Predicting the Yield of Cereal-Legume Intercrops for Plant Density Optimization

Christian Vincent Cabral
cvcabral@gbox.adnu.edu.ph
Ateneo de Naga University
Naga City, PH

Selwyn Guiruela
sguiruela@gbox.adnu.edu.ph
Ateneo de Naga University
Naga City, PH

Reiven Lee
relee@gbox.adnu.edu.ph
Ateneo de Naga University
Naga City, PH

Kyla Ronquillo
kyronquillo@gbox.adnu.edu.ph
Ateneo de Naga University
Naga City, PH

Keywords

Plant density, Yield prediction, Data Analysis

1 Introduction

Global food demand is projected to rise significantly by 2050 due to population growth, rising incomes, and changes in consumption patterns. In an income-driven scenario, world crop calorie production is expected to increase by 47% from 2011 levels, driven primarily by yield improvements (intensification), with a smaller contribution from cropland expansion (intensification)[5].

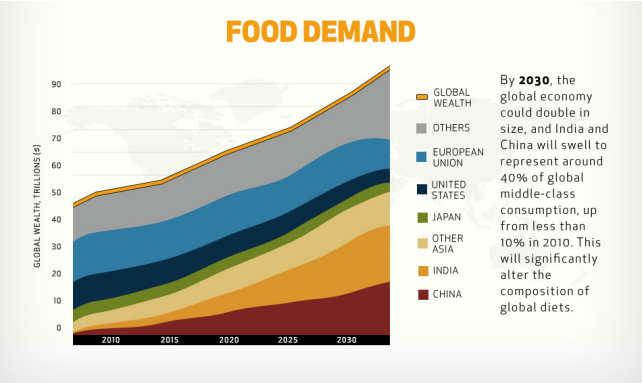

Figure 1: Projected Food Demand and Global Wealth Distribution by 2030.[1]

To meet this rising demand sustainably, agricultural systems must adopt more efficient and effective practices such as intercropping. Intercropping is the practice of planting two or more crops simultaneously in the same area. The goal of this method is to increase overall yield by efficiently using resources that a single crop might not fully utilize. This has been a method for centuries and has been proven to increase efficiency in utilizing resources [3]. One of the common intercropping systems is cereals and legumes.

Cereals (e.g., wheat, maize, barley) are high nitrogen consumers, while legumes (e.g., faba bean, lentil, chickpea) can fix atmospheric nitrogen through their symbiosis with Rhizobium bacteria. This biological nitrogen fixation (BNF) not only reduces the need for synthetic nitrogen fertilizers but also enhances soil fertility. Additionally, legumes can improve micronutrient availability, such as iron and zinc, by releasing phytosiderophores that solubilize poorly available nutrients in the soil [2, 6]. In addition, grain legumes are less preferred in general by farmers, even in organic crop rotations, due to their reputation for low and inconsistent yields. As a result, grain legumes are frequently cultivated in various cereal intercropping systems to enhance yield stability and maintain soil fertility .

However, one of the concurring challenges in cereal-legume intercropping is determining the optimal plant density for each component crop. Plant density is the number of plants per unit area in which it is a key variable for overall efficiency. With a density too high, it may lead to excessive competition for resources while density that is too low may result in poor canopy closure, increasing weed pressure and reducing photosynthetic efficiency and yield. Without predictive tools, farmers will rely on trial-and-error or generalized recommendations, which may not account for differences in species traits and these methods are inefficient[1].

Moreover, the optimal plant density is not static; it varies depending on crop species, environmental factors (e.g. soil type, rainfall, temperature), planting arrangements (row vs. mixed), and management practices (e.g. fertilization). Without predictive tools, farmers will rely on trial-and-error or generalized recommendations, which may not account for differences in species traits and these methods are inefficient[1]. There are two main types of relationships between planting density and crop yield (quantity of harvested crop in a unit area). The first is a progressive relationship, where yield increases with planting density up to a certain point and then levels off, maintaining a stable maximum at higher densities. The second type is a parabolic relationship, in which yield initially increases with density, reaches a peak, and then gradually declines as density continues to rise. Holliday's work served as a foundation for further studies on yield-density interactions[9].

The primary objective of this study is to develop a predictive approach for optimizing plant density in cereal-legume intercropping systems. This is crucial for improving the overall efficiency, yield, and sustainability of such systems, especially under varying environmental and management conditions.

[1]Source: U.S. Department of Agriculture, Economic Research Service (2023). https://doi.org/10.32747/2023.8134356.ers

2 Data Gathering and Preparation

2.1 Data Description

We obtained our data through Zenodo—a free, open-source, multidisciplinary research repository maintained by CERN, OpenAIRE and funded by the European Commission. The direct source of the data can be found through a **global dataset** gathering 37 field experiments involving cereal-legume intercrops and their corresponding sole crops.

2.2 Pre-Processing

The data was filtered to only include **wheat_turgidum_-fababean** intercrop from data_traits.csv to minimize sample for model training. The data of multiple CSV files were then merged to create a unified dataframe. The dataframe was then cleaned by removing/-filling missing data, removing duplicates, reducing noisy data, and checking for inconsistencies in data types and formats.

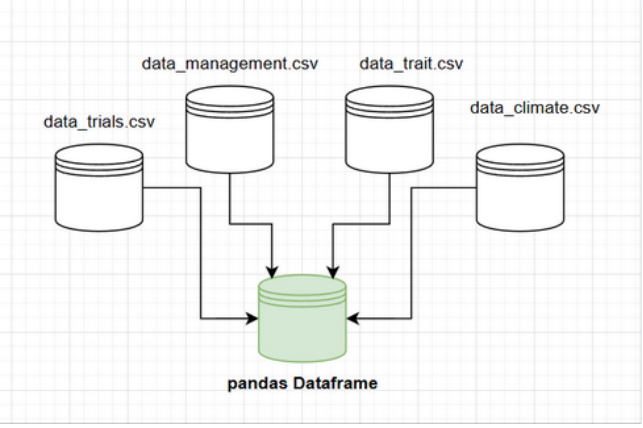

Figure 2: Merging of Multiple CSV files

2.3 Transformation and Reduction

To address the curse of dimensionality in the dataframe, which consists of complex agroecological features with potentially hidden interactions, a dual strategy was employed. First, tree-based models (Gradient Boosted Decision Trees) was used with repeated feature importance evaluation to identify and prioritize the most predictive variables while inherently capturing non-linear relationships. Second, dimensionality reduction techniques, specifically Partial Least Squares (PLS) regression, was applied which is particularly suited for high-dimensional datasets with multicollinearity, as it projects both predictors and responses into a latent space while preserving covariance structure. Where appropriate, Principal Component Analysis (PCA) was also used to further isolate orthogonal sources of variation. This combined approach not only mitigates overfitting risks but also enhances model interpretability by revealing key feature interactions governing plant density outcomes in cereal-legume intercrops.

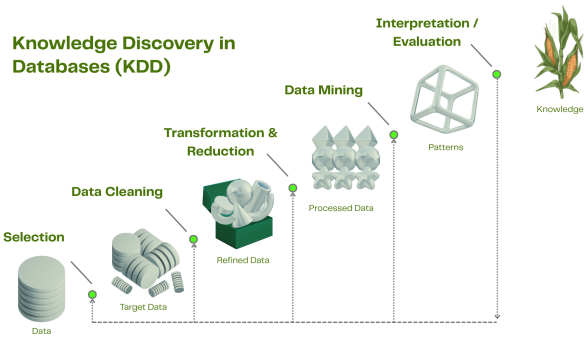

Figure 3: Theoretical Framework based on the Knowledge Discovery in Databases (KDD) Process.

3 Theoretical Framework

The theoretical framework of this study is grounded in the Knowledge Discovery in Databases (KDD) methodology. KDD is a structured and systematic approach that guides the extraction of meaningful patterns and insights from large volumes of data. As shown in Figure 2, the framework is composed of several stages: Selection, Data Cleaning, Transformation and Reduction, Data Mining, and Interpretation/Evaluation.

The process starts with the Selection phase, where important data is collected from sources like databases or live data. Since raw data may have errors or unnecessary information, we will proceed to Data Cleaning. This step fixes errors, fills in missing values, and removes noise, resulting in cleaner data for analysis. After cleaning, the data goes through Transformation and Reduction. This step changes the data into a better format by using techniques like normalization or feature engineering. Then, we have Data Mining, where algorithmic models are used to find patterns, trends, or relationships in the data. Finally, in the Interpretation and Evaluation stage, the patterns are checked to make sure they are useful and accurate. Only patterns that provide meaningful insights are kept and used to support conclusions and decisions in the study.

In summary, the KDD framework helps researchers turn large amounts of data into clear, helpful knowledge. It's especially useful in modern fields like agriculture, AI, and data science, where understanding complex data is essential.

4 Methodology

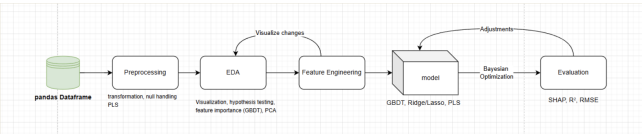

Figure 4: Expanded Form of the Methodology Process.

4.1 Exploratory Data Analysis (EDA)

EDA provides insights into the dataset through visualization, statistical testing, and dimensionality reduction.

4.1.1 Visualization. Visual tools such as histograms, scatter plots, box plots, and heatmaps are used to explore distributions, relationships, and patterns in the data. Pairplots are created to analyze correlations between variables.

4.1.2 Hypothesis Testing. Statistical tests (e.g., t-tests, ANOVA) are conducted to evaluate the relationships and differences between groups. These tests help identify significant variables and validate assumptions.

4.1.3 Feature Importance (GBDT). Gradient Boosted Decision Trees (GBDT) are used to determine feature importance. The algorithm ranks features based on their contribution to reducing error in predictions, guiding feature selection.

4.1.4 Principal Component Analysis (PCA). PCA reduces the dimensionality of the dataset by transforming correlated variables into uncorrelated principal components. This aids in visualization and removes noise, while retaining most of the information.

4.2 Feature Engineering

Feature engineering involves creating new features or modifying existing ones to improve model performance. Domain knowledge and insights from EDA inform the design of these features. Techniques include:

- Standardization of features.
- Interaction terms to capture relationships between variables.
- Polynomial features for capturing non-linear patterns.
- Binning continuous variables into categories.
- Aggregation of temporal or spatial data to create summary features.

4.3 Model Algorithms

The model algorithms used in the study were:

4.3.1 Gradient Boosted Decision Trees (GBDT). GBDT is a robust ensemble learning method that combines multiple weak learners to minimize prediction errors. Its ability to handle non-linear relationships and interactions makes it highly effective.

4.3.2 Ridge/Lasso Regression. Regularization techniques such as Ridge (L2 penalty) and Lasso (L1 penalty) are applied to linear regression models to reduce overfitting and enhance interpretability. Lasso is particularly useful for feature selection due to its ability to shrink coefficients of irrelevant features to zero.

4.3.3 Partial Least Squares (PLS). PLS regression is used as a supplementary model for datasets with multicollinearity or when dimensionality reduction is crucial.

4.4 Model Training and Validation

For the model training and validation, the dataset was split into: training (60%), validation (20%), and testing (20%). The training set and validation set was combined for the final model.

4.5 Tree-structured Parzen Estimator (TPE)

To efficiently tune model hyperparameters, we employ the Tree-structured Parzen Estimator (TPE), a Bayesian optimization technique well-suited for complex and high-dimensional search spaces.

Unlike traditional Gaussian Process-based methods, TPE models the probability density of hyperparameters leading to good and bad performance separately. It constructs two surrogate models: one for the top-performing configurations (exploitation) and another for the rest (exploration). By evaluating the ratio of these densities, TPE selects new hyperparameter values that balance exploration of new areas and exploitation of known good regions.

This approach is particularly effective for optimizing models with mixed types of hyperparameters (continuous, categorical, and discrete) and has demonstrated efficiency in various machine learning tasks.

4.6 Evaluation

Model performance is assessed using quantitative metrics and interpretability tools.

4.6.1 SHapley Additive exPlanations (SHAP). SHAP values provide insights into the contribution of each feature to individual predictions, enhancing model transparency.

4.6.2 R-squared (R^2). R^2 measures the proportion of variance in the target variable explained by the model. A higher R^2 indicates better model performance.

4.6.3 Root Mean Squared Error (RMSE). RMSE quantifies the average magnitude of errors between predicted and actual values. Lower RMSE values signify better model accuracy.

5 Results and Discussion

5.1 Nitrogen Availability by Season

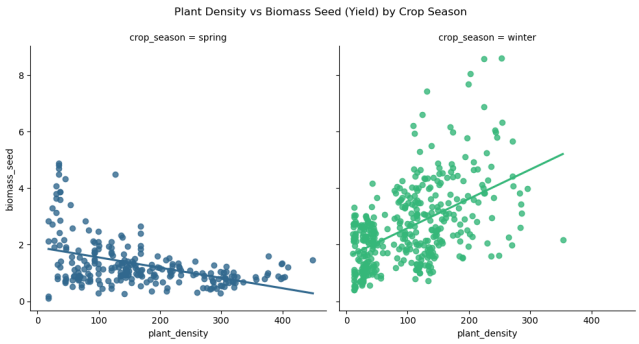

Figure 5: Plant Density Vs Biomass Seed by Crop Season

In Figure 6, we observe that nitrogen availability is significantly lower during the spring season compared to winter. This finding is especially important, as nitrogen is a key nutrient driving biomass accumulation. Plants rely heavily on nitrogen for vegetative growth, particularly during early developmental stages when biomass formation is most responsive to nutrient supply [8].

The limited nitrogen availability in spring provides a compelling explanation for the negative relationship between plant density

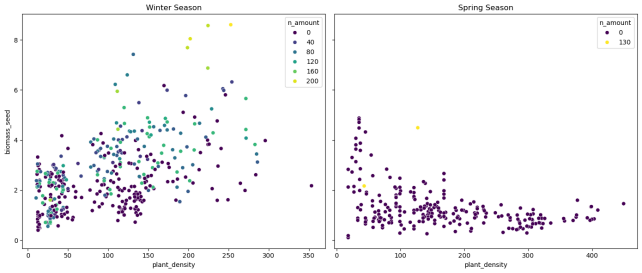

Figure 6: Plant Density Vs Biomass Seed by Nitrogen Amount for Spring and Winter

and biomass seed observed in that season as seen in Figure 5. As plant density increases, competition for nitrogen increases as well.

In contrast, during winter, nitrogen levels are more sufficient, possibly due to slower microbial decomposition and better nitrogen retention in cooler, moister soils [4]. This somehow supports the observed positive correlation between plant density and biomass seed in winter in Figure 5.

5.2 Effect of Planting Accuracy on Yield

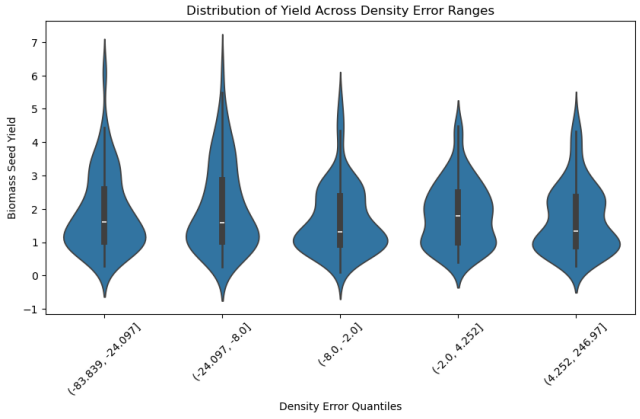

Figure 7: Distribution of Yield Across Density Error Ranges

The violin plot above is divided into 5 groups, in which each of them represent a range of density error, which is the difference between the actual and target plant count. The 4th group from the left shows a density error where the density error is the smallest, meaning it hit the target stand almost exactly. This portrays the highest median seed biomass and the narrowest spread, where most of the trials in this bin cluster tightly around strong yields.

This implies that density error harms yield, that when the closer it is to stick to planned density, the better the yields. Both under-density and overdensity depress biomass seed compared with a properly established stand.

5.3 Yield Response to Nitrogen Application

Figure 8 somehow shows the possibility and likelihood of nitrogen being a major factor in increasing yield. The application of nitrogen

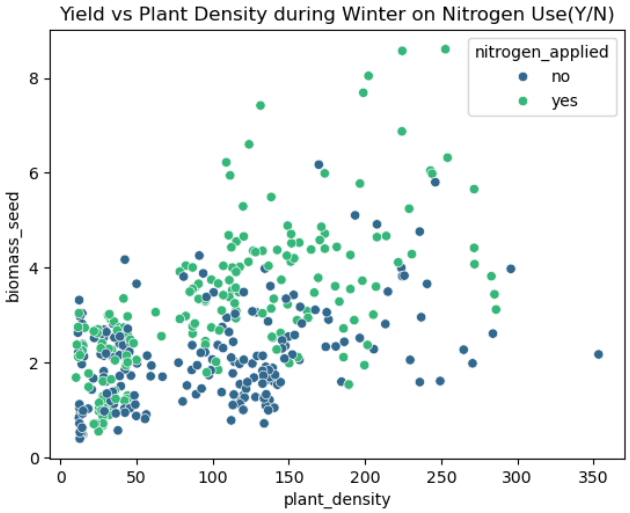

Figure 8: Yield Vs Plant Density during Winter on Nitrogen Use

consistently leads to higher biomass yields at nearly every level of plant density. This is evident from the upward shift in the green points.

5.4 Effect of Stress Index on Yield

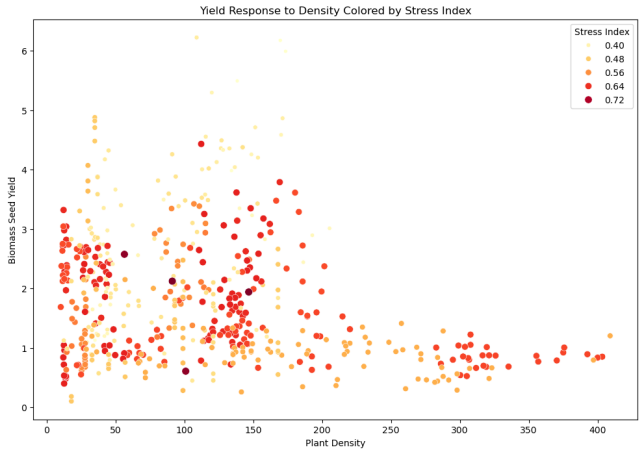

Figure 9: Yield Response to Density by Stress Index

Stress index is a powerful yield modulator. Low-stress plots (pale yellow, 0.40–0.48) produce the highest yields (up to 9 kg/ha) and show a clear positive density–yield trend up to 200–250 plants/unit. High-stress plots (deep red, 0.64–0.72) never exceed 4 kg/ha—and above 150 plants/unit their yields collapse toward <1.5 kg/ha. In addition, density-yield slope flips under stress. Under low stress, yields climb steadily with density. Under high stress, increasing density actually reduces yield—likely because stressed plants can't compete for scarce resources. Intermediate stress levels (orange) sit

between these extremes, with modest density responses (peak 4–5 kg/ha at 150 plants/unit).

5.5 Role of Soil Water Availability

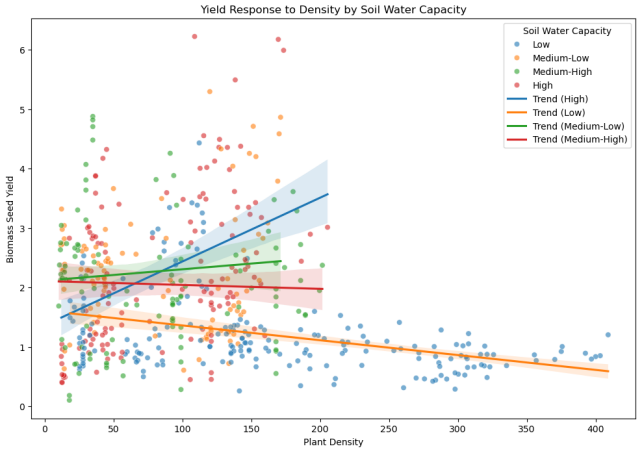

Figure 10: Yield Response to Density by Soil Water Capacity

Based on Figure 10, plant density and biomass seed yield are inversely proportional in terms of soil water. This suggests that lower densities reduce competition, which is beneficial under drought or dry conditions. This study [7] also found the same results in which it stated that, "progressive lowering of crop density increased seed production".

5.6 Distribution Normalization

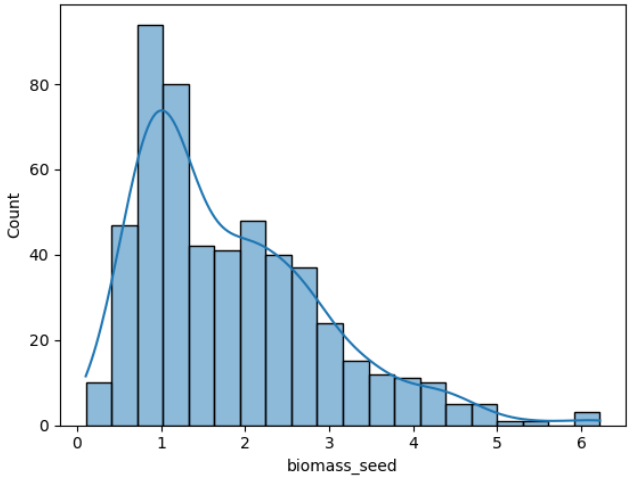

Figure 11: Target Variable Distribution

In Figure 11, the initial distribution of the target variable, biomass seed, exhibited a strong right skew, with most observations concentrated between 0.5 and 2.0. A long tail extending toward higher values suggests outliers that may affect model performance. Hence, a logarithmic transformation was applied to the biomass seed variable to address it. The result is more symmetric and closely approximates a normal distribution, as seen in Figure 12.

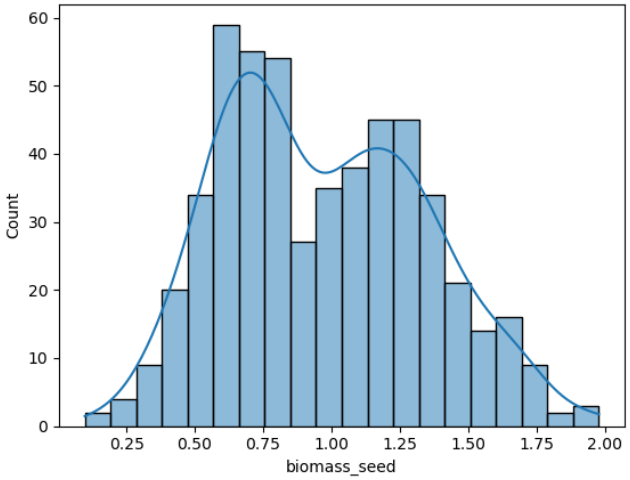

Figure 12: Target Variable Distribution with Log Transformation

5.7 Model Performancce

Following the log transformation of the target variable, to establish a baseline, we trained an XGBoost Regression model using the full set of engineered features, without hyperparameter tuning. To establish a reference point, we trained an initial **XGBoost regression model** using the full set of engineered features, without hyperparameter tuning. The model was evaluated on a held-out test set using the coefficient of determination (R^2) and Root Mean Squared Error (RMSE):

- Full features R^2: 0.763
- Full features RMSE: 0.614

This baseline provided a strong starting point for assessing feature importance and future refinements.

To understand the model's decision-making, we applied **SHAP (SHapley Additive exPlanations)** to the trained XGBoost model. SHAP summary plots reveals us understand which features are driving the model's decisions

This set preserved model relevance while simplifying complexity. We retrained the XGBoost model using only the selected features and evaluated its performance:

- Full features R^2: 0.763
- Top features R^2: 0.753
- Top features RMSE: 0.627

The model that utilized all features achieved an R^2 score of 0.763, indicating a strong fit between the predicted and actual values. When the model was simplified to include only the top features, the R^2 slightly decreased to 0.753, showing only a minimal drop in predictive performance. Additionally, the top-features model had a root mean square error (RMSE) of 0.627, which suggests a relatively low average prediction error. With that, the results demonstrate that

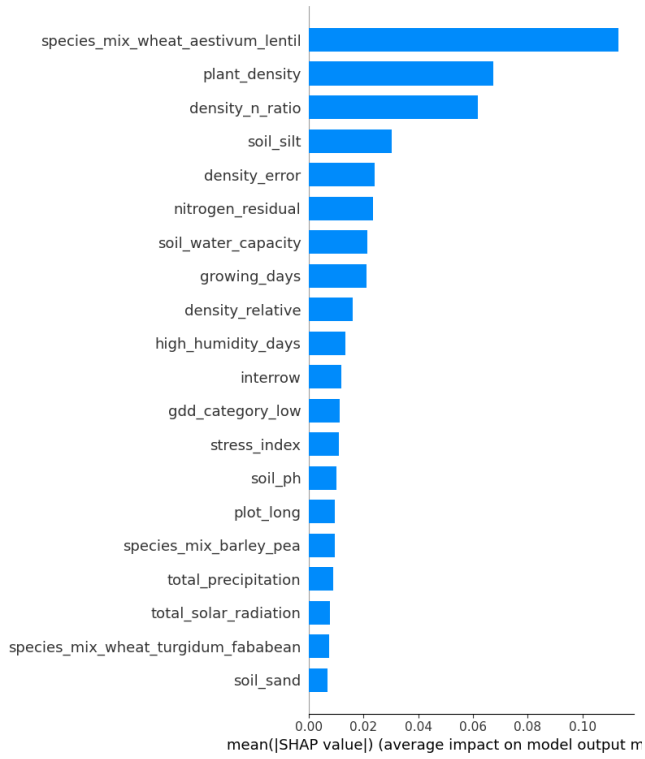

Figure 13: Feature Importance

the top features alone are sufficient to retain most of the model's accuracy.

Meanwhile, the p-value of zero from the baseline comparison suggests that the difference between the full-feature and top-feature models is statistically significant. However, we still choose to proceed with the top-feature model to maintain simplicity and reduce the risk of overfitting. This decision is brought by the negligible difference in R^2.

Before training, we verified that all the selected important features were present in the prepared training dataset. This ensures that the model is trained with the complete set of relevant features.

```
good: bool = True
for feature in training_features:
    if feature not in X_test_prepared_df.columns:
        print(f"{feature} not in X_train_prepared_df")
        good = False

if good:
    print("All features are present in
        X_train_prepared_df")
else:
    print("Not all features are present in
        X_train_prepared_df")
```

Listing 1: Feature presence check

Using the combined dataset of training and testing samples, we performed 5-fold cross-validation to robustly evaluate the model's performance. The model was trained and tested on different splits of the data, allowing a more comprehensive assessment.

```
from sklearn.model_selection import KFold
from xgboost import XGBRegressor
from sklearn.metrics import r2_score
import numpy as np
import pandas as pd

kf = KFold(n_splits=5, shuffle=True, random_state=42)
scores = []

for train_index, test_index in kf.split(combined_df):
    X_train_fold, X_test_fold = combined_df.iloc[
        train_index], combined_df.iloc[test_index]
    y_train_fold, y_test_fold = y_test_combined.iloc[
        train_index], y_test_combined.iloc[test_index]

    model = XGBRegressor(**params)
    model.fit(X_train_fold, y_train_fold)

    y_pred = np.expm1(model.predict(X_test_fold))
    y_true = np.expm1(y_test_fold)

    score = r2_score(y_true, y_pred)
    scores.append(score)

print(f"Mean R  : {np.mean(scores):.4f}")
print(f"Standard Deviation: {np.std(scores):.4f}")
```

Listing 2: K-Fold Cross-Validation with XGBoost

The average coefficient of determination across the cross-validation folds was found to be $\mathbf{R^2 = 0.6793}$, indicating that the model explains approximately 68% of the variance in the target variable, which reflects strong predictive performance. The standard deviation of the R^2 scores was SD = **0.1479**, suggesting moderate variability in model performance across different data splits; however, the overall results demonstrate reasonable consistency.

5.8 SHAP Analysis

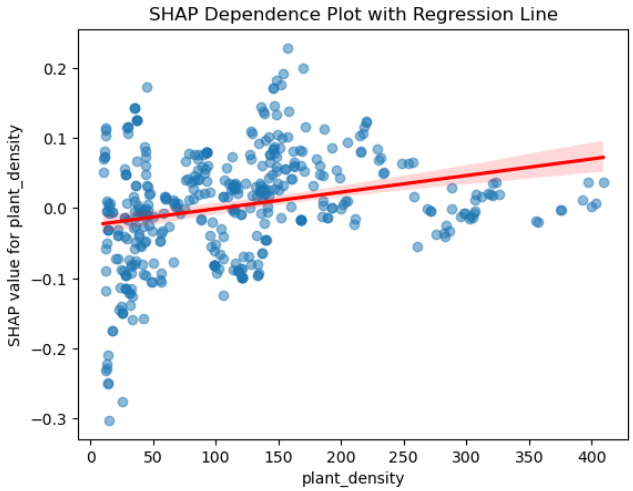

Figure 14: SHAP Dependence Plot with Regression Line

It starts low then as `plant_density` increases, chances are yield will increase but reaches a certain cap.

Diminishing Returns Beyond 200 Plants/m^2:

- From ~200 to 400 plants/m^2, SHAP values converge near zero or remain slightly positive.
- This implies increasing plant density beyond this point yields little added benefit or influence on the model's target prediction.

Initially (0–100 plants/m^2), the impact on the model prediction is both positive and negative, suggesting sensitivity to context (e.g., weather, soil).

Here are other interactions that depend on various factors:

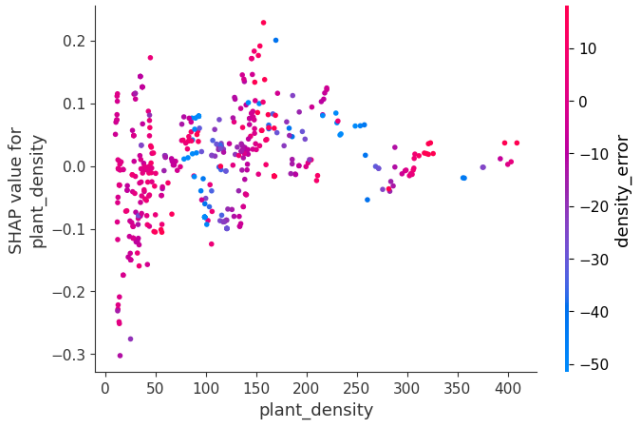

Figure 15: SHAP Dependence Plot for Plant Density with Density Error Coloring

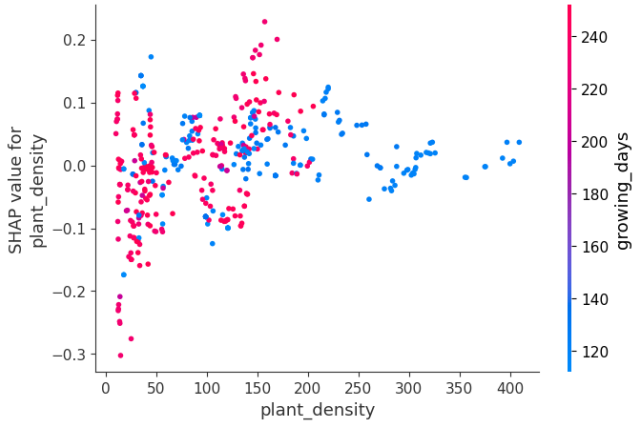

Figure 16: SHAP Dependence Plot for Plant Density with Growing Days

6 Conclusion and Recommendation

Our results highlight the central role of soil nitrogen dynamics in driving seasonal and management-induced variation in seed biomass. In spring, when nitrogen availability is significantly lower than in winter, plants compete more intensely for this critical nutrient, leading to a negative relationship between plant density and biomass accumulation. By contrast, winter's cooler, moister soils retain nitrogen more effectively, supporting a positive density–yield relationship under lower nutrient stress.

Likewise, nitrogen application was shown to shift the entire density–yield curve upward at all densities, reaffirming nitrogen's role as a primary driver of biomass accumulation. However, the interaction with environmental stress reveals that under low-stress conditions, yields increase steadily with density, whereas under

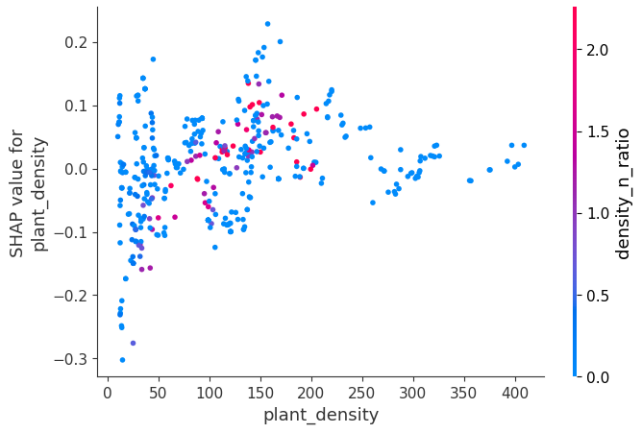

Figure 17: SHAP Dependence Plot for Plant Density with Density-Nitrogen Ratio

high stress, increasing density increases competition for resources and therefore reduces yield.

Water relations further modulate these outcomes: when soil moisture is limiting, lower crop densities favor seed production, suggesting that density management can partially compensate for water stress. From a data-analysis perspective, applying a logarithmic transformation to address the right skew in biomass distribution yielded a more symmetric, approximately normal target variable, improving the reliability of our modeling efforts. An XGBoost regression model trained on the full set of engineered features achieved a strong fit ($R^2 = 0.763$, RMSE = 0.614), and when retrained using only the top features identified via SHAP, it retained nearly identical predictive power ($R^2 = 0.753$, RMSE = 0.627).

These findings suggest that optimizing seed biomass requires a multifaceted approach: season-specific nitrogen management, rigorous control of planting density, and stress mitigation strategies are all critical. Future research should explore dynamic fertilization routines tailored to seasonal nutrient fluxes, develop finer-scale stress index mapping, and integrate remote sensing data to further refine and validate the predictive model.

In light of these observations, we can conclude that while plant density contributes to yield, there is a certain threshold where it does not get any better. This suggests that there is a global optimum, and maximizing this improves the chances of having more yield. However, based on the interactions between the current features, plant density optimization requires a much complex configuration. This study attempted to capture the complex patterns and provide insights for optimization, but with the lack of much-needed data, the findings are less reliable. As such, we recommend future researchers with a similar approach to gather a significant amount of data to reveal conclusive pieces of evidence.

References

[1] A. Attallah, W. Hamdi, A. Souid, M. Farissi, B. L'taief, A. J. Messiga, N. Y. Rebouh, S. Jellali, and M. F. Zagrarni. 2024. Impact of Cereal–Legume Intercropping on Changes in Soil Nutrients Contents under Semi–Arid Conditions. *Sustainability* 16, 7 (2024), 2725. https://doi.org/10.3390/su16072725

[2] A. Ebbisa. 2022. Mechanisms Underlying Cereal/Legume Intercropping as Nature-Based Biofortification: A Review. *Food Production, Processing and Nutrition* 4 (2022), 19. https://doi.org/10.1186/s43014-022-00096-y

[3] Francis Ofori and W.R. Stern. 1987. Cereal–Legume Intercropping Systems. In *Advances in Agronomy*, N.C. Brady (Ed.). Vol. 41. Academic Press, 41–90. https://doi.org/10.1016/S0065-2113(08)60802-0

[4] Clare Robinson. 2002. Controls on decomposition and soil nitrogen availability at high latitudes. *Plant and Soil* 242 (05 2002). https://doi.org/10.1023/A:1019681606112

[5] Ronald Sands, Birgit Meade, Jr. Seale, James L., Sherman Robinson, and Rachel Seeger. 2023. *Scenarios of Global Food Consumption: Implications for Agriculture.* Economic Research Report ERR-323. U.S. Department of Agriculture, Economic Research Service. https://doi.org/10.32747/2023.8134356.ers

[6] Demie Dereje T., Döring Thomas F., Finckh Maria R., van der Werf Wopke, Enjalbert Jérôme, and Seidel Sabine J. 2022. Mixture × Genotype Effects in Cereal/Legume Intercropping. *Frontiers in Plant Science* 13 (2022). https://doi.org/10.3389/fpls.2022.846720

[7] B. J. WILSON, K. J. WRIGHT, P. BRAIN, M. CLEMENTS, and E. STEPHENS. 1995. Predicting the competitive effects of weed and crop density on weed biomass, weed seed production and crop yield in wheat. *Weed Research* 35, 4 (1995), 265–278. https://doi.org/10.1111/j.1365-3180.1995.tb01789.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-3180.1995.tb01789.x

[8] Jie Wu, Ying Song, Guang-Yu Wan, Liang-Qi Sun, Jing-Xian Wang, Zi-Sheng Zhang, and Cheng-Bin Xiang. 2025. Boosting crop yield and nitrogen use efficiency: the hidden power of nitrogen-iron balance. *New Crops* 2 (2025), 100047. https://doi.org/10.1016/j.ncrops.2024.100047

[9] Rui Zhu, Zhen Zhang, Yujie Cao, Zhiyuan Hu, Yujie Li, Hao Cao, Zhiyong Zhao, Dongxue Xin, and Qijun Chen. 2023. CPDOS: A Web-Based AI Platform to Optimize Crop Planting Density. *Agronomy* 13, 10 (2023), 2465. https://doi.org/10.3390/agronomy13102465