RMSE (Root Mean Squared Error): Determines how big prediction errors are on average. Better crop yield prediction accuracy is indicated by a lower RMSE.

•**Mean Absolute Error (MAE):** provides information about the accuracy of the model's predictions by capturing the average prediction deviation.

R-squared: explains the dependent variable's variance, which aids in evaluating how well the model fits the data.

Why RMSE Matters: The main indicator used to assess how accurate crop production prediction models are is RMSE. Since continuous values are being forecasted, RMSE offers a clear indicator of how closely the anticipated crop yields match the actual results. Reducing RMSE guarantees more precise and trustworthy forecasts, which are essential for maximising agricultural methods and allocating resources.

Implications

Practical Applications:
Agricultural Planning: Crop yields can be predicted using the model, which will enable farmers and agribusinesses more effectively manage resources like fertilizer and water.

Sustainability Initiatives: The model can be used as a guide by governments and non-governmental organizations to develop policies that will increase food security and encourage sustainable farming methods.

Potential Impact:
Reducing waste, optimizing agricultural outputs, and improving resource efficiency are all possible with accurate crop yield forecasts, which would support environmental and economic sustainability. Better management of the world's agricultural resources is made possible by improved forecasting, which can also boost food security.

Recommendations

Actionable Insights:
Resource Optimization: Apply the machine learning model to optimize the distribution of resources, especially fertilizer and water, according to anticipated crop yields.
•**Sustainability Practices:** Enable farmers to embrace sustainable farming methods by offering model insights on the most effective use of resources.
Limitations:
•**Data Coverage:** Under-represented variables or missing data, such as particular geographical features or newly emerging environmental circumstances, may constrain the current model and have an impact on yield forecasts.
Future Work:
•**Real-Time Data Integration:** Increase the accuracy and scalability of predictions by integrating real-time data sources, such as Internet of Things sensors for soil and weather conditions.

Model Enhancement: To increase the model's resilience and capacity to adjust to changing agricultural conditions, investigate adding other characteristics, such as information on crop rotation or insect pressure.

Conclusion

Summary: A prediction model with low RMSE was created as a result of our effort, which also identified important factors affecting crop yields and offered practical advice for enhancing agricultural sustainability and resource allocation.

Final Thought: A more effective use of resources, increased productivity, and a more optimistic outlook for global food security are all made possible by accurate crop yield prediction, which enables farmers, agribusinesses, and policymakers to shift from reactive to proactive tactics.