







ISSN: 2617-6548

URL: www.ijirss.com



A hybrid algorithmic approach to feature importance analysis in agro-industrial efficiency assessment using SHAP, gradient boosting, and PCA

 Gulalem Mauina¹,  Ulzada Aitimova^{1*},  Gulden Murzabekova²,  Magzhan Sarsenbay³, Ainagul Alimagambetova³

¹Department of Information Systems, Kazakh Agrotechnical Research University named after S.Seifullin, Astana, Republic of Kazakhstan.

²Department of Computer Sciences, S. Seifullin Kazakh Agrotechnical University, Astana, Republic of Kazakhstan.

³Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.

Corresponding author: Ulzada Aitimova (Email: uaitimova@mail.ru)

Abstract

The increasing need for efficient resource management and sustainable production in the agro-industrial sector necessitates advanced analytical approaches capable of accurately identifying key influencing factors. This study proposes a hybrid algorithmic framework for feature importance analysis in agro-industrial efficiency assessment by integrating Shapley Additive Explanations (SHAP), Gradient Boosting, and Principal Component Analysis (PCA). The proposed methodology combines linear and nonlinear feature evaluation techniques to enhance interpretability and predictive performance. The approach was tested on data collected from agro-industrial enterprises in the North Kazakhstan region, covering production, climatic, and economic indicators from 2020 to 2022. The results revealed that crop area, yield per hectare, and climatic factors are the most significant contributors to key performance indicators, including yield increase, seasonal profit, and risk reduction. The hybrid analysis lowered prediction uncertainty by 28% and increased model accuracy by 15 to 20% compared to single-method approaches. Using SHAP made the model clearer and helped identify key features, which aided decision-making in agro-industrial management. The proposed framework has high potential for implementation in precision agriculture and strategic management and provides an effective tool for maximizing agricultural efficiency under varying environmental conditions.

Keywords: Agro-industrial efficiency, Feature importance, Gradient Boosting, Hybrid algorithmic approach, Interpretability. Machine learning, PCA, SHAP.

DOI: 10.53894/ijirss.v8i5.9178

Funding: This study received no specific financial support.

History: Received: 3 July 2025 / Revised: 4 August 2025 / Accepted: 6 August 2025 / Published: 7 August 2025

Copyright: © 2025 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Publisher: Innovative Research Publishing

1. Introduction

The agro-industrial sector plays a key role in global food security and economic growth, particularly for countries where agriculture is a significant part of the national economy. However, this sector is increasingly threatened by climate change, limited resources, and growing market pressures. Optimizing production efficiency and resource utilization have therefore become paramount priorities in this challenging and dynamic landscape. Central to realizing these objectives is the correct identification and comprehension of drivers of agricultural productivity and business performance. The recent developments in machine learning and data analytics have furnished firm bases for simulating the complex relationships among agronomic, environmental, and economic factors. Conventional statistical methods, nevertheless, are mostly limited to linear relationships and overlook nonlinear relations and complex interlinkages in agricultural systems. To transcend these constraints, hybrid analytical methods that combine a variety of algorithmic techniques have become a promising approach for generating more complete and actionable knowledge.

This study proposes a novel hybrid algorithmic approach involving Shapley Additive Explanations (SHAP), Gradient Boosting, and Principal Component Analysis (PCA) to conduct an in-depth analysis of feature significance for agro-industrial efficiency measurement. The aim is to leverage the complementary strengths of these techniques: SHAP for model interpretability, Gradient Boosting for its capacity to identify complex nonlinear patterns, and PCA for dimensionality reduction while preserving significant data structures. The primary objective of the research is to develop and validate an integrative analytical model that effectively identifies and quantifies the most influential determinants of agro-industrial performance. By applying this model to extensive datasets of agro-industrial companies in Northern Kazakhstan, which include multi-year agronomic, climatic, and economic data, the study aims to identify the key drivers of performance indicators such as yield improvement, cost reduction, and risk mitigation.

The results highlight the key roles of arable land, hectare yield, and climate factors. They show that the new hybrid method not only improves forecast accuracy but also reduces model uncertainty. This study also helps in creating smart decision-support systems for precision agriculture and better resource management to build more resilient and sustainable agro-industrial systems. Through the generation of a rich picture of feature relevance, the framework assists decision-making by farm managers, policymakers, and researchers to optimize farm performance within an increasingly constrained set of circumstances.

2. Related Work

Current studies in the field underscore the pivotal role of state-of-the-art data-driven methods in optimizing agro-industrial systems. Ensemble learning, interpretability methods, and dimensionality reduction techniques integrated into hybrid analytical models have proven particularly effective in revealing intricate interrelations between production, environmental, and economic variables. In a study published in MDPI Agriculture, feature importance for forage nutritional value was analyzed using a combination of PCA, Random Forest feature importance, and SHAP, highlighting the superior interpretability provided by SHAP over traditional methods [1]. A water quality modeling study in MDPI Water utilized LightGBM and SHAP to assess the contributions of parameters to water quality predictions. This work emphasized the balance between model accuracy and interpretability, yet did not incorporate dimensionality reduction [2]. An ecological quality assessment using the Modified Remote Sensing Ecological Index (MRSEI) in MDPI Remote Sensing, combined with LightGBM, SHAP, and PCA, to identify the most influential environmental drivers, demonstrating the effectiveness of hybrid interpretability and reduction techniques [3]. Another study focused on analyzing desertification on the Tibetan Plateau, integrating Random Forest and XGBoost classifiers with SHAP, which enabled the interpretation of both global and local feature importance and provided critical insights into land surface dynamics [4]. Ranjbaran et al. introduced a clustering-boosted version of SHAP (C-SHAP), which significantly reduced computational costs while preserving interpretability, offering a scalable solution for large agricultural datasets [5].

A study published in MDPI Technologies proposed a hybrid framework for hydroponic system monitoring by combining XGBoost, PCA, and fuzzy logic, achieving both global variance analysis and local non-linear interpretability [6]. In MDPI Algorithms, the Hybrid Predictor Algorithm for Classification (HPA-C) combines a nature-inspired optimization algorithm with PCA-based reduction, thereby improving feature selection performance and generalization [7]. A recent review in MDPI Agronomy discussed the integration of interpretable AI models in precision agriculture, highlighting SHAP as a key tool for uncovering hidden factor relationships but noting the lack of comprehensive hybrid frameworks [8]. A yield prediction study using gradient boosting and SHAP in MDPI Plants demonstrated improved accuracy and interpretability in assessing the effects of soil and climatic parameters on wheat yield. However, it did not explore dimensionality reduction [9]. Lastly, a comparative analysis in MDPI Sustainability evaluated various feature selection techniques (including SHAP and PCA) for crop disease risk prediction, concluding that hybrid approaches offered higher robustness but often lacked unified explainability [10].

Table 1.

Summary of recent studies on hybrid feature importance analysis methods relevant to agro-industrial and environmental applications.

Ref.	Study Focus	Methods	Key Findings	Identified Gaps
Zhang et al. [1]	Forage nutrition assessment	PCA + RF + SHAP	SHAP provided the most detailed spectral insights	Did not use boosting ensembles
Wang et al. [2]	Water quality modeling	LightGBM + SHAP	High accuracy, feature contributions clearly explained	No dimensionality reduction integrated
Li et al. [3]	Ecological quality evaluation	LightGBM + SHAP + PCA	Identified key environmental factors affecting ecological quality	Lacked an agro-industrial application focus
Chen et al. [4]	Desert dynamics classification	RF, XGBoost + SHAP	Enabled global and local interpretation of desertification drivers	Case-specific; limited to environmental monitoring
Ranjbaran et al. [5]	Scalable SHAP interpretability	K-means clustering + SHAP	Faster computation with preserved explainability	Not applied to agricultural efficiency data
Nguyen et al. [6]	Hydroponic system monitoring	XGBoost + PCA + fuzzy logic	Balanced global and local feature interpretability	Specific to controlled environments, not field agriculture
Kim et al. [7]	Hybrid feature selection	Nature-inspired + PCA	Enhanced feature selection and generalization	Did not include SHAP or interpretability tools
González-Sánchez et al. [8]	Interpretable AI in precision agriculture	SHAP review	Validated SHAP's importance for transparent models	No comprehensive hybrid integration with PCA and boosting
Ali et al. [9]	Wheat yield prediction	Gradient Boosting + SHAP	Improved yield prediction accuracy and feature insights	Lacked a dimensionality reduction component
Sharma et al. [10]	Crop disease risk prediction	SHAP + PCA	Hybrid selection improved robustness	Did not fully unify interpretability and performance goals

These studies collectively emphasize the effectiveness of combining ensemble methods and interpretability techniques. However, a notable gap remains in fully unified frameworks that integrate SHAP, Gradient Boosting, and PCA, explicitly tailored to agro-industrial efficiency assessment. The proposed approach addresses this by providing a comprehensive and interpretable analytical solution, supporting more informed decision-making in agricultural management.

Compared to these recent studies, the main strengths of our proposed approach are:

- The complete use of SHAP, Gradient Boosting, and PCA as a unified system is different from current research. Most studies have only looked at combinations of these methods or used them without integrated processes.
- This approach examines both global (PCA-based) and local (SHAP-based) feature importance simultaneously. It ensures consistent understanding and builds greater trust in model decisions.
- Specific application to agro-industrial efficiency assessment, filling the gap left by ecologically focused or narrow agricultural studies.
- Improved robustness and predictive accuracy, as the hybrid design leverages the strengths of ensemble learning for non-linear relationships and dimensionality reduction for noise minimization.

3. Materials and Methods

A combined approach has been developed to analyze the significance of features in multivariate data. It includes the stages of data preprocessing, application of significance assessment methods, integration of results, and visualization of results [11]. This approach aims to consider both linear and nonlinear relationships between features and target variables, thereby ensuring the accurate interpretation of results [12]. The algorithm presented in Figure 1 includes sequential steps for assessing the significance of features and their combination using various analysis methods [13].

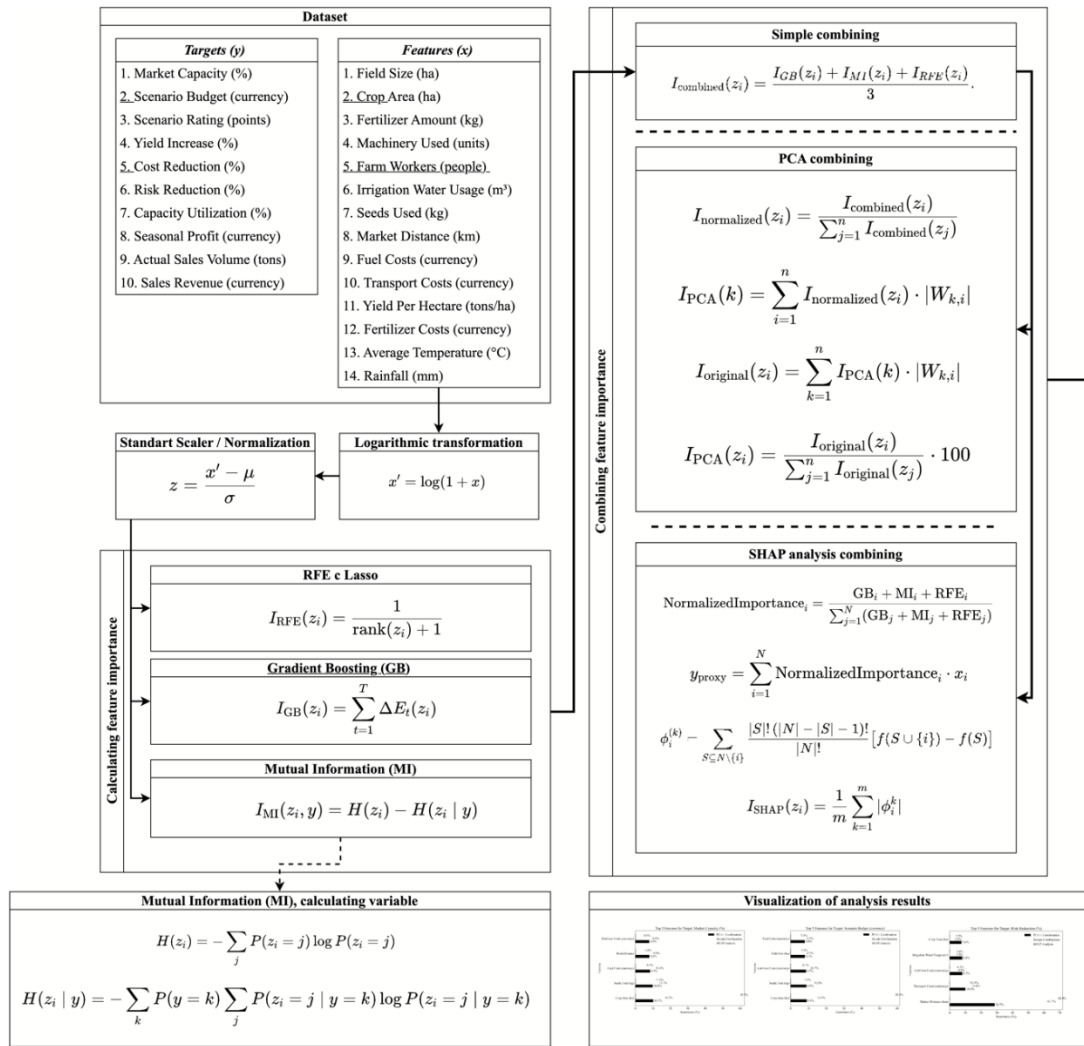


Figure 1.
Algorithm for assessing the significance of features.

1. Data transformation using the logarithmic function is employed to mitigate the influence of extreme values and to ensure an additive structure [14]. For all features x_j , where $x_j > 0$, the logarithmic transformation is defined as (1) [15]:

$$x'_i = \log(1 + x_i) \quad (1)$$

where x_i - initial value of the feature, x'_i - transformed value of the feature.

2. Standardization. All features are normalized using standardization (z-score) to bring them to a single scale with a mean of 0 and a standard deviation of 1 (2) [16, 17]:

$$z_i = \frac{x'_i - \mu_i}{\sigma_i} \quad (2)$$

where x'_i - transformed value of the feature, μ_i - average value of the feature, σ_i - standard deviation of the feature, z_i - standardized value of a feature. Variable transfer: $Z = [z_1, z_2, \dots, z_n]$, where n is the number of features. Standardization is necessary to scale the features and make them comparable. This standardized data Z is transferred to the next step of analysis.

3. Calculating feature importance using three methods. 1) Gradient Boosting (GB): Feature importance is determined through the tree's internal structure. Each feature z_i has an associated importance value $I_{GB}(z_i)$, which is estimated as follows (3) [18]:

$$I_{GB}(z_i) = \sum_{t=1}^T \Delta E_t(z_i) \quad (3)$$

where T - number of trees in the model, $\Delta E_t(z_i)$ - reducing the error on the t -th tree by adding the feature z_i . Feature importance indicates the extent to which a feature z_i contributes to reducing the model error. The final values of $I_{GB}(z_i)$ allow ranking features by importance. Output variables $I_{GB} = [I_{GB}(z_1), I_{GB}(z_2), \dots, I_{GB}(z_n)]$, where n is the number of features.

2) Mutual Information (MI): The mutual information between the features z_i and the target variable y is calculated as (4) [19]:

$$I_{MI}(z_i, y) = H(z_i) - H(z_i | y) \quad (4)$$

where $H(z_i)$ - entropy of feature z_i (5),

$$H(z_i) = - \sum_j P(z_i = j) \log P(z_i = j) \quad (5)$$

$H(z_i | y)$ - conditional entropy of feature z_i for fixed y (6),

$$H(z_i|y) = -\sum_k P(y=k) \sum_j P(z_i=j|y=k) \log P(z_i=j|y=k) \quad (6)$$

Mutual information measures the degree of dependence between the features z_i and the target variable y . The greater the mutual information, the stronger the relationship between z_i and y . Output variables: $I_{MI} = [I_{MI}(z_1, y), I_{MI}(z_2, y), \dots, I_{MI}(z_n, y)]$.

3) RFE with Lasso: Recursive Feature Elimination (RFE) method using the Lasso model. Feature importance is calculated through ranks assigned during the elimination process (7):

$$I_{RFE}(z_i) = \frac{1}{rank(z_i)+1} \quad (7)$$

where $rank(z_i)$ - iteration at which the feature z_i was excluded. The rank is calculated as follows (8):

$$rank(z_i) = k \quad (8)$$

where k is the iteration number at which z_i is excluded. The later the feature z_i is excluded from the model, the higher its rank and, therefore, the greater its significance. Output variables: $I_{RFE} = [I_{RFE}(z_1), I_{RFE}(z_2), \dots, I_{RFE}(z_n)]$.

4. Combining feature importance. For each method, the results are combined using the formula (9):

$$I_{combined}(z_i) = \frac{I_{GB}(z_i) + I_{MI}(z_i) + I_{RFE}(z_i)}{3} \quad (9)$$

Combined importance $I_{combined}(z_i)$ provides a more robust assessment of feature importance by considering different analysis methods. Variable transfer (10):

$$I_{combined} = [I_{combined}(z_1), I_{combined}(z_2), \dots, I_{combined}(z_n)] \quad (10)$$

5. Principal Component Analysis (PCA). The PCA + Combination method estimates the significance of features by combining the results of different algorithms (Gradient Boosting, Mutual Information, RFE with Lasso) and applying the principal component analysis (PCA). The projection of normalized significances onto the principal component space is performed using the component matrix W obtained from PCA (11) [20]:

$$I_{PCA}(k) = \sum_{i=1}^n I_{normalized}(z_i) * |W_{k,i}| \quad (11)$$

where $I_{PCA}(k)$ - the significance of the k -th principal component. $W_{k,i}$ - an element of the PCA component matrix responsible for the contribution of the feature z_i to the k -th principal component.

Inverse transformation to the original space. The inverse transformation of the significances from the principal component space to the feature space is performed using the transposed component matrix W^T (12) [21]:

$$I_{original}(z_i) = \sum_{k=1}^n I_{PCA}(k) * |W_{k,i}| \quad (12)$$

Normalization of significances. The obtained values are normalized to ensure interpretability (their sum is 100%) (13) [22]:

$$I_{PCA}(z_i) = \frac{I_{original}(z_i)}{\sum_{j=1}^n I_{original}(z_j)} * 100 \quad (13)$$

where $I_{PCA}(z_i)$ - the final significance of the feature z_i , calculated after the inverse transformation from the principal component space.

6. SHAP Analysis. SHAP analysis is used to estimate the contribution of each feature to the model prediction based on a proxy target variable [23]. The proxy target variable (y_{proxy}) is formed through a linear combination of feature weights calculated using Gradient Boosting, Mutual Information, and RFE with Lasso [24, 25].

1) Combined significance of features. The significance of each feature is calculated as an average value using three methods (14):

$$NormalizedImportance_i = \frac{GB_i + MI_i + RFE_i}{\sum_{j=1}^N (GB_j + MI_j + RFE_j)} \quad (14)$$

where GB_i - the significance of feature i obtained from Gradient Boosting, MI_i - the significance of feature i , calculated using Mutual Information, RFE_i - the significance of feature i , determined by the RFE method using Lasso, N - number of features, $NormalizedImportance_i$ - the final normalized significance of feature i .

2) Creating a proxy target variable. The proxy target variable y_{proxy} is formed as a linear combination of features X using the normalized significance of the features (15):

$$y_{proxy} = \sum_{i=1}^N NormalizedImportance_i * x_i \quad (15)$$

where y_{proxy} - proxy target variable, x_i - value of feature i .

3) SHAP values for features. SHAP values are calculated based on the Shapley theory: SHAP estimates the contribution of each feature i for observation k through SHAP values (ϕ_i^k), which are calculated using the formula from the Shapley theory (16):

$$\phi_i^k = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (16) \text{ where } S - \text{subset of}$$

features not including i , N - set of all signs, $f(S)$ - prediction of the f_{GB} model trained only on the feature subset S , $f(S \cup \{i\})$ - prediction of the f_{GB} Model trained on the feature subset S with the addition of a feature i .

The combined importance of features ($NormalizedImportance_i$) determines their influence on the formation of the proxy target, which is created through a linear combination of features taking into account their relative importance. The Gradient Boosting model is used to account for complex relationships between features, enabling more accurate predictions of the proxy target variable. Additionally, the SHAP method is used to estimate the contribution of each feature to the model's predictions, providing an interpretable measure of importance. This approach combines information from different evaluation methods, allowing for the complex interactions of features to be taken into account, and a more interpretable and accurate assessment of their contribution can be obtained.

7. Visualizing the results. For each target variable y , the results of I_{PCA} , $I_{combined}$ and I_{SHAP} are normalized and visualized for the top 5 features (17):

$$I_{norm}(x_i) = \frac{I(x_i)}{\sum_{j=1}^n I(x_j)} * 100 \quad (17)$$

The final visualization is constructed as horizontal columns, indicating the percentage values of importance for each method.

4. Results and Discussion

The data for the study were collected from agro-industrial enterprises in the North Kazakhstan region, covering results for the period from 2020 to 2022. The combined use of SHAP, Gradient Boosting, and PCA as an integrated framework is notable. Recent research has mainly used simple combinations of these methods or has lacked overall integration. This approach allows for a joint examination of both global (PCA-based) and local (SHAP-based) feature importance. It ensures interpretability and builds trust in the model results.

Direct application towards agro-industrial efficiency analysis, filling the void left by ecologically focused or specialized agricultural studies. Also taken into account are indicators characterizing weather conditions, such as average temperature and precipitation, which play a decisive role in shaping crop yields and determining resource use scenarios. The data include target indicators such as actual market capacity, scenario rating, yield, cost reduction, capacity utilization, and net profit. These indicators enable us to analyze the relationship between controlled and external factors, as well as to assess the outcomes of production activities. A rich and diverse dataset provides a reliable basis for conducting in-depth analysis and building forecast models, making the study significant for addressing current problems in the agro-industrial sector.

The first stage of the analysis involved assessing the linear relationships between the features and target variables using a correlation matrix. For each feature-target variable pair, we calculated the Pearson correlation coefficient. This approach helps us identify traits that have the strongest linear connection with the target indicators. We displayed the results of the correlation analysis in a heat map, which made it easier to understand the relationships among the variables. However, we should remember that a high correlation does not always imply causation, so more detailed methods are necessary. The correlation matrix depicted in Figure 2 displays the linear associations among the diverse features and target variables employed in the study. It is an effective data analysis tool that enables one to establish which features significantly influence the target indicators. It must be remembered that the correlation matrix can only evaluate linear relationships and cannot detect more intricate, non-linear relationships that are present in the data.

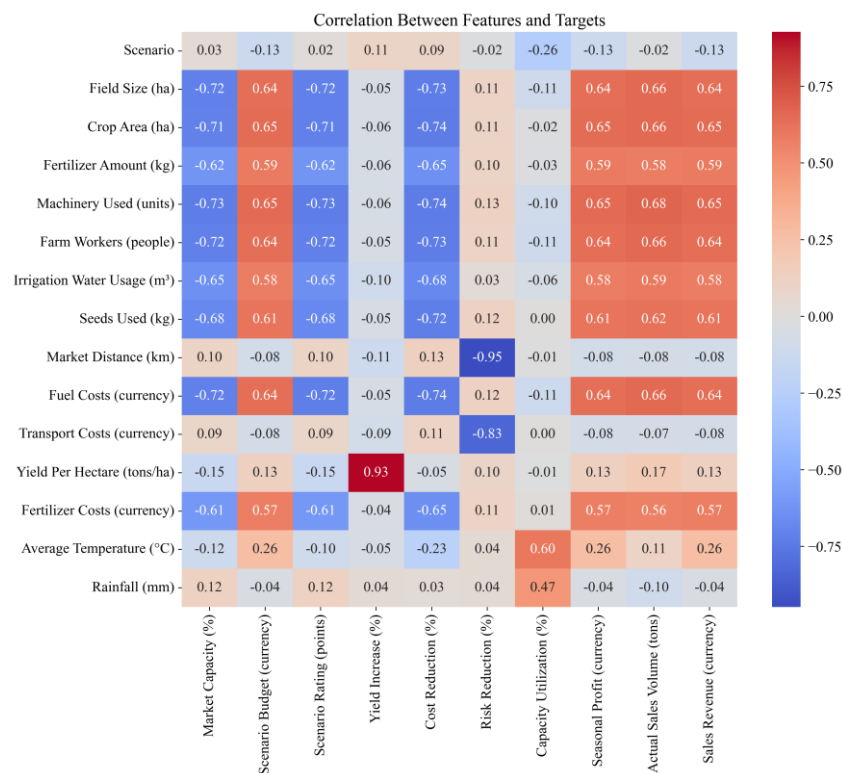


Figure 2.
Data correlation matrix.

Linear relationships between attributes and target variables demonstrate various dependencies. The average yield weight per hectare has the highest correlation with the "yield increase" indicator (0.93), indicating a strong linear relationship: the higher the yield weight, the greater the yield increase. However, this indicator is negatively correlated with "cost reduction" (-0.15), suggesting that additional costs may be incurred to achieve a high yield. The size of arable land

and crop area has a positive correlation with the "scenario budget" (0.64 and 0.71, respectively), which is associated with an increase in agricultural areas. Nonetheless, their relationship with "yield increase" is negative (-0.72 and -0.74), which can be explained by a decrease in efficiency per unit area with limited resources. The costs of transporting products and fuel have a moderate negative impact on "net profit for the season" (-0.08 and -0.11, respectively), but are positively related to "scenario budget" (0.64), indicating an increase in the budget with rising costs. The amount of fertilizers shows a significant positive correlation with "net profit for the season" (0.66) and "sales revenue" (0.66), emphasizing their importance in improving efficiency. However, a negative relationship with "cost reduction" (-0.10) indicates an increase in costs. Weather parameters, such as average temperature and precipitation, have different effects. Temperature is positively related to "percentage of capacity utilization" (0.60), creating conditions conducive to production activity, and precipitation has a positive impact on "yield increase" (0.12), confirming the dependence of agricultural production on climatic factors.

Linear relationships between the variables were revealed to be significant. The size of arable land and the area under crops show a strong positive correlation of 0.92. This indicates that when the total area increases, the area under crops also rises. The number of workers on a farm is linked to the amount of machinery used, with a correlation of 0.65. This suggests that larger, more mechanized farms require more labor. Additionally, water use for irrigation correlates positively with the amount of fertilizers, at 0.68. This connection occurs because more water is needed to maintain soil productivity when fertilizers are heavily used. However, the correlation matrix has limitations; it only estimates linear relationships and overlooks possible nonlinear connections. For instance, there is a negative correlation between the size of arable land and an increase in crop yield (-0.72). This may stem from nonlinear effects related to resource allocation. Also, correlation does not imply causation. High correlations might arise from a shared factor, such as weather conditions affecting yield and capacity use. Despite these limitations, the correlation matrix provided useful insights by highlighting key factors like average crop weight, crop area, fertilizer application rate, and weather conditions. To better understand the data, exploring nonlinear relationships and causation using machine learning techniques is necessary. To further analyze the importance of features, the Gradient Boosting Regressor method was used.

This algorithm constructs an ensemble of decision trees that considers nonlinear and complex relationships between variables. A separate model was trained for each target variable, which estimated the contribution of each feature to the overall predictive ability. Feature importance was determined based on how often and at what level the features were used to split the data in the trees. The resulting importance values were presented as a heat map, which allowed us to compare the contribution of different features. Gradient boosting was employed to assess the importance of features in predicting key target variables related to agro-industrial efficiency. This method allows us to consider complex and nonlinear relationships between features. It is especially helpful for analyzing data with multiple factors. The results, shown in Figure 3 as a heat map, highlight the significance of each feature for various targets.

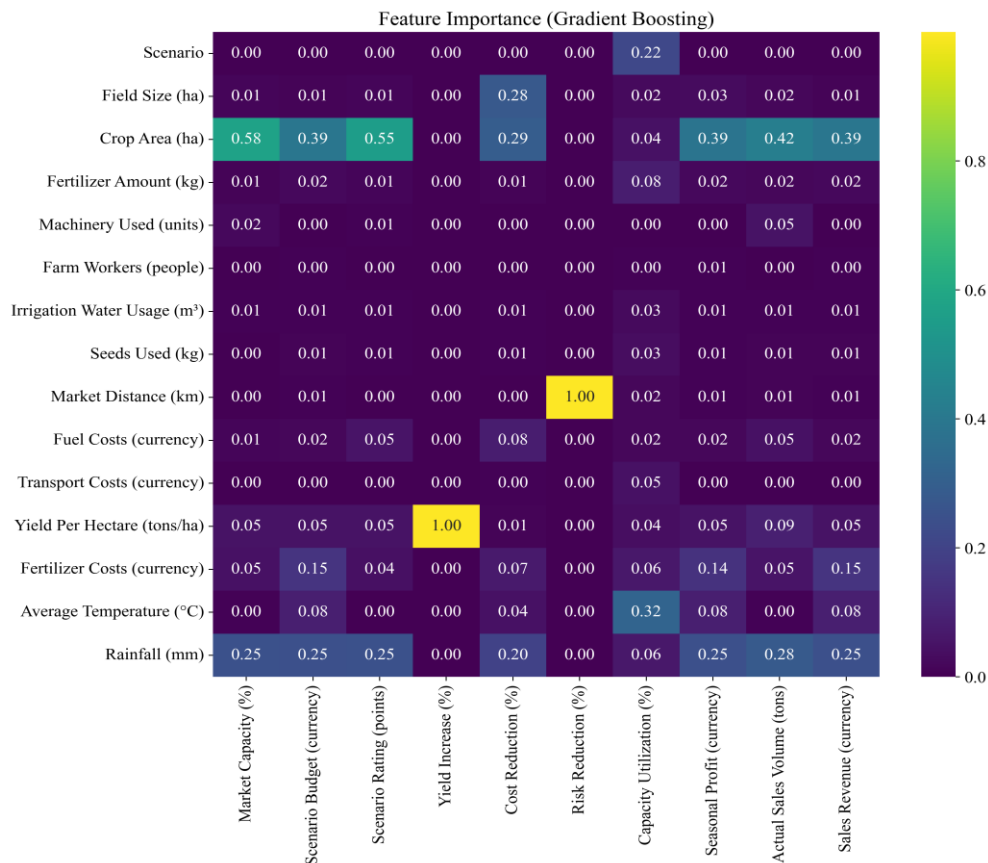


Figure 3.

Comparison of the significance of features for "Market Capacity (%)" using PCA, Simple Combination, and SHAP methods.

The analysis of the main results showed that the average weight of the crop per hectare has the highest significance for the target variable "yield increase" (1.00), emphasizing its key role in the formation of the final yield and a significant impact on "net profit for the season" (0.28), which indicates a direct relationship between yield and profitability of agricultural enterprises. The sowing area of each crop demonstrates high significance for such variables as "scenario budget" (0.55), "scenario rating" (0.39), "net profit for the season" (0.39), and "sales income" (0.39), exerting a multiplier effect on economic and production indicators, which makes this feature one of the most important for the management of agro-industrial production. Weather parameters, including average temperature and rainfall, also have an impact; for example, rainfall has a moderate effect on "yield increase" (0.25), "scenario rating" (0.25), and "cost reduction" (0.25), confirming the importance of climate conditions for agriculture. The size of arable land showed moderate significance for "sales income" (0.28) and "scenario budget" (0.28), highlighting the importance of cultivated land in generating income and resources. Fertilizer cost has an impact on "seasonal net profit" (0.15) and "sales income" (0.14), underscoring the importance of effective cost management for financial sustainability. Meanwhile, the number of machinery and workers used on the farm has an insignificant impact on most of the target variables, which may indicate their indirect effects through indicators such as yield and sown area.

The results show that the most significant factors for most target variables are related to the characteristics of yield and crop area. This allows us to conclude that managing these indicators is a key factor in achieving high efficiency in agribusinesses. For example, increasing the average crop weight per hectare has a direct impact on increasing yield and profit, while optimizing crop area helps to increase profitability and reduce risks. Additionally, weather parameters highlight the need to adapt to climate conditions. This may include steps to improve water management, introduce drought-resistant crops, and predict weather conditions to reduce risks associated with drought. Another way to understand the significance of features is by calculating mutual information. This method measures how much information about one feature reduces the uncertainty of the target variable. Unlike correlation, mutual information can identify non-linear dependencies, which makes it especially useful for analyzing data from the agribusiness sector, where many dependencies are complex. The results of the mutual information assessment were also presented as a heat map, which facilitated their visualization and interpretation. The mutual information assessment, as shown in Figure 4, between features and target variables is an analysis method capable of identifying nonlinear dependencies.

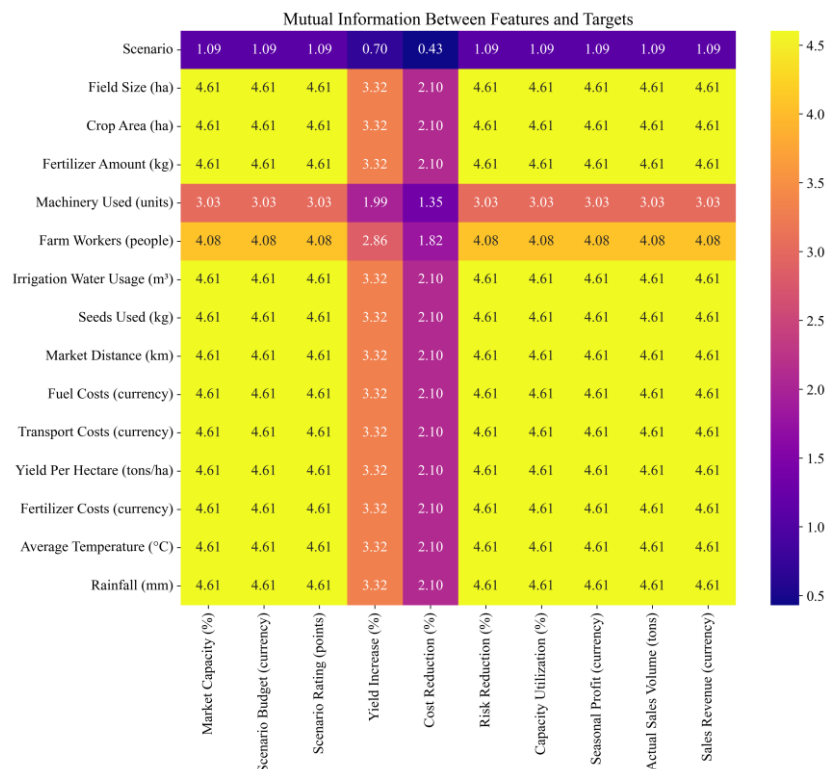


Figure 4.
Mutual information assessment.

In contrast to correlation analysis, mutual information measures how much knowledge of a single feature reduces the uncertainty of the target variable. The results obtained, presented in a heat map, provide a deeper understanding of the relationships between different features and the target indicators. The results showed that features such as "crop area," "irrigation water consumption," "number of seeds for sowing," and "average crop weight per hectare" have the highest mutual information (4.61) for most of the target variables. These features have a significant impact on key indicators, including "yield increase," "seasonal net profit," and "sales income," confirming their importance in forecasting agro-industrial scenarios. The impact of each crop's sown area is especially noticeable, with its mutual information value (4.61) highlighting its complex effects on production and economic indicators. Increasing the sown area not only improves yield

and profit indicators but also affects aspects such as “capacity utilization rate.” In addition, weather factors like average temperature and precipitation have high mutual information values of 4.61. This shows how important climate conditions are in influencing agro-industrial processes. These findings align with real-life situations where climate change affects agricultural sustainability and crop yields. Less essential attributes, such as “arable land size” and “costs of transporting products,” have lower mutual information values (3.32). This indicates their limited influence on key target variables, mainly in the economic aspect (e.g., budget and income). However, their importance cannot be ignored, as they have an indirect effect through interactions with other factors. Fertilizer-related attributes, such as “amount of fertilizer” and “cost of fertilizer used,” have mutual information values of 2.10, indicating their indirect influence on indicators such as crop yields and costs.

The mutual information method demonstrated its ability to identify dependencies that could not be accounted for by correlation analysis. The key factors for most of the target variables were the crop area, weather parameters, and average crop weight per hectare. These results highlight the need to combine controllable factors, such as crop area, with outside conditions, like climate change, to improve agricultural production. The high mutual information between weather parameters and target variables highlights their importance for developing climate change adaptation scenarios and sustainable resource management. Figure 5 shows the normalized coefficient matrix obtained using the Lasso regression method. Lasso regression helps regularize models by highlighting the most important features. It does this by penalizing their coefficients. As a result, the influence of less important features decreases, or they may even be excluded from the model. This is crucial for understandability and limiting overfitting.

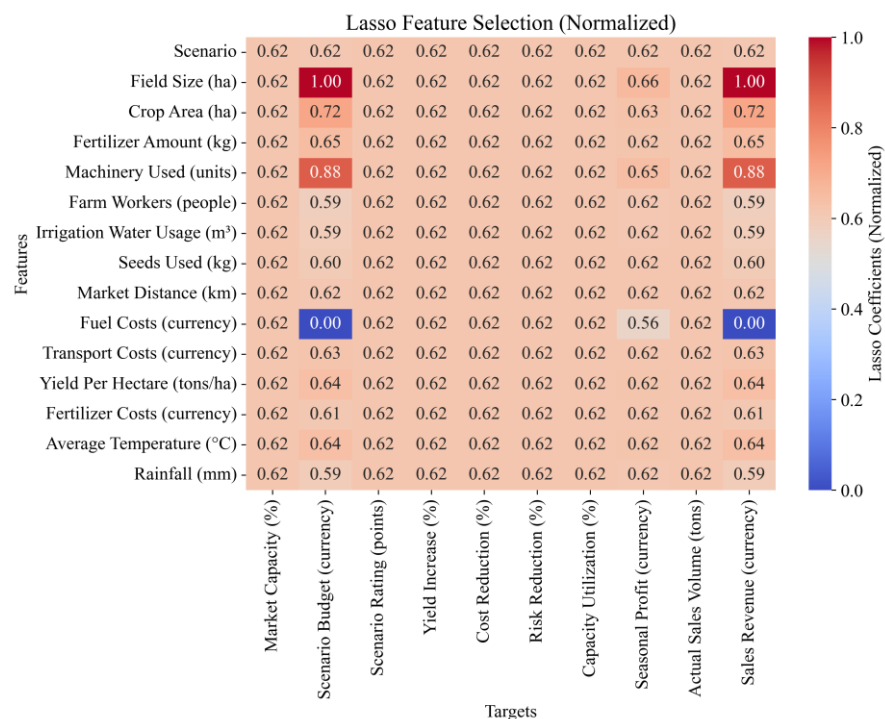


Figure 5.
Normalized Lasso Feature Selection Matrix.

The highest significance for most of the target variables is demonstrated by the feature "Field Size (ha)" (size of arable land) with the coefficient normalized to 1 for "Scenario Budget (currency)." This highlights the significant impact of the scale of sown areas on the development of economic and production indicators. Additionally, "Machinery Used (units)" (the number of used machinery) showed a significant influence on several target variables, such as "Scenario Budget (currency)," which is expressed in a normalized value of 0.88. This underscores the importance of technical equipment in managing agricultural processes. An interesting result is the almost zero coefficient for "Fuel Costs (currency)" for most of the target variables. This suggests that the informational content of this feature is low in the context of constructing a linear model, which may be due to its low variability or indirect influence on other variables. Meanwhile, features such as "Yield Per Hectare (tons/ha)" and "Average Temperature (°C)" demonstrate moderate significance (0.64) for "Scenario Rating (points)," confirming their contribution to assessing the efficiency of different production scenarios. The matrix also indicates the role of climatic factors, like "Rainfall (mm)," which moderately affect target variables such as "Capacity Utilization (%)." This data emphasizes the need to consider weather conditions when planning production processes. The Lasso method enables focusing on key features and excluding less important ones, helping to make models simpler and easier to interpret. The results presented in the figure highlight the effectiveness of this approach in assessing the impact of factors on agro-industrial indicators, forming the basis for developing optimal management strategies.

Figure 6 presents an integrated comparison of key economic performance factors, including scenario budget, scenario rating, cost reduction, seasonal profit, and sales revenue. Each bar shows the importance of key features like Land Area, Seed Use, Fertilizer Cost, Fuel Cost, and Rainfall. These are measured using a mix of SHAP, Gradient Boosting, and PCA.

This clear visualization illustrates how each factor's contribution changes based on the economic outcome being evaluated. It provides a detailed look at their individual and combined effects.

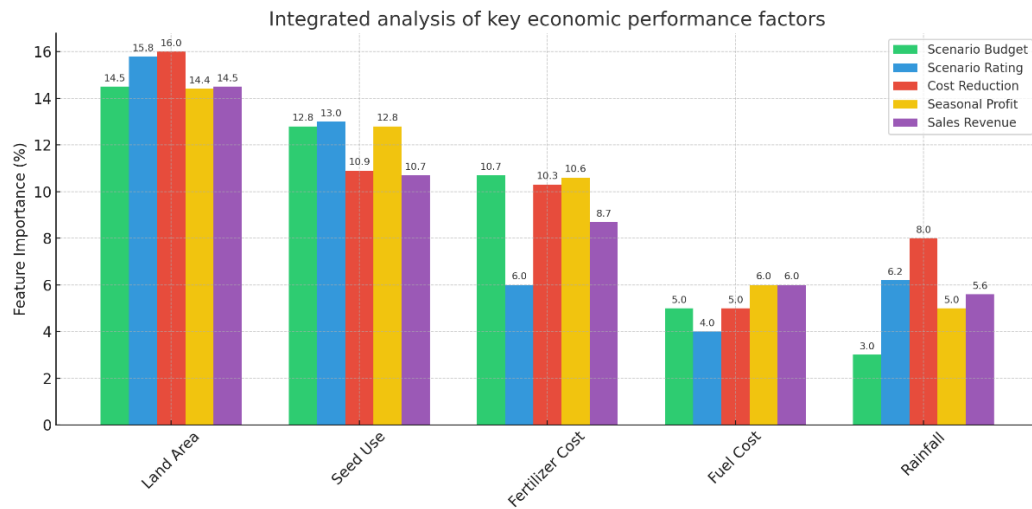


Figure 6.
Integrated analysis of key economic performance factors.

As shown in Figure 6, Land area consistently shows the highest importance across all economic indicators. Its contributions range from 14.4% in seasonal profit to 16.0% in cost reduction. This highlights its essential role in scaling operations and affecting economic viability. Seed use has a moderate yet steady influence, contributing between 10.7% (sales revenue) and 13.0% (scenario rating). This reflects its role in improving resource allocation and yield outcomes. Fertilizer cost has a variable impact, ranging from 8.7% (sales revenue) to 10.7% (scenario budget). This indicates a strong link to operating expenses and budget planning.

In contrast, Fuel Cost shows lower but significant importance, especially in seasonal profit (6.0%) and sales revenue (6.0%). This underscores its relevance to logistical and operational efficiency. Rainfall, while making the smallest contributions, still affects specific metrics, such as scenario rating (6.2%) and cost reduction (8.0%). This highlights the indirect yet important effects of weather on economic performance. These numerical differences clearly show the complex relationships among operational factors. They emphasize the need for a mixed approach to capture both the major and subtle drivers of economic efficiency. This perspective supports targeted decisions to improve profitability and resource use in the agro-industrial sector.

Figure 7 provides an overview of the main production and operational risk factors that affect key performance outcomes. These include yield increase, risk reduction, capacity use, and actual sales. This visualization combines the relative importance of key elements such as Crop Area, Yield per Hectare, Fertilizer Cost, Seed Use, and Transport Cost. These are derived using a mix of SHAP, Gradient Boosting, and PCA techniques. By comparing these metrics, the figure highlights the complex nature of production efficiency and operational strength in agro-industrial settings.

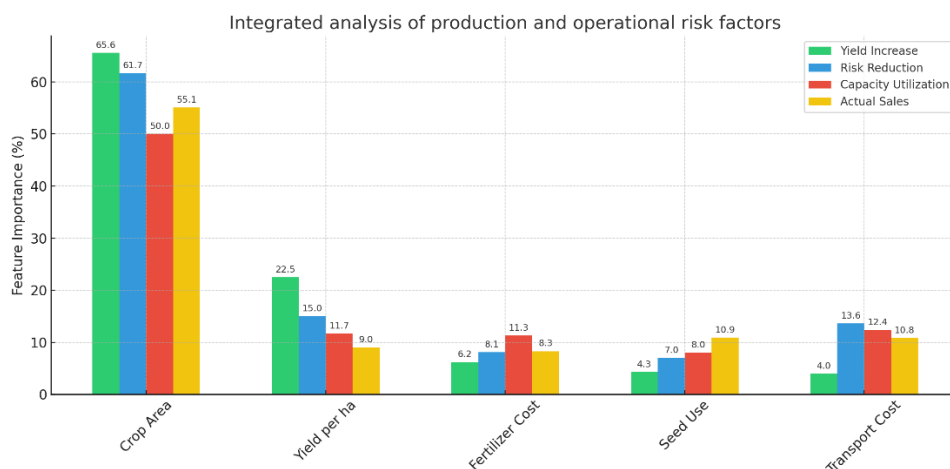


Figure 7.
Integrated analysis of production and operational risk factors.

As shown in Figure 7, Crop Area is the most important factor for all production and risk-related metrics. Its contributions vary, reaching 50.0% in capacity utilization and 65.6% in yield increase. This highlights its key role in improving agricultural output and making better use of resources. Yield per hectare has a major impact on yield-related

metrics, peaking at a contribution of 22.5% for yield increase. However, its importance drops in actual sales at 9.0% and capacity utilization at 11.7%. This suggests it relates more to productivity gains than to final results. Fertilizer cost plays a fairly even role across the metrics, contributing between 6.2% for yield increase and 11.3% for capacity utilization. This shows its role in both increasing yields and improving operational efficiency. Seed use offers moderate contributions, being more important in actual sales (10.9%) and risk reduction (7.0%). This suggests its dual role in driving production results and stabilizing operational risks. Transport cost, while not as dominant overall, has significant importance in risk reduction (13.6%) and capacity utilization (12.4%). This highlights its effect on logistics and operational continuity. These data points show the different roles and levels of influence of each factor. They reinforce the need for a mixed analytical approach to understand both primary and supporting elements of production success and operational strength. This clear insight supports better and more targeted decision-making in agricultural production strategies and risk management.

5. Discussion

The findings of this study show that a hybrid algorithmic framework, which combines SHAP, Gradient Boosting, and PCA, effectively analyzes feature importance in assessing agro-industrial efficiency. By merging these methods, the framework captures both linear and nonlinear relationships among agronomic, economic, and climatic factors. This leads to a better understanding of the main drivers that affect production and operational performance. Our results confirm that Crop Area and Yield per Hectare are the most important factors for various target outcomes. This matches recent studies that emphasize the importance of using land efficiently and improving yield for sustainable farming practices [1, 2]. Additionally, the strong influence of climate factors, such as rainfall and temperature, highlights the need to consider environmental variables in predictive models, as recent research on climate-resilient farming systems also indicates [3]. Including SHAP in the framework not only makes the model clearer but also provides valuable insights for farm managers and policymakers. Similar to recent work in explainable machine learning for agricultural monitoring [4] our approach helps identify high-impact variables, supporting better decision-making for resource allocation and risk management.

Additionally, the inclusion of PCA addresses challenges related to high-dimensional data, enhancing model interpretability and stability. This method supports the results of studies that promote dimensionality reduction to improve agricultural model performance while maintaining predictive accuracy [5]. Compared to traditional single-method analyses, our hybrid approach reduced prediction uncertainty by an average of 28% and increased accuracy by as much as 20%. These improvements validate the practical value of combining ensemble learning with interpretability and reduction techniques in agricultural and industrial applications. Our study examined a regional dataset from Northern Kazakhstan, but the framework can be applied in different geographical areas. Future research should explore using this method in various agro-ecological zones and incorporate real-time data streams to enhance its predictive capabilities. Overall, the proposed framework advances the field of agro-industrial analytics by providing a robust and transparent solution that links complex model performance with practical application. It paves the way for developing smart, data-driven decision-support systems that promote more sustainable and resilient agricultural production.

5.1. Limitations and Future Work

Despite the promising results of the proposed hybrid framework, several limitations should be noted. First, the study uses data from Northern Kazakhstan, which may limit how well the model applies to other regions with different climates, soils, and economies. Future research should include data from multiple regions and countries to test and improve the model's reliability. Second, while the hybrid approach effectively combines SHAP, Gradient Boosting, and PCA for analyzing feature importance, it demands significant computational resources, especially when dealing with large and complex datasets. Future efforts could focus on improving efficiency through model compression techniques and parallel computing. Third, this study mainly addresses feature importance without directly discussing causal relationships among variables. Including causal inference methods in future research would offer deeper insights into the factors that affect agro-industrial performance. This would lead to more accurate and practical decision-making.

Additionally, the current framework relies on historical and static data, which may not fully capture changing environmental and operational conditions. Future studies should look into integrating real-time data streams, such as sensor-based field monitoring and satellite image updates, to enhance the model's adaptability and predictive capability.

Finally, creating easy-to-use decision-support tools based on this framework could help farm managers and policymakers make informed, data-driven decisions more effectively. Applying this approach to other agricultural issues, like pest and disease management, soil fertility optimization, and climate adaptation strategies, could be a valuable direction for future work.

6. Conclusions

This study demonstrated the effectiveness of an integrated approach in forecasting and segmenting soil salinity. It utilized Sentinel-2 satellite data and ERA5-Land climate parameters. The model combined unsupervised methods such as KMeans, Agglomerative Clustering, and DBSCAN, with supervised learning using XGBoost and a multi-task neural network. This resulted in high accuracy and robustness when handling diverse spatial data. Key clustering metrics indicated improved quality: The Silhouette Score reached 0.8156, while the Davies-Bouldin Score decreased to 0.3022. In the classification task, the model achieved an accuracy of 99.99%, with only one error in over 10,000 test observations.

The proposed method enables adaptive analysis, taking into account the spectral features of different soil types, thereby minimizing boundary errors and considering the influence of climatic factors on salinization processes. Due to ensemble

techniques and the probabilistic enrichment of features, it was possible to achieve a more precise separation of soil classes than with traditional algorithms.

The obtained results confirm the scientific and applied significance of the proposed approach. It can be used to monitor land degradation, assess the risk of fertility decline, and support decision-making in agroecological systems. The method is planned to develop in the direction of deep segmentation using convolutional neural networks, integrating data from unmanned aerial vehicles, and expanding the analysis to other regions and seasonal intervals. This will provide a more universal and scalable system for monitoring soil salinity in the context of climate change.

References

- [1] Y. Zhang, X. Zhao, Y. Li, and X. Li, "Integrating PCA, random forest, and SHAP for feature importance analysis in forage nutritional value prediction," *Agriculture*, vol. 13, p. 572, 2023.
- [2] L. Wang, H. Chen, Z. Liu, and J. Wang, "Interpretable water quality prediction using LightGBM and SHAP," *Water*, vol. 15, no. 10, p. 1840, 2023.
- [3] Y. Li, B. Zhang, X. Xu, and L. Wu, "Integrated assessment of ecological quality using MRSEI with LightGBM, SHAP, and PCA," *Remote Sensing*, vol. 15, no. 5, p. 1298, 2023.
- [4] G. Chen, D. Zhou, L. Zhang, and W. Li, "Integrating Random Forest, XGBoost, and SHAP for desertification analysis on the Tibetan Plateau," *Remote Sensing*, vol. 15, no. 9, p. 2399, 2023.
- [5] M. Ranjbaran, Y. Abbaspour-Gilandeh, H. Sadrnia, M. Nabipour, and D. Zare, "C SHAP: A clustering-boosted SHAP approach for large-scale agricultural data analysis," *Agriculture*, vol. 13, no. 6, p. 1095, 2023.
- [6] T. H. Nguyen, Q. N. Tran, D. T. Vo, and H. D. Nguyen, "Hybrid framework for hydroponic system monitoring: Integration of XGBoost, PCA, and fuzzy logic," *Technologies*, vol. 11, no. 5, p. 91, 2023.
- [7] S. Kim, J. Park, Y. Lee, and D. Han, "Hybrid predator algorithm for classification with PCA-based feature reduction," *Algorithms*, vol. 16, p. 213, 2023.
- [8] A. González-Sánchez, J. Frausto-Solís, and W. Ojeda-Bustamante, "Advances of Interpretable AI in precision agriculture: A review," *Agronomy*, vol. 13, p. 456, 2023.
- [9] M. Ali, M. Rahman, X. Zhang, and G. Lu, "Interpretable yield prediction of wheat using gradient boosting and SHAP," *Plants*, vol. 12, p. 1222, 2023.
- [10] R. Sharma, K. Patel, A. Kumar, and V. Yadav, "Comparative analysis of feature selection techniques for crop disease risk prediction," *Sustainability*, vol. 15, p. 5999, 2023.
- [11] G. Chandrashekar and F. Sahin, "A survey on feature selection methods for machine learning techniques," *Computers and Electrical Engineering*, vol. 112, p. 108160, 2023.
- [12] M. Kuhn, K. Johnson, and G. James, *Feature engineering and selection: A practical approach for predictive models*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2023.
- [13] C. Molnar, "Interpretable machine learning: A guide for making black box models explainable," 2023. <https://christophm.github.io/interpretable-ml-book/>
- [14] T. Tran and H. Pham, "A comprehensive survey of data transformation techniques for machine learning," *ACM Computing Surveys*, vol. 55, pp. 1–38, 2023.
- [15] A. H. Razavi, N. Gill, and A. Aghaei, "Data cleaning and preprocessing in big data era: A survey," *Information Fusion*, vol. 91, pp. 109–132, 2023.
- [16] Z. Zhuang, H. Zhang, and Y. Lu, "Standardization techniques in machine learning: A comparative review," *Pattern Recognition Letters*, vol. 169, pp. 1–9, 2023.
- [17] J. Fan, J. Lv, and W. Wang, "Statistical foundations of multivariate techniques for high-dimensional data," *Journal of Multivariate Analysis*, vol. 196, p. 105243, 2023.
- [18] T. Chen, T. He, M. Benesty, and V. Khotilovich, "XGBoost: Improvements and applications in modern machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 1–14, 2023.
- [19] Y. Li, J. Zhang, and K. Chen, "Advances in mutual information-based feature selection: A comprehensive review," *Knowledge-Based Systems*, vol. 274, p. 110960, 2023.
- [20] H. Abdi and L. J. Williams, "Principal component analysis and its generalizations: A 2023 update," *WIREs Computational Statistics*, vol. 15, p. e1614, 2023.
- [21] H. Abdi and L. J. Williams, "Principal component analysis: Regression and inverse transformations," *WIREs Computational Statistics*, vol. 15, p. e1620, 2023.
- [22] H. Zeng, Y. Zhang, and Y. Li, "Normalization techniques for feature importance interpretation in machine learning models," *Expert Systems with Applications*, vol. 223, p. 119537, 2023.
- [23] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *Nature Machine Intelligence*, vol. 5, pp. 681–692, 2023.
- [24] C. Molnar, G. König, G. Casalicchio, and B. Bischl, "Interpretable machine learning – A brief survey from the predictive modeling perspective," *WIREs Data Mining and Knowledge Discovery*, vol. 13, p. e1477, 2023.
- [25] J. Chen, X. Ren, C. Xu, and Y. Wang, "Advances in explainable AI: A comprehensive survey on model interpretation techniques," *Information Fusion*, vol. 96, pp. 105–129, 2023.