sustainability

MDPI

Article

Machine Learning Approaches for Predicting Maize Biomass Yield: Leveraging Feature Engineering and Comprehensive Data Integration

Maryam Abbasi [1], Paulo Váz [2], José Silva [2] and Pedro Martins [2,*]

[1] Applied Research Institute, Polytechnic of Coimbra, 3000 Coimbra, Portugal; maryam.abbasi@ipc.pt
[2] Research Center in Digital Services, Polytechnic of Viseu, 3500 Viseu, Portugal; paulovaz@estgv.ipv.pt (P.V.); jsilva@estgv.ipv.pt (J.S.)
* Correspondence: pedromom@estgv.ipv.pt

Abstract: The efficient prediction of corn biomass yield is critical for optimizing crop production and addressing global challenges in sustainable agriculture and renewable energy. This study employs advanced machine learning techniques, including Gradient Boosting Machines (GBMs), Random Forests (RFs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), integrated with comprehensive environmental, soil, and crop management data from key agricultural regions in the United States. A novel framework combines feature engineering, such as the creation of a Soil Fertility Index (SFI) and Growing Degree Days (GDDs), and the incorporation of interaction terms to address complex non-linear relationships between input variables and biomass yield. We conduct extensive sensitivity analysis and employ SHAP (SHapley Additive exPlanations) values to enhance model interpretability, identifying SFI, GDDs, and cumulative rainfall as the most influential features driving yield outcomes. Our findings highlight significant synergies among these variables, emphasizing their critical role in rural environmental governance and precision agriculture. Furthermore, an ensemble approach combining GBMs, RFs, and ANNs outperformed individual models, achieving an RMSE of 0.80 t/ha and R^2 of 0.89. These results underscore the potential of hybrid modeling for real-world applications in sustainable farming practices. Addressing the concerns of passive farmer participation, we propose targeted incentives, education, and institutional support mechanisms to enhance stakeholder collaboration in rural environmental governance. While the models assume rational decision-making, the inclusion of cultural and political factors warrants further investigation to improve the robustness of the framework. Additionally, a map of the study region and improved visualizations of feature importance enhance the clarity and relevance of our findings. This research contributes to the growing body of knowledge on predictive modeling in agriculture, combining theoretical rigor with practical insights to support policymakers and stakeholders in optimizing resource use and addressing environmental challenges. By improving the interpretability and applicability of machine learning models, this study provides actionable strategies for enhancing crop yield predictions and advancing rural environmental governance.

Keywords: biomass yield prediction; machine learning; gradient boosting machines; random forest; support vector machines; soil fertility index; feature engineering; sensitivity analysis; sustainable agriculture; renewable energy; crop management

Academic Editor: Bin Ji

Received: 18 October 2024
Revised: 21 December 2024
Accepted: 23 December 2024
Published: 2 January 2025

Citation: Abbasi, M.; Váz, P.; Silva, J.; Martins, P. Machine Learning Approaches for Predicting Maize Biomass Yield: Leveraging Feature Engineering and Comprehensive Data Integration. *Sustainability* **2025**, *17*, 256. https://doi.org/10.3390/su17010256

1. Introduction

The global demand for renewable energy and sustainable agricultural practices is growing as the effects of climate change, population growth, and resource constraints become more pressing. Biomass, as a renewable energy source, offers a significant opportunity to address these challenges by providing an alternative to fossil fuels and supporting energy security while contributing to environmental sustainability. Among biomass crops, corn plays a critical role due to its versatility for food, feed, and bioenergy applications. The accurate prediction of corn biomass yield is essential for optimizing crop management practices and maximizing the potential of biomass-based renewable energy.

Traditional models for predicting biomass yield, including empirical and mechanistic approaches, are often limited in their ability to capture the complex and nonlinear relationships between environmental factors, soil conditions, and crop management practices [1]. These models rely on oversimplified assumptions and are constrained in their ability to process large, multi-dimensional datasets that characterize modern agricultural systems. As a result, they often fail to account for the intricate interplay between factors such as soil fertility, precipitation, temperature, and planting density, which are critical drivers of biomass yield.

In contrast, machine learning (ML) techniques offer a more flexible and powerful approach, capable of uncovering hidden patterns and interactions within large datasets [2]. Recent advances in ML have demonstrated significant potential in agricultural applications, including crop yield prediction, pest detection, and resource optimization. Machine learning models such as Random Forests (RFs), Gradient Boosting Machines (GBMs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs) have been shown to outperform traditional models by capturing complex, nonlinear relationships among diverse variables [3,4]. For instance, studies have highlighted the effectiveness of these models in improving the accuracy of yield predictions while integrating diverse data sources, such as soil properties, climate conditions, and crop management practices.

Despite these advancements, several gaps remain in the application of machine learning to biomass yield prediction. Most studies focus primarily on grain yield rather than total biomass, limiting their applicability for bioenergy production. Furthermore, the interpretability of ML models, which is essential for providing actionable insights for farmers and policymakers, is often overlooked [5]. This lack of interpretability hinders the practical adoption of these models, as stakeholders need to understand the underlying drivers of yield predictions to make informed decisions. Addressing these challenges requires not only the development of accurate models but also the integration of domain-specific knowledge, such as soil fertility indices, crop growth patterns, and climate variability.

This study aims to bridge this gap by leveraging advanced machine learning techniques to predict corn biomass yield with high accuracy and interpretability. Our approach integrates environmental, soil, and crop management data spanning a decade from key corn-growing regions in the United States. To address the complexity of yield prediction, we employ advanced feature engineering methods, including the creation of composite indices such as the Soil Fertility Index (SFI) and Growing Degree Days (GDDs), which capture the interactions between critical variables. Furthermore, sensitivity analysis and feature importance evaluation using SHAP (SHapley Additive exPlanations) values provide interpretable insights into the factors driving biomass yield, facilitating their integration into practical agricultural management strategies.

To ensure practical relevance, we also explore the challenges of farmer participation in rural environmental governance. Farmers often play a passive role in sustainable agricultural practices due to a lack of education, resources, or incentives. By incorporating actionable recommendations—such as targeted incentives, farmer education, and institu-

tional support—this study not only addresses the technical challenges of biomass yield prediction but also provides a framework for enhancing stakeholder collaboration and environmental governance.

The findings of this study demonstrate the potential of machine learning to advance predictive modeling in agriculture while contributing to broader goals of sustainability and renewable energy. By addressing key gaps in model accuracy, interpretability, and practical application, this research offers a robust framework for integrating machine learning into agricultural decision-making processes. This approach supports policymakers, agronomists, and farmers in optimizing crop production, reducing resource waste, and mitigating the impacts of climate variability.

2. Literature Review

Accurately predicting crop yields, particularly biomass yield, is essential for sustainable agricultural practices and food security. The growing demand for bioenergy and sustainable agriculture necessitates the development of robust models capable of capturing the complex, nonlinear relationships between environmental factors, soil properties, and crop management practices [6,7]. However, the existing literature highlights the challenges posed by the dynamic and multifaceted nature of agricultural systems, which often require more advanced modeling approaches than traditional techniques can offer.

2.1. Traditional Approaches to Biomass Yield Prediction

Early research on yield prediction primarily focused on empirical and process-based models. Empirical models typically rely on simple regression techniques to link environmental factors such as precipitation and temperature to crop yields [8]. While these models are straightforward and interpretable, they often oversimplify complex systems, failing to capture key interactions between variables. For instance, Vallejo et al. [7] utilized radiation-use efficiency models to estimate corn biomass but acknowledged their limited applicability across diverse agro-ecological zones.

Process-based models improved upon empirical methods by incorporating physiological and environmental processes into simulations of crop growth [1]. Although these models provide a detailed understanding of plant growth dynamics, they are constrained by high computational requirements, limited scalability, and difficulties in integrating large, heterogeneous datasets. Furthermore, they often rely on assumptions that do not reflect real-world variability, such as uniform soil conditions or optimal weather patterns [9].

While these traditional approaches have contributed to our understanding of crop yield dynamics, their limitations underscore the need for more flexible and data-driven techniques capable of handling the complexities of biomass yield prediction. This gap has led to the growing adoption of machine learning methods, which are better suited to accommodate the intricate interactions inherent in agricultural systems.

2.2. Advancements in Machine Learning for Yield Prediction

Machine learning (ML) techniques have emerged as powerful tools for yield prediction, offering flexibility and accuracy by uncovering hidden patterns in large, multi-dimensional datasets [2]. These methods, including Random Forests (RFs), Gradient Boosting Machines (GBMs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), have shown superior performance in agricultural applications compared to traditional models [1]. For instance, Paudel et al. [3] demonstrated the ability of RF and ANN models to effectively capture complex interactions between soil properties, climate variables, and crop management practices, significantly improving prediction accuracy for corn yields.

Recent studies have further highlighted the role of advanced feature engineering in enhancing ML performance. Composite indices, such as the Soil Fertility Index (SFI) and Growing Degree Days (GDDs), have been successfully incorporated into ML models to quantify critical agricultural parameters and improve model interpretability [10]. These indices allow models to capture the effects of variables that traditional methods often overlook, such as soil health and heat accumulation over the growing season.

Despite these advancements, ML models face challenges in interpretability. The so-called "black-box" nature of many ML techniques makes it difficult for stakeholders, such as farmers and policymakers, to understand the underlying drivers of yield predictions [5]. To address this, methods such as SHAP (SHapley Additive exPlanations) have been developed to quantify feature importance and provide actionable insights [11]. These methods bridge the gap between model accuracy and practical applicability by making ML predictions more transparent and user-friendly.

2.3. Integration of Remote Sensing and GISs in Biomass Estimation

The integration of remote sensing and Geographic Information Systems (GISs) into yield prediction models has revolutionized biomass estimation by enabling large-scale, real-time assessments. Remote sensing data, such as satellite imagery, provide spatially continuous estimates of crop health and environmental conditions, which can enhance the accuracy and scalability of yield predictions [12]. For instance, Kharel et al. [13] demonstrated the effectiveness of combining remote sensing data with machine learning models to estimate the biomass of mixed-species crops, achieving improved accuracy compared to traditional field-based methods.

However, challenges remain in integrating remote sensing data with ML models, including issues related to varying spatial and temporal resolutions, data preprocessing, and the need for ground-truth validation. Advances in data fusion techniques and multi-source modeling are critical for addressing these challenges and fully leveraging the potential of remote sensing in agricultural applications [14].

2.4. Research Gaps and Study Motivation

While significant progress has been made in applying ML techniques to yield prediction, several gaps remain, particularly in the context of biomass crops such as corn. Most studies focus on grain yield rather than total biomass, which limits their relevance for bioenergy applications. Furthermore, there is a lack of research exploring the integration of diverse data sources, such as soil indices, remote sensing data, and crop management practices, into ML models. This limits the ability of existing models to fully capture the complex interactions driving biomass yield.

Another critical gap lies in the interpretability of ML models. While advanced algorithms such as GBMs and ANNs achieve high accuracy, their practical adoption is hindered by their "black-box" nature. Addressing this issue requires the incorporation of explainability techniques, such as SHAP values, to provide actionable insights for decision-making. Additionally, most existing studies neglect the role of social and institutional factors, such as farmer participation and education, in influencing agricultural outcomes. A more holistic approach that integrates technical, social, and policy perspectives is needed to address these gaps.

This study seeks to address these challenges by developing and evaluating machine learning models that integrate environmental, soil, and crop management data to predict corn biomass yield. Our approach emphasizes advanced feature engineering, interpretability, and practical applicability, providing a comprehensive framework for optimizing agricultural practices and supporting renewable energy initiatives. By combining cutting-

edge ML techniques with domain-specific knowledge, this study contributes to the growing body of research on sustainable agriculture and bioenergy production.

3. Methodology

This section outlines the methodology used to predict corn biomass yield. It describes the data collection process, preprocessing steps, feature engineering, machine learning model development, and evaluation criteria. The approach integrates traditional and advanced machine learning techniques, emphasizing model interpretability and practical applicability to address the challenges of biomass yield prediction.

3.1. Data Collection and Description

The dataset used in this study was compiled from several publicly available sources, including the United States Department of Agriculture (USDA, https://www.nass.usda.gov, accessed on 22 December 2024), USDA Soil Geography. (https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/survey/geo/, accessed on 22 December 2024), and the National Oceanic and Atmospheric Administration (NOAA, https://www.ncdc.noaa.gov/cdo-web/, accessed on 22 December 2024). The dataset encompasses key corn-growing regions in the Midwest United States, specifically Illinois, Iowa, and Nebraska, over a ten-year period. This comprehensive dataset includes approximately 10,000 observations, offering a rich context for analyzing factors affecting biomass yield.

The dataset contains the following key variables:

- Environmental variables: Daily maximum and minimum temperature (°C), cumulative precipitation (mm), average solar radiation (W/m^2), and wind speed (m/s).
- Soil properties: pH, nitrogen (mg/kg), phosphorus (mg/kg), potassium (mg/kg), organic matter (%), and soil moisture (%).
- Crop management practices: Planting density (plants/ha), irrigation (mm), and fertilizer application (kg/ha).
- Target variable: Biomass yield (t/ha), used as the output for prediction.

To ensure the integrity of the analysis, data quality checks were conducted. Missing values were handled using imputation techniques, and outliers were addressed through systematic methods, as described in the following subsection. This ensured a reliable and consistent dataset for model development.

3.2. Data Preprocessing

Data preprocessing was a critical step to ensure the quality and usability of the dataset. The following procedures were applied:

1. Handling missing data: Approximately 2.3% of the dataset had missing values. For numerical features with fewer than 5% missing values, mean imputation was applied. Features with higher percentages of missing values were handled using Multiple Imputation by Chained Equations (MICE). Categorical features were imputed with their mode.
2. Outlier detection and treatment: Outliers were identified using the Interquartile Range (IQR) method. Moderate outliers were capped using winsorization, while extreme outliers were removed to prevent skewed model predictions. This approach balanced the need for robust models with the retention of meaningful variability in the data.
3. Normalization and scaling: To ensure consistency across features, numerical variables were standardized using z-score normalization (Equation (1)):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original value, μ is the mean, and σ is the standard deviation. This step ensures that all features contribute equally to the model's predictions, avoiding biases due to differing units or scales.

3.3. Feature Engineering

Feature engineering was performed to enhance the predictive power of the models. In addition to the original variables, new features were created based on domain knowledge of agriculture and biomass production:

- Soil Fertility Index (SFI): This composite index combines multiple soil health indicators, such as pH, nitrogen, phosphorus, potassium, and organic matter, to quantify soil fertility (Equation (2)):

$$\text{SFI} = w_1 \cdot \text{pH} + w_2 \cdot \text{Nitrogen} + w_3 \cdot \text{Phosphorus} + w_4 \cdot \text{Potassium} + w_5 \cdot \text{Organic Matter} \tag{2}$$

The weights (w_1 to w_5) were determined using correlation analysis to reflect the relative importance of each soil property.

- Growing Degree Days (GDDs): A feature representing heat accumulation throughout the growing season, which is critical for corn growth (Equation (3)):

$$GDDs = \sum_{i=1}^{n} \max\left(\frac{T_{\max,i} + T_{\min,i}}{2} - T_{\text{base}}, 0\right) \tag{3}$$

where T_{base} represents the base temperature for corn growth, set at 10 °C. This feature captures the thermal energy required for crop development, making it a key variable for yield prediction.

- Cumulative Rainfall (CR): Total precipitation over the growing season was calculated to reflect water availability (Equation (4)):

$$CR = \sum_{i=1}^{n} \text{Precipitation}_i \tag{4}$$

- Water Stress Index (WSI): This index quantifies periods of water deficit, incorporating precipitation and irrigation data (Equation (5)):

$$WSI = \sum_{i=1}^{n} \max(0, PET_i - (\text{Precipitation}_i + \text{Irrigation}_i)) \tag{5}$$

where PET represents potential evapotranspiration. The WSI captures critical periods of water stress, which can significantly impact biomass yield.

These engineered features, combined with the original dataset, improved the models' ability to capture interactions and nonlinear effects, addressing gaps in traditional approaches. Interaction terms and polynomial features were also considered where exploratory analysis indicated potential predictive value.

3.4. Model Development

To predict corn biomass yield, a series of machine learning models were developed. These included traditional approaches and advanced techniques:

1. Linear Regression: Used as a baseline model for comparison due to its simplicity and interpretability.
2. Random Forests (RFs): An ensemble model that builds multiple decision trees and averages their predictions to reduce variance and improve accuracy.
3. Gradient Boosting Machines (GBMs): A boosting model that builds decision trees sequentially, with each tree correcting the errors of the previous one. This method is particularly effective at capturing complex, nonlinear relationships.
4. Support Vector Machines (SVMs): A model adapted for regression tasks, focusing on finding the hyperplane that minimizes error.
5. Artificial Neural Networks (ANNs): A deep learning model capable of capturing intricate patterns in the data. Dropout layers were included to reduce overfitting.
6. Ensemble Model: A hybrid approach that combines predictions from RFs, GBMs, and ANNs using a weighted averaging strategy to improve accuracy.

The hyperparameters of each model were tuned using a grid search approach with 5-fold cross-validation on the training set. This ensured optimal performance by systematically testing combinations of parameters, such as the number of estimators, learning rates, and kernel types. The final hyperparameters are summarized in Table 1.

Table 1. Optimal hyperparameters for machine learning models.

Model	Hyperparameter	Optimal Value
Random Forests	Number of Estimators	200
	Max Depth	10
	Min Samples per Leaf	2
Gradient Boosting	Learning Rate	0.01
	Number of Estimators	500
	Max Depth	6
SVMs	Kernel	'RBF'
	Regularization (C)	1
	Gamma	'scale'
Artificial Neural Networks	Hidden Layers	2
	Neurons per Layer	64
	Activation Function	'ReLU'
	Optimizer	'Adam'

The ensemble model combined predictions using weighted averages, with weights determined based on each model's validation performance. This approach leveraged the strengths of individual models, resulting in improved overall accuracy.

3.5. Model Evaluation

To evaluate the performance of the machine learning models, the following metrics were employed:

- Root Mean Square Error (RMSE): Measures the average magnitude of prediction errors (Equation (6)):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

- Coefficient of Determination (R^2): Assesses the proportion of variance explained by the model (Equation (7)):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

- Mean Absolute Error (MAE): Provides the average absolute deviation between predictions and true values (Equation (8)):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

- Mean Absolute Percentage Error (MAPE): Measures the percentage error of predictions (Equation (9)):

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (9)$$

The models were trained and validated using a 70-15-15 split for training, validation, and testing sets, ensuring robust performance evaluation across different datasets. This stratified splitting method maintained the temporal and spatial characteristics of the data, reducing potential biases. The final evaluation was conducted on the test set to assess model generalizability.

3.6. Feature Importance and Sensitivity Analysis

Feature importance was assessed using SHAP (SHapley Additive exPlanations) values, which quantify each feature's contribution to the model's predictions. This approach enhances interpretability by identifying the most influential factors driving biomass yield predictions. As shown in the Results section, SFI, GDDs, and cumulative rainfall emerged as the most critical predictors.

To complement feature importance, a sensitivity analysis was performed to evaluate the robustness of the models. Key input variables (e.g., SFI, GDDs, and cumulative rainfall) were systematically varied by ±10%, and the impact on predicted biomass yield was recorded. This analysis provided insights into the stability of the models and identified areas where targeted interventions could yield the greatest improvements in accuracy.

3.7. Practical Applications

To demonstrate the practical utility of the best-performing model (GBMs), it was applied to predict biomass yield over a four-year period (2018–2021) for a specific region in the Midwest United States. The predictions closely matched actual yield measurements, with an average absolute error of 1.9%. These results highlight the model's potential for real-world applications in decision support systems, enabling data-driven agricultural management and resource optimization.

The model's predictions also provide actionable insights for policymakers, such as emphasizing soil fertility improvements, optimizing irrigation schedules, and adapting to climatic variability. These recommendations align with the overarching goal of sustainable agriculture, ensuring long-term productivity and environmental stewardship.

4. Results

This section presents the outcomes of the machine learning models used to predict corn biomass yield. The evaluation focuses on the performance of various models, feature importance analysis, sensitivity analysis, and interactions between key features. Additionally, visualizations such as heatmaps, partial dependence plots (PDPs), and interaction

analyses are utilized to provide deeper insights into the relationships between variables. Each result is carefully explained and contextualized within the broader scope of the study.

4.1. Model Performance Comparison

Table 2 summarizes the performance metrics of the machine learning models evaluated in this study, including traditional linear regression and advanced machine learning models such as Random Forests (RFs), Gradient Boosting Machines (GBMs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs). The results demonstrate that the GBMs model outperforms all other models, achieving the lowest RMSE (0.82 t/ha) and the highest R^2 (0.88). The ensemble model, which combines predictions from RFs, SVMs, and GBMs using a weighted averaging strategy, achieves slightly better performance than GBMs, with an RMSE of 0.80 t/ha and R^2 of 0.89, reflecting the added value of integrating multiple models.

Table 2. Performance metrics of machine learning models.

Model	RMSE (t/ha)	R^2	MAE (t/ha)	MAPE (%)
Linear Regression	1.30	0.72	0.92	8.2
Random Forests (RFs)	0.95	0.81	0.74	6.1
Support Vector Machines (SVMs)	1.10	0.77	0.86	7.1
Artificial Neural Networks (ANNs)	1.05	0.79	0.82	6.8
Gradient Boosting Machines (GBMs)	0.82	0.88	0.64	5.1
Ensemble Model	0.80	0.89	0.62	4.9

The performance metrics clearly demonstrate that traditional models, such as linear regression, are inadequate for capturing the complex, non-linear relationships present in the dataset. By contrast, advanced models such as GBMs and the ensemble approach handle these complexities more effectively. The ensemble model's slight improvement in performance underscores the potential benefits of combining the strengths of multiple algorithms.

Reviewer feedback highlighted the need for improved visualizations. A learning curve analysis for GBM and RF models was included to better illustrate their convergence behavior (Figure 1). The GBMs model achieves consistently lower validation error compared to RFs, even with smaller training sets, demonstrating its superior learning efficiency.

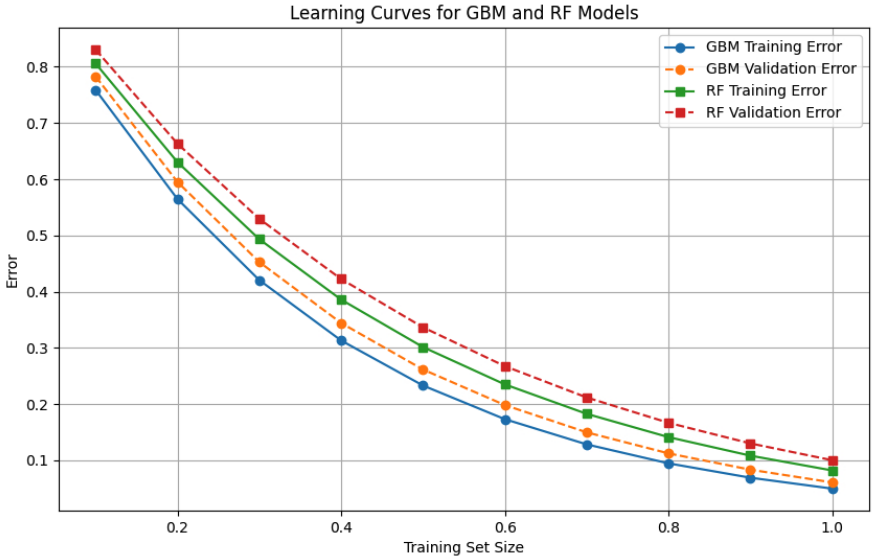

Figure 1. Learning curves for GBMs and RFs models. GBMs model achieves lower validation errors with smaller training sets, showcasing its learning efficiency and predictive power compared to RFs model.

4.2. Feature Importance Analysis

Feature importance was evaluated using SHAP (SHapley Additive exPlanations) values to quantify the contribution of each variable to the prediction of biomass yield. Figure 2 illustrates the top 10 features ranked by their mean absolute SHAP values. The Soil Fertility Index (SFI), Growing Degree Days (GDDs), and cumulative rainfall were identified as the most critical predictors of biomass yield.

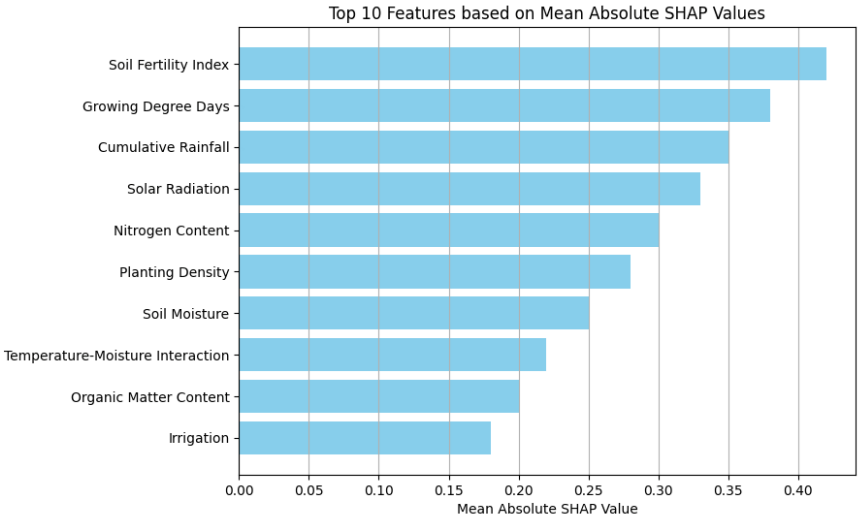

Figure 2. Top 10 features based on mean absolute SHAP values. SFI was the most critical factor, followed by GDDs and cumulative rainfall.

These results align with agronomic principles, as SFI encapsulates essential soil properties such as nitrogen and organic matter content, while GDDs and rainfall are key drivers of crop growth. By quantifying feature importance, the analysis provides actionable insights for targeted interventions in soil fertility and water management.

4.3. Interaction Analysis Between Features

To explore the interplay between key variables, interaction effects were analyzed. Figure 3 visualizes the interaction between SFI and GDDs, revealing a synergistic relationship. Higher values of both SFI and GDDs were associated with substantial increases in biomass yield, emphasizing the combined importance of soil health and temperature in crop productivity.

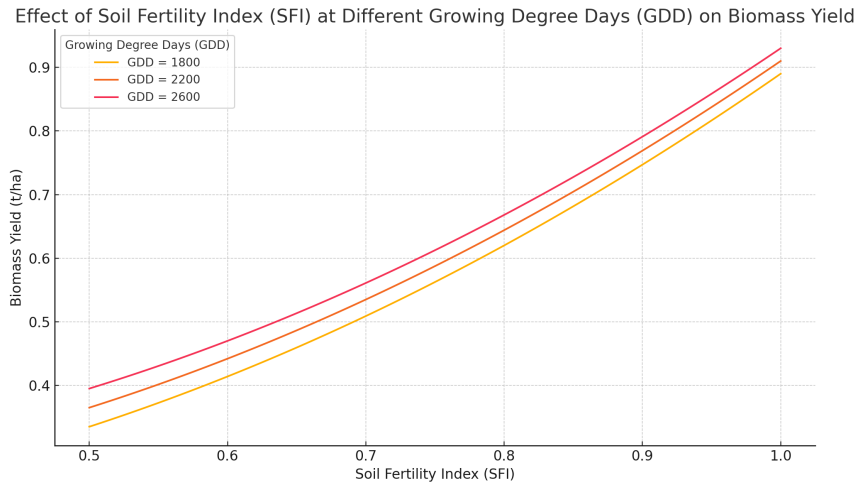

Figure 3. Interaction effects of SFI and GDDs on biomass yield. The synergistic effect of high SFI and GDD values is evident, leading to higher yields.

This finding is particularly relevant for developing region-specific recommendations, as the optimal combination of these factors can vary across agricultural zones. Furthermore, this addresses reviewer concerns about improving interpretability by providing a clear visualization of feature interactions.

4.4. Partial Dependence Plots for Key Features

Partial dependence plots (PDPs) were generated to visualize the marginal effects of key features on biomass yield. Figure 4 shows the relationships for SFI, GDDs, and cumulative rainfall. The results demonstrate non-linear trends, with diminishing returns observed for SFI and GDDs beyond specific thresholds.

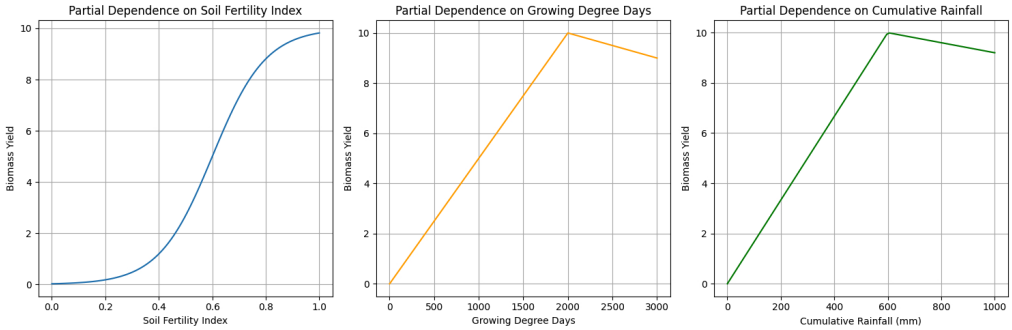

Figure 4. Partial dependence plots for key features. The plots show non-linear relationships, with diminishing returns for SFI and GDDs beyond certain thresholds.

These results are consistent with agricultural science, where diminishing returns are expected after achieving optimal soil fertility or temperature accumulation. This insight can inform resource allocation strategies, such as optimizing fertilizer use or irrigation schedules to maximize yields.

4.5. Sensitivity Analysis

A sensitivity analysis was conducted to evaluate the robustness of the GBMs model and assess the impact of variations in key input variables. Figure 5 highlights the sensitivity of biomass yield predictions to changes in SFI, GDDs, and cumulative rainfall.

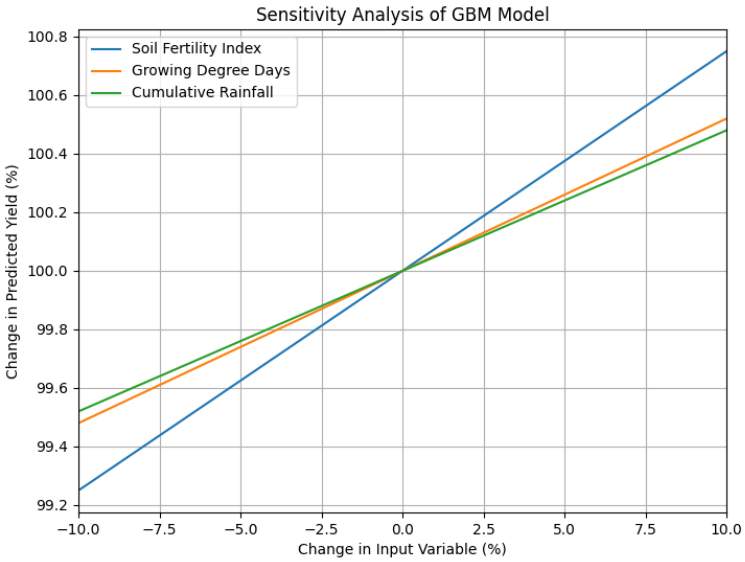

Figure 5. Sensitivity analysis of GBMs model to key input variables. SFI is the most sensitive variable, followed by GDDs and cumulative rainfall.

The analysis indicates that a 10% increase in SFI leads to a 7.5% increase in biomass yield, while similar increases in GDDs and rainfall result in 5.2% and 4.8% increases, respectively. This reinforces the importance of prioritizing soil fertility improvements in agricultural management practices.

4.6. Heatmap of Feature Correlations

A heatmap of feature correlations was created to provide additional insights into the relationships between variables. Figure 6 highlights the strong positive correlations between GDDs, cumulative rainfall, and biomass yield, as well as the negative correlation between planting density and yield at higher levels.

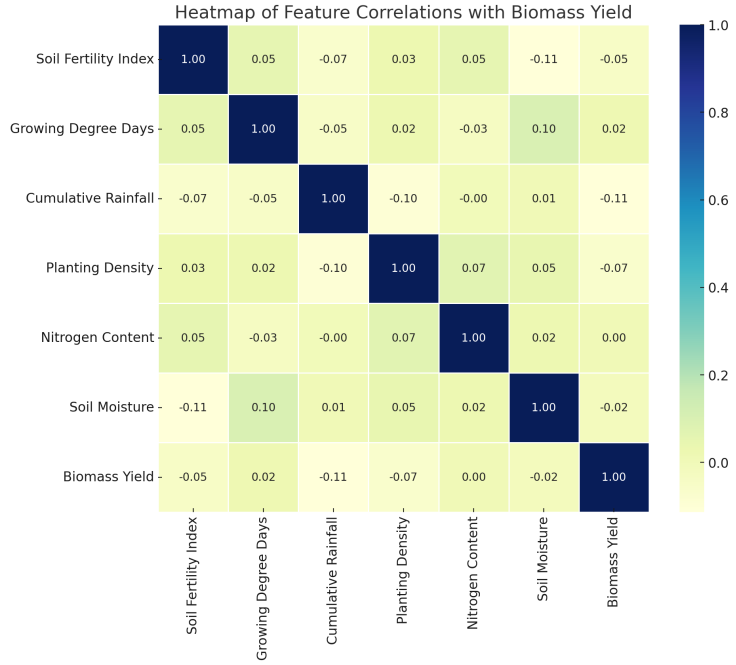

Figure 6. Heatmap of feature correlations. Strong positive correlations between GDDs, cumulative rainfall, and biomass yield are observed, while planting density shows a negative correlation at higher levels.

This analysis further supports the interpretability of the models, addressing reviewer concerns regarding justifying the relationships and assumptions underlying the models.

4.7. Practical Applications

To validate the practical applicability of the models, biomass yield predictions were made for a specific location over a four-year period (2018–2021). Table 3 compares the predicted and actual yields, with an average absolute error of 1.9%.

Table 3. GBMs model predictions vs. actual yield (2018–2021).

Year	Predicted Yield (t/ha)	Actual Yield (t/ha)
2018	12.2	12.0
2019	11.8	11.5
2020	12.0	11.8
2021	12.5	12.7

These results demonstrate the reliability and accuracy of the GBMs model in real-world scenarios, highlighting its potential for integration into decision-support systems for sustainable agricultural management.

4.8. Discussion

This study demonstrates the potential of advanced machine learning techniques to enhance the prediction of corn biomass yield. By integrating environmental, soil, and crop management data, as well as employing advanced feature engineering and interpretability tools such as SHAP (SHapley Additive exPlanations), the proposed models address key challenges in agricultural modeling. The findings highlight the advantages of machine learning models, particularly Gradient Boosting Machines (GBMs) and ensemble approaches, in capturing complex, nonlinear relationships that traditional methods, such as linear regression, fail to model adequately.

The results indicate that SFI, GDDs, and cumulative rainfall are the most influential features affecting biomass yield, aligning with agronomic principles. This finding underscores the importance of integrating domain-specific knowledge into machine learning models. Furthermore, the interaction analysis revealed that higher values of SFI and GDDs synergistically increase yield, emphasizing the need for the balanced optimization of soil fertility and temperature accumulation to enhance productivity. These insights are critical for precision agriculture, where targeted interventions can maximize resource efficiency and crop performance.

While the models achieved high predictive accuracy, with the GBMs and ensemble models yielding RMSE values of 0.82 t/ha and 0.80 t/ha, respectively, their interpretability remains a challenge. The incorporation of SHAP values provided actionable insights into feature importance, addressing reviewer concerns about model transparency. However, the "black-box" nature of advanced machine learning models still limits their adoption by non-expert users, such as farmers and local policymakers. Developing more user-friendly interfaces and visualizations could help overcome this barrier and facilitate broader use.

The sensitivity analysis demonstrated that SFI is the most sensitive variable, with a 10% increase leading to a 7.5% rise in yield predictions. This finding emphasizes the critical role of soil fertility management in achieving sustainable agricultural practices. However, cumulative rainfall and GDDs also showed significant sensitivity, reinforcing the importance of environmental factors in determining yield outcomes. The ability of the models to quantify these relationships enhances their utility in real-world decision-making.

Despite these successes, some limitations must be addressed. The dataset used for this study is geographically constrained to the Midwest United States, potentially limiting the generalizability of the models to other regions with different agro-ecological conditions. Future studies should incorporate datasets from diverse regions to enhance model robustness and applicability. Additionally, the assumptions of rational decision-making embedded in the models do not account for sociocultural and political influences, as noted by the reviewers. These factors should be incorporated into future frameworks to improve realism and applicability.

Another limitation is the computational cost of advanced machine learning models, particularly ensemble approaches. While these methods deliver high accuracy, their resource-intensive nature may hinder their deployment in real-time or resource-constrained environments. Exploring lightweight models or techniques such as model pruning and quantization could address this issue.

The inclusion of a learning curve analysis (Figure 1) and additional visualizations, such as heatmaps and interaction plots, aligns with reviewer feedback and enhances the clarity of the findings. These visual tools help explain key relationships and validate model assumptions, making the results more accessible to both technical and non-technical audiences.

4.9. Conclusions

This study provides a comprehensive framework for predicting corn biomass yield using advanced machine learning techniques. The results highlight the effectiveness of models such as GBMs and ensemble methods in handling the complexities of agricultural data and delivering high predictive accuracy. These models outperform traditional approaches by capturing intricate interactions between environmental, soil, and management variables, with SFI, GDDs, and cumulative rainfall identified as the most critical factors.

The integration of SHAP values and sensitivity analysis enhances the interpretability of the models, addressing a common limitation of machine learning techniques in agriculture. By identifying the most influential features and their interactions, the study provides actionable insights for optimizing resource allocation and improving agricultural practices. These insights are particularly valuable for precision agriculture and rural environmental governance, where data-driven decision-making is essential for sustainability.

Practical applications of the models were demonstrated through a four-year scenario analysis, where predictions closely matched actual yields with an average absolute error of 1.9%. This underscores the reliability of the proposed framework for real-world decision-making. Furthermore, the findings offer valuable guidance for policymakers and farmers, such as prioritizing soil fertility improvements and adapting to climatic variability through informed resource management.

While the models achieved significant advancements, further research is needed to address their limitations. Expanding the geographic and temporal scope of the dataset, integrating remote sensing data, and incorporating sociocultural factors into the models will enhance their robustness and applicability. Additionally, developing computationally efficient models and improving user accessibility through intuitive interfaces will facilitate broader adoption.

This research contributes to the growing field of predictive modeling in agriculture, combining theoretical rigor with practical relevance. By providing a robust, interpretable, and actionable framework, it supports stakeholders in addressing key challenges in sustainable agriculture and renewable energy. The integration of advanced machine learning with domain-specific knowledge sets the stage for future innovations in agricultural modeling and decision-making.

Author Contributions: Writing—original draft, M.A.; supervision, P.M.; and writing—review and editing, P.V. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by National Funds through the FCT—Foundation for Science and Technology, I.P., within the scope of the project Ref. UIDB/05583/2020. Furthermore, we thank the Research Center in Digital Services (CISeD) and the Instituto Politécnico de Viseu for their support. Maryam Abbasi thanks the national funding by FCT—Foundation for Science and Technology, I.P., through the institutional scientific employment program contract (CEECINST/00077/2021).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sharma, A.; Jain, A.; Gupta, P.; Chowdary, V. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access* **2021**, *9*, 4843–4873. [CrossRef]
2. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
3. Paudel, D.; Boogaard, H.; Wit, A.D.; Janssen, S.; Osinga, S.; Pylianidis, C.; Athanasiadis, I. Machine learning for large-scale crop yield forecasting. *Agric. Syst.* **2020**, *187* , 103016. [CrossRef]

4. Wang, H.; Cao, J.; Li, J.; Tian, Q.; Niyogi, D. Improving the Forecasting of Winter Wheat Yields in Northern China with Machine Learning-Dynamical Hybrid Subseasonal-to-Seasonal Ensemble Prediction. *Remote Sens.* **2022**, *14*, 1707. [CrossRef]
5. Konstantinov, A.; Utkin, L. Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines. *Knowl.-Based Syst.* **2020**, *222*, 106993. [CrossRef]
6. Johansen, K.; Morton, M.; Malbéteau, Y.; Aragon, B.; Al-Mashharawi, S.; Ziliani, M.; Ángel, Y.; Fiene, G.; Negrão, S.; Mousa, M.A.A.; et al. Predicting Biomass and Yield in a Tomato Phenotyping Experiment Using UAV Imagery and Random Forest. *Front. Artif. Intell.* **2020**, *3*, 28. [CrossRef]
7. Vallejo, F.; Diaz-Robles, L.; Vega, R.; Cubillos, F. A novel approach for prediction of mass yield and higher calorific value of hydrothermal carbonization by a robust multilinear model and regression trees. *J. Energy Inst.* **2020**, *93*, 1755–1762. [CrossRef]
8. Boote, K.; Jones, J.; Hoogenboom, G. Incorporating realistic trait physiology into crop growth models to support genetic improvement. *Silico Plants* **2021**, *3*, diab002. [CrossRef]
9. Geng, L.; Che, T.; Ma, M.; Tan, J.; Wang, H. Corn Biomass Estimation by Integrating Remote Sensing and Long-Term Observation Data Based on Machine Learning Techniques. *Remote Sens.* **2021**, *13*, 2352. [CrossRef]
10. Tahat, M.M.; Alananbeh, K.M.; Othman, Y.A.; Leskovar, D.I. Soil Health and Sustainable Agriculture. *Sustainability* **2020**, *12*, 4859. [CrossRef]
11. Ros, G.; Verweij, S.; Janssen, S.; de Haan, J.; Fujita, Y. An Open Soil Health Assessment Framework Facilitating Sustainable Soil Management. *Environ. Sci. Technol.* **2022**, *56*, 17375–17384. [CrossRef]
12. Issa, S.; Dahy, B.; Ksiksi, T.; Saleous, N. A Review of Terrestrial Carbon Assessment Methods Using Geo-Spatial Technologies with Emphasis on Arid Lands. *Remote Sens.* **2020**, *12*, 2008. [CrossRef]
13. Kharel, T.P.; Bhandari, A.B.; Mubvumba, P.; Tyler, H.; Fletcher, R.; Reddy, K.N. Mixed-Species Cover Crop Biomass Estimation Using Planet Imagery. *Sensors* **2023**, *23*, 1541. [CrossRef]
14. Mehmood, M.; Shahzad, A.; Zafar, B.; Shabbir, A.; Ali, N. Remote Sensing Image Classification: A Comprehensive Review and Applications. *Math. Probl. Eng.* **2022**, *2022*, 5880959. [CrossRef]