Seed Yield Prediction Using Machine Learning

Abstract

This project predicts soybean seed yield per unit area (SYUA) using machine learning techniques. By analyzing various factors like plant height, chlorophyll content, protein percentage, and biological weight, we determine the most important features affecting yield. We employ **Random Forest** for feature selection and use **Linear Regression** to predict the yield based on selected features. This study helps farmers and researchers optimize soybean growth and productivity using data-driven insights.

Problem Statement

Soybean is a vital agricultural crop, and accurately predicting its yield is crucial for efficient resource management. Traditional yield estimation relies on expert knowledge and field measurements, which are time-consuming and costly. Machine learning offers an automated, accurate, and scalable approach to predicting soybean yield based on various plant attributes.

Objectives

1. **Preprocess the dataset** by handling missing values, scaling features, and dropping irrelevant columns.
2. **Perform exploratory data analysis (EDA)** to visualize relationships between different plant attributes and seed yield.
3. **Identify important features** using Random Forest to understand the factors affecting yield.
4. **Train a Linear Regression model** to predict soybean seed yield based on selected features.
5. **Evaluate model performance** using standard metrics like RMSE and R^2 score.

Dataset

The dataset used in this study is the **Advanced Soybean Agricultural Dataset**, compiled at Tikrit University in 2025. It consists of **55,450 rows and 13 columns** capturing key agricultural parameters, including:

- **Plant Height (PH)**

- **Number of Pods (NP)**
- **Biological Weight (BW)**
- **Sugars (Su)**
- **Relative Water Content in Leaves (RWCL)**
- **ChlorophyllA663**
- **Chlorophyllb649**
- **Protein Percentage (PPE)**
- **Weight of 300 Seeds (W3S)**
- **Leaf Area Index (LAI)**
- **Seed Yield per Unit Area (SYUA)** (Target Variable)
- **Number of Seeds per Pod (NSP)**
- **Protein Content (PCO)**

	Parameters	Random	Plant Height (PH)	Number of Pods (NP)	Biological Weight (BW)	Sugars (Su)	Relative Water Content in Leaves (RWCL)	ChlorophyllA663	Chlorophyllb649	Protein Percentage (PPE)	Weight of 300 Seeds (W3S)	Leaf Area Index (LAI)	Seed Yield per Unit Area (SYUA)	Number of Seeds per Pod (NSP)	Protein Content (PCO)
0	C1S1G5	R1	50.5	130.3	111.0	0.433	0.732	1.4	3.1	33.2	33.6	0.08	5567.4	1.86	0.82
1	C2S3G4	R1	44.5	132.0	80.0	0.334	0.674	7.1	2.2	38.5	34.4	0.09	2245.5	1.87	0.13
2	C2S1G6	R1	52.2	150.0	83.0	0.490	0.677	1.8	1.3	33.6	35.2	0.07	4326.7	2.04	0.15
3	C1S1G1	R1	50.5	140.8	66.0	0.163	0.745	1.1	3.0	33.5	52.3	0.09	6214.5	2.20	0.14
4	C1S2G6	R2	49.2	175.6	73.0	0.795	0.725	7.5	2.1	39.6	31.2	0.10	3897.8	1.88	0.60

Data Cleaning

The **Parameters** column is removed because it contains encoded categorical values like "C1S1G5", which are not needed for numerical modeling. The **Random** column is also dropped as it contains arbitrary values like "R1, R2, R3" that do not provide meaningful information.

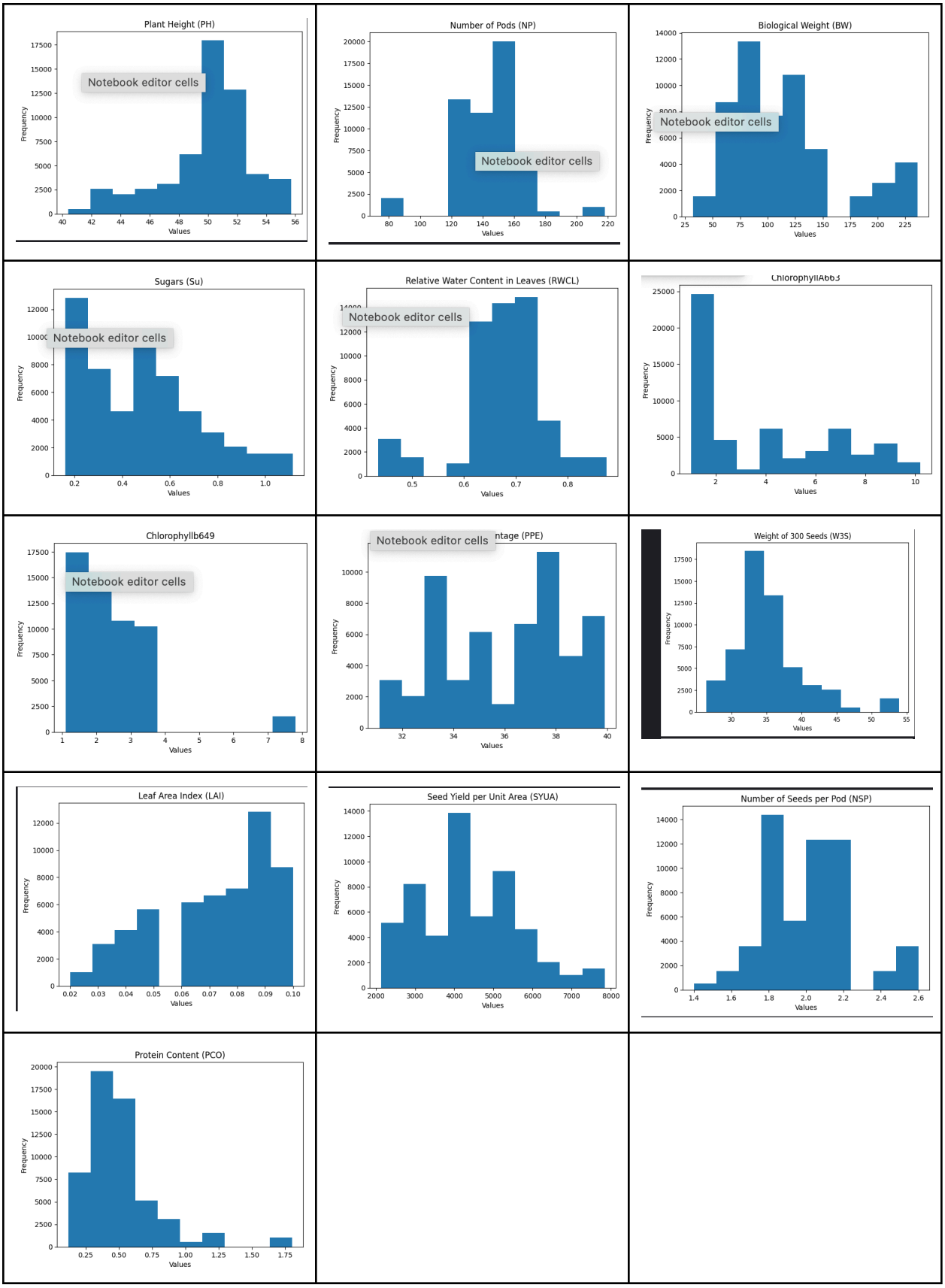
Methodology

1. Data Preprocessing

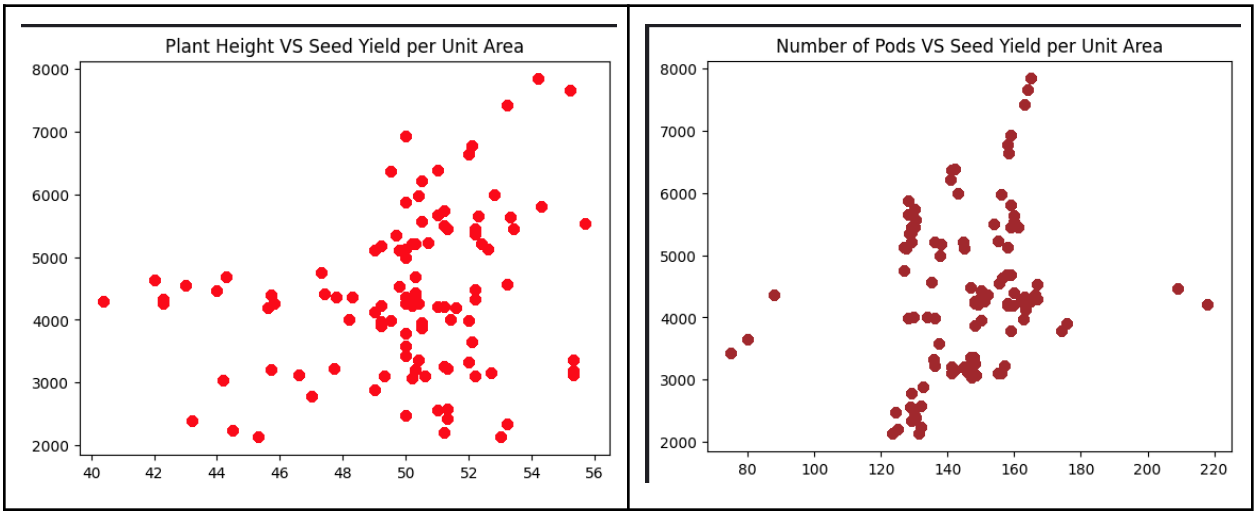
- **Handling Missing Values**: Check for missing values and handle them appropriately.
- **Scaling Data**: Standardize numerical features using **StandardScaler** to improve model performance.
- **Feature Selection**: Use **Random Forest** to identify the most influential features affecting yield.

2. Data Visualization

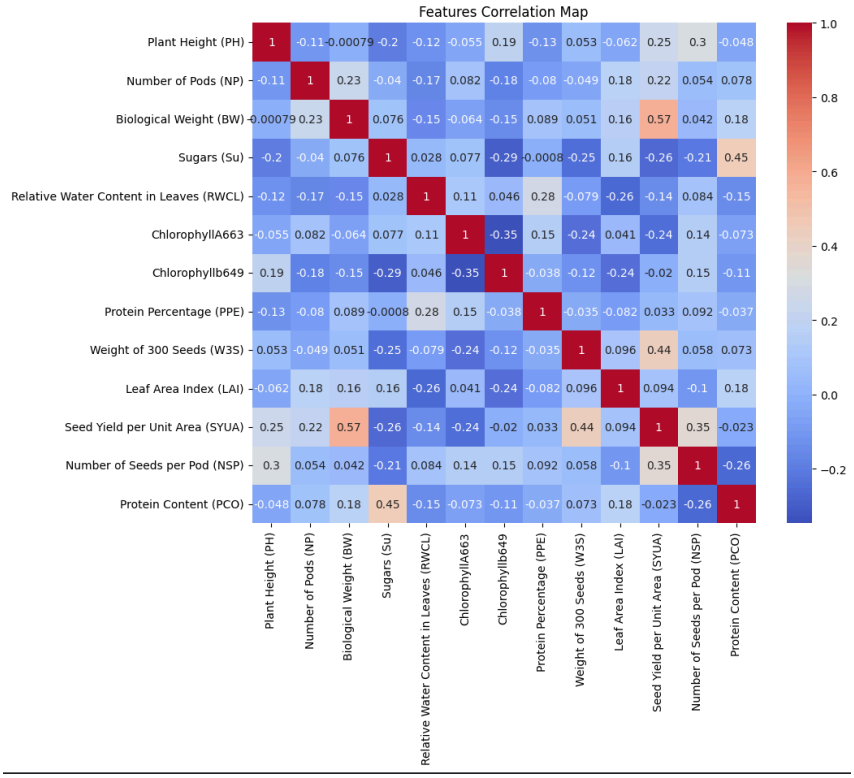
- **Histograms**: Display the distribution of key features.


Plant Height (PH)
Number of Pods (NP)
Biological Weight (BW)
Sugars (Su)
Relative Water Content in Leaves (RWCL)
Chlorophyllb649
Weight of 300 Seeds (W3S)
Leaf Area Index (LAI)
Seed Yield per Unit Area (SYUA)
Number of Seeds per Pod (NSP)
Protein Content (PCO)
Frequency
Values
Notebook editor cells

- **Scatter Plots**: Explore relationships between plant attributes and yield.

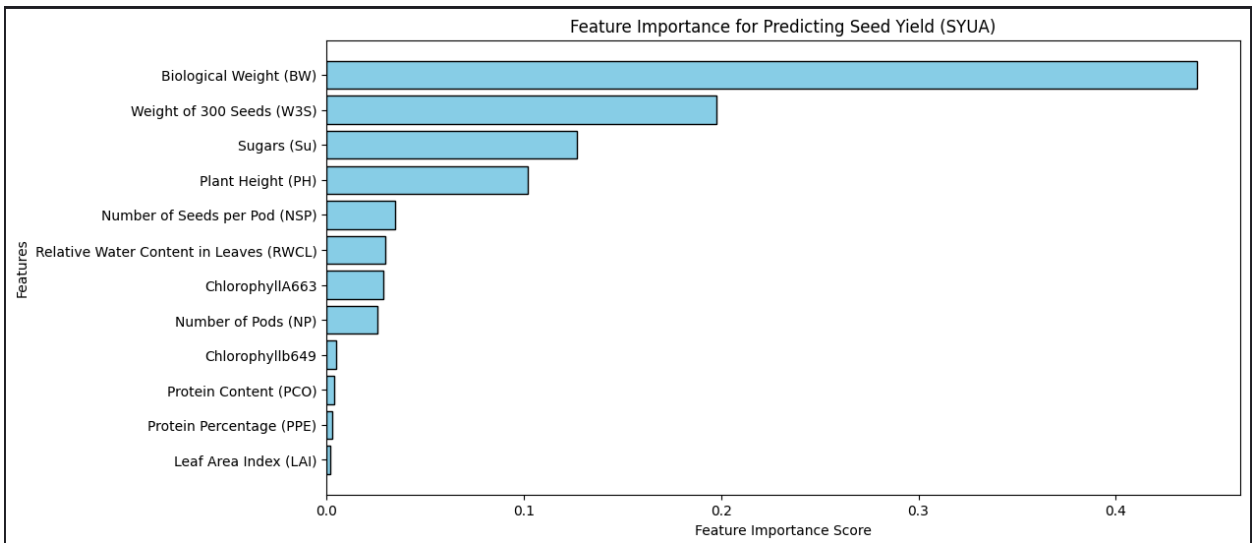


- **Correlation Heatmap**: Identify the most correlated variables with SYUA.



3. Feature Extraction Using Random Forest

Random Forest calculates feature importance scores. This helps determine which factors have the most significant impact on yield, enabling better predictions while reducing noise.



4. Prediction Using Linear Regression

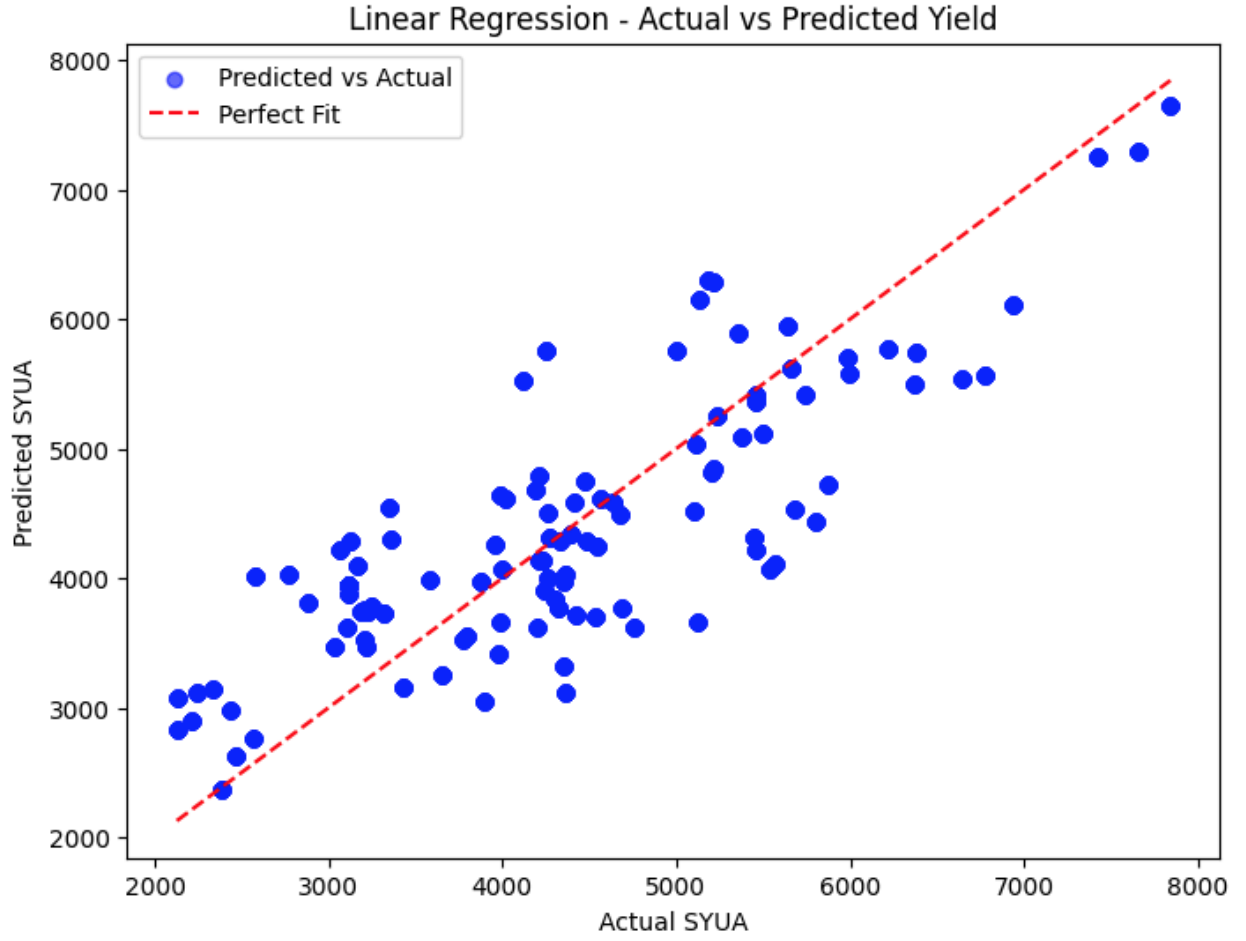
Linear Regression is chosen as it is a simple yet effective model for predicting continuous numerical values like seed yield. The model is trained using the selected important features and evaluated for performance.

5. Evaluation Methods

To assess model accuracy, we use the following metrics:

- **R^2 Score**: Measures how well the independent variables explain the variability in the target variable.
- **Root Mean Squared Error (RMSE)**: Evaluates the difference between actual and predicted values.

```
Mean Absolute Error (MAE): 598.1102
Mean Squared Error (MSE): 532934.0187
R² Score: 0.6571
```



Conclusion

This study demonstrates how machine learning predicts soybean yield efficiently. By selecting the most influential features using Random Forest and applying Linear Regression, we develop a robust predictive model. This approach aids farmers in optimizing crop production and making data-driven agricultural decisions.