Advanced Soybean Yield Prediction

Optimizing Agricultural Practices with Data-Driven Models



Executive Summary of Soybean Yield Prediction Project

Objective: The objective of this project is to develop an accurate model for predicting soybean seed yield based on key agronomic traits such as plant height, biological weight, and protein content. The goal is to provide actionable insights that can optimize farming practices and improve soybean yield predictions.



Importance of the Study

- **Key Features:** Focus on traits like plant height, biological weight, and protein content, all of which are critical in yield prediction.
- Importance: Improving soybean yield is crucial for agricultural productivity. This study will help optimize farming practices and enhance food security by providing precise yield predictions.
- Impact: Farmers can make data driven decisions to improve productivity and sustainability in soybean farming.

Overview of Soybean Dataset & Features

- Dataset Size: 55,450 entries.
- Features: 15 columns representing agronomic characteristics of soybeans, including:
 - > Plant Height (PH)
 - > Number of Pods (NP)
 - > Biological Weight (BW)
 - > Protein Content (PCO)
 - > Seed Yield per Unit Area (SYUA) (target variable).
- Target Variable: Seed Yield per Unit Area (SYUA), representing the total yield per unit area of the farm.

Key Insights from Feature Correlations

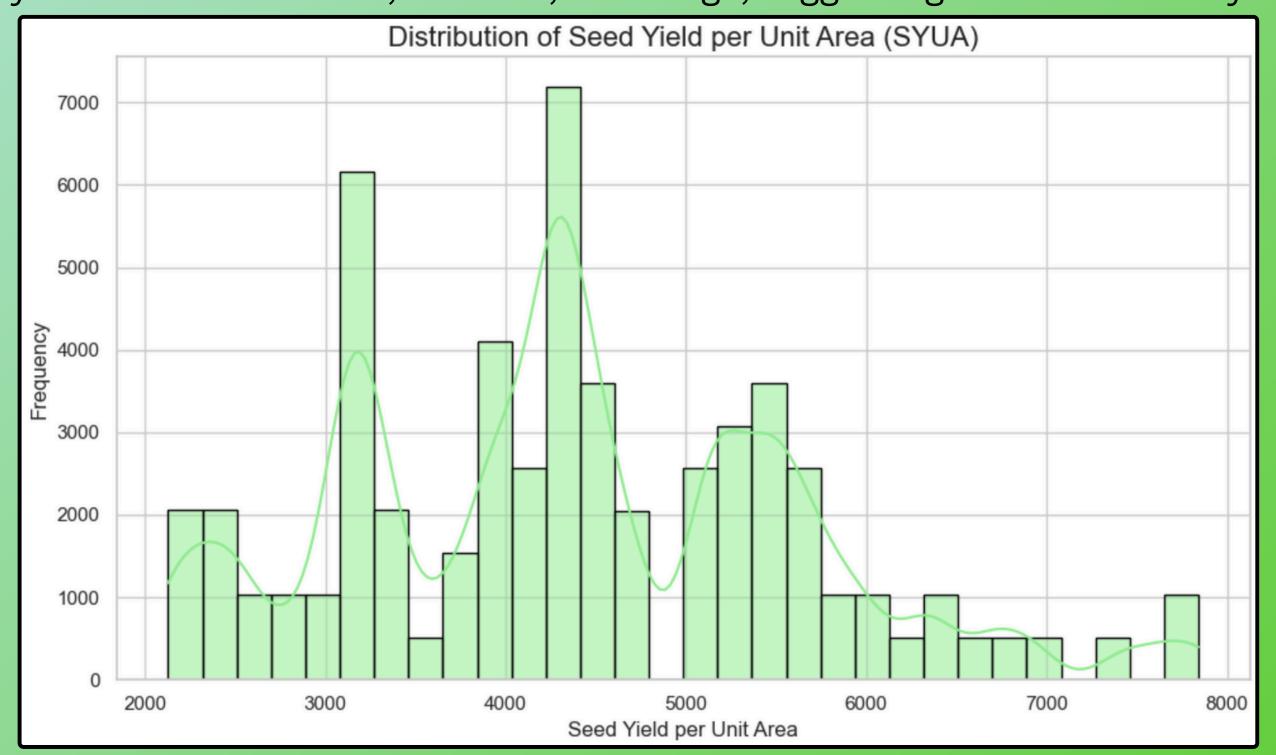
- ➡ Biological Weight (BW) and Number of Seeds per Pod (NSP) have strong correlations with Seed Yield (SYUA), highlighting their importance as predictors.
- Plant Height (PH) and Sugars (Su) show weaker correlations, suggesting they are less significant for predicting yield.

Plant Height (PH)	1.00	-0.11	-0.00	-0.20	-0.12	-0.06	0.19	-0.13	0.05	-0.06	0.25	0.30	-0.05
Number of Pods (NP)	-0.11	1.00	0.23	-0.04	-0.17	0.08	-0.18	-0.08	-0.05	0.18	0.22	0.05	0.08
Biological Weight (BW)	-0.00	0.23	1.00	0.08	-0.15	-0.06	-0.15	0.09	0.05	0.16	0.57	0.04	0.18
Sugars (Su)	-0.20	-0.04	0.08	1.00	0.03	0.08	-0.29	-0.00	-0.25	0.16	-0.26	-0.21	0.45
elative Water Content in Leaves (RWCL)	-0.12	-0.17	-0.15	0.03	1.00	0.11	0.05	0.28	-0.08	-0.26	-0.14	0.08	-0.15
ChlorophyllA663	-0.06	0.08	-0.06	0.08	0.11	1.00	-0.35	0.15	-0.24	0.04	-0.24	0.14	-0.07
Chlorophyllb649	0.19	-0.18	-0.15	-0.29	0.05	-0.35	1.00	-0.04	-0.12	-0.24	-0.02	0.15	-0.11
Protein Percentage (PPE)	-0.13	-0.08	0.09	-0.00	0.28	0.15	-0.04	1.00	-0.04	-0.08	0.03	0.09	-0.04
Weight of 300 Seeds (W3S)	0.05	-0.05	0.05	-0.25	-0.08	-0.24	-0.12	-0.04	1.00	0.10	0.44	0.06	0.07
Leaf Area Index (LAI)	-0.06	0.18	0.16	0.16	-0.26	0.04	-0.24	-0.08	0.10	1.00	0.09	-0.10	0.18
Seed Yield per Unit Area (SYUA)	0.25	0.22	0.57	-0.26	-0.14	-0.24	-0.02	0.03	0.44	0.09	1.00	0.35	-0.02
Number of Seeds per Pod (NSP)	0.30	0.05	0.04	-0.21	0.08	0.14	0.15	0.09	0.06	-0.10	0.35	1.00	-0.26
Protein Content (PCO)	-0.05	0.08	0.18	0.45	-0.15	-0.07	-0.11	-0.04	0.07	0.18	-0.02	-0.26	1.00
	Plant Height (PH)	Number of Pods (NP)	Biological Weight (BW)	Sugars (Su)	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)

Seed Yield Distribution and Key Insights

Seed Yield per Unit Area (SYUA) varies significantly across the dataset, with values ranging from 2,000 to 8,000.

Most seed yields fall within the 3,000 to 5,000 range, suggesting a concentrated yield zone.

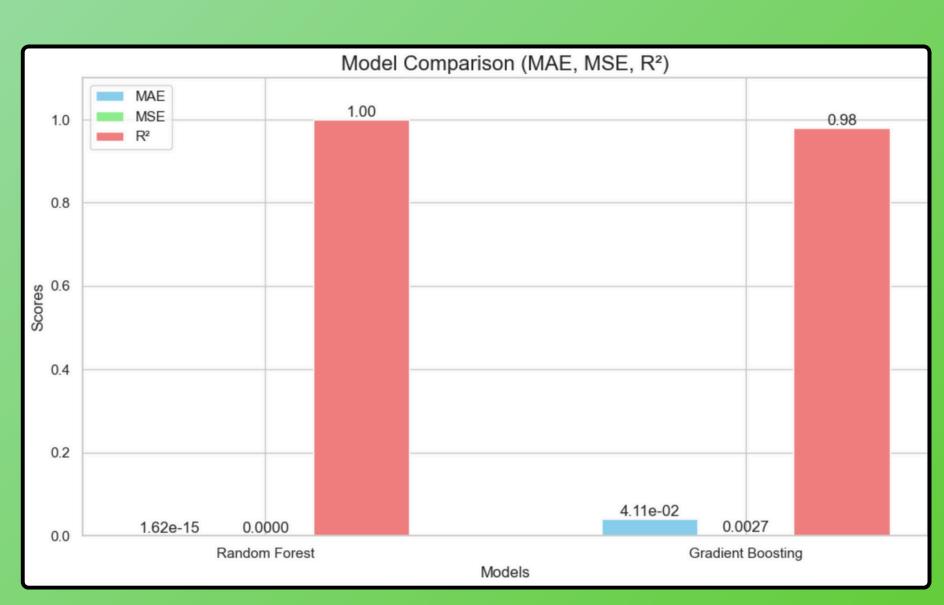


Model Performance Evaluation - Random

Forest vs Gradient Boosting

Key Insights:

- Random Forest shows near-perfect performance:
- > MAE (Mean Absolute Error): 1.62e-15 (indicating almost no error).
- > MSE (Mean Squared Error): 6.66e-30 (effectively negligible).
- > R² (Coefficient of Determination): 1 (perfect prediction).
- Gradient Boosting performs well but with slightly higher error metrics:
- **MAE**: 0.0411.
- **MSE:** 0.0027.
- > R²: 0.98 (indicating very strong predictive accuracy).



Key Findings & Actionable Recommendations

- Actionable Insights for Farmers: Farmers should prioritize Biological Weight and Protein Content.
- ◆ Actionable Insights for Researchers: Researchers should focus on genetic optimization of key traits like Biological Weight and Protein Content.
- Policy Recommendations: Policymakers can encourage research into more productive soybean varieties.

Conclusion & Future Work

Validation with Real-World Data:

- >Objective: Test the model with real farm data to ensure its accuracy and applicability under varying environmental conditions.
- >Action: Collaborate with farms or agricultural organizations to gather real-world yield data for model validation.

Feature Expansion:

- > **Objective:** Enhance prediction accuracy by incorporating additional factors like weather patterns, soil quality, and irrigation.
- >Action: Partner with meteorological and soil health agencies to integrate weather and soil data into the model.

Use of the Model for Other Crops:

- >Objective: Extend the model to other crops like corn, wheat, and rice for broader agricultural applications.
- >Action: Gather data for different crops and adapt the model to account for crop-specific features, expanding its utility.



Let's collaborate to use data and technology to improve crop yields and support sustainable farming.

