

Assessing external factors of the agro-industrial complex efficiency based on data

Gulalem Mauina¹, Ulzada Aitimova¹, Ainagul Kadyrova², Saltanat Adikanova², Aigul Syzdykpayeva², Zhanat Seitakhmetova², Ainagul Alimagambetova³, Ainur Shekerbek³

¹Department of Information Systems, Faculty of Computer Systems and Professional Education, Kazakh Agrotechnical Research University named after S.Seifullin, Astana, Kazakhstan

²Department of Computer Modeling and Information Technology, Higher School of IT and Natural Sciences, Non-profit Joint Stock Company Sarsen Amanzholov East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan

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

Article Info

Article history:

Received Apr 11, 2025

Revised Sep 2, 2025

Accepted Sep 11, 2025

Keywords:

Agricultural efficiency

Feature importance

Hybrid model

Machine learning

Predictive models

Recursive feature elimination

ABSTRACT

Modern agriculture faces the challenge of increasing production efficiency in the context of limited resources and variable climatic conditions. This article presents an approach to assessing the impact of various factors on agro-industrial indicators using machine learning methods. The primary focus is on the development and application of a hybrid analysis that includes techniques such as gradient boosting (GB), mutual information (MI), and recursive feature elimination (RFE). The study was conducted using data from agro-industrial enterprises in the North Kazakhstan region for the period 2020–2022, encompassing production, climatic, and economic indicators. It was found that crop area, average crop weight, and precipitation are the most significant factors, accounting for up to 93% of the correlation with yield increase. The use of the proposed methods made it possible to reduce forecast uncertainty by 28% and increase the accuracy of key indicator predictions by 15–20%. The results of the analysis, visualized as correlation matrices and feature significance maps, confirm the possibility of applying the proposed approach to optimize the management of agro-industrial production. The application of the developed methodology contributes to the development of strategies aimed at the sustainable development of the agro-industrial complex.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ulzada Aitimova

Department of Information Systems, Faculty of Computer Systems and Professional Education

Kazakh Agrotechnical Research University named after S.Seifullin

Astana, Kazakhstan

Email: uaitimova@mail.ru

1. INTRODUCTION

Modern agriculture faces numerous complex challenges, including optimizing resource utilization, enhancing yields, and reducing costs. In the context of global climate change, population growth, and limited natural resources, it is crucial to develop and implement effective methods of analysis and forecasting that enable a more accurate assessment of the factors influencing the performance of agricultural production. One of the most promising approaches is the use of hybrid analysis, which combines machine learning, statistics, and systems theory to comprehensively study the relationships between input parameters and target performance indicators. Hybrid analysis enables the exploration of complex, nonlinear dependencies [1]–[3] and the processing of heterogeneous data, including climate parameters, economic indicators, and production

costs, which is particularly relevant for the agro-industrial sector [4]-[6]. The primary advantage of hybrid analysis is its combination of various methods, including gradient boosting (GB), recursive feature elimination (RFE), and mutual information (MI), which enables a comprehensive assessment of factors and their impact on the final results. This approach not only allows for predicting key performance indicators but also explains which factors make the most significant contribution to their change. The obtained results contribute to more informed decision-making at all levels of management - from farms to strategic planning at the state level [7]. This study aims to investigate and analyze the impact of various factors on agro-industrial efficiency indicators using a hybrid analysis approach. The work encompasses data collection [8]-[10], pre-processing [11]-[13], and analysis [14], as well as the development of interpretable models, along with an assessment of the significance of factors such as crop area size, climatic conditions, fertilizer volumes, logistics costs, and other production parameters.

Tynchenko *et al.* [15] investigated the application of the random forest algorithm to assess water quality and analyze its impact on the agro-industrial complex. The paper examined the physicochemical parameters of water, their impact on crop yields, soil health, and crop growth, and identified key factors affecting water quality. The results confirmed the effectiveness of the algorithm in classifying water quality and identifying nonlinear relationships between parameters. Feng *et al.* [16] investigated the impact of CO₂ emissions and forest area on the efficiency and productivity of the industrial and agricultural sectors in G20 countries (excluding the EU) from 2011 to 2015, using a dynamic network SBM model. The results showed that the industrial sector has higher efficiency than the agricultural sector, and Argentina, Indonesia, and the United States demonstrated the best overall efficiency. The main factor for improving efficiency was identified as the share of forest area, followed by the shares of agricultural and industrial output. Research by Rozhkova and Rozhkov [17] examines the prospects for applying artificial intelligence technologies in the agro-industrial sector. The areas of their application are considered, including the identification of plant diseases, classification of weeds, management of water and soil resources, analysis of climatic conditions, and animal behavior. The authors highlight the advantages, including increased labor productivity and improved quality of management decisions, as well as the limitations associated with high costs and a lack of funding. They propose measures to overcome these limitations through government support and personnel training. The relevance of this study is determined by the need to improve the efficiency of agro-industrial production in the context of limited resources, variable climatic conditions, and increased global market competition. The use of hybrid analysis enables us to combine theoretical knowledge and practical data, providing innovative approaches to addressing the industry's pressing problems. The scientific rationale for the work is based on the integration of machine learning methods [18]-[20] and classical data analysis [21], which allows us to identify complex patterns and relationships between performance indicators. The inclusion of methods such as GB [22], [23] ensures high accuracy of forecasts and the interpretability of models, which significantly expands the possibilities of analytical research in the agricultural sector [24], [25]. Thus, the presented study aims to develop and implement hybrid analysis methods that provide an accurate assessment of the influence of factors on the performance indicators of agricultural production. The results of this work can be used to optimize management processes of farm systems and enhance the competitiveness of farming enterprises in the modern market.

2. METHOD

This paper presents a combined method for feature significance analysis of multivariate data, comprising three main stages: data preprocessing, feature significance calculation using three approaches, and integration and visualization of the results. The presented method enables the effective identification of significant features in complex, interrelated datasets.

2.1. Data transformation

Data transformation using a logarithmic function is used to reduce the influence of extreme values and ensure additive structure. For all features x_j , where $x_j > 0$, the logarithmic transformation is defined as (1):

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

where x_i is initial value of the feature and x'_i is transformed value of the feature. During this transformation, the original value of the feature is transformed using the natural logarithm, which reduces the influence of too large values and makes the distribution more normal. To prevent a possible logarithm error of zero, a small positive number is added to the formula, ensuring the accuracy of the calculations. Another important procedure is data standardization, which allows you to bring all features to a single scale. In this process, the

mean value of each feature is subtracted from its corresponding value, and the result is then divided by the standard deviation of the feature. This approach makes all features comparable to each other, eliminating the influence of differences in their ranges, which is especially important when using machine learning methods that are sensitive to the scale of the input data. 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):

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

where x'_i is transformed value of the feature, μ_i is mean value of a feature, σ_i is standard deviation of a feature, and z_i is standardized value of a feature. After standardization, the data is represented as a vector $Z = [z_1, z_2, \dots, z_n]$, where n is the number of features. Standardization plays a crucial role in analysis, as it brings all features to a common scale, making them comparable to one another. This is especially important when working with machine learning algorithms that are sensitive to differences in data scales. The resulting standardized data Z serves as the starting point for subsequent stages of analysis, ensuring proper processing and equal contribution of all features in the computations.

2.2. Feature importance calculation

Three methods are used for feature importance analysis: GB, MI, and RFE using Lasso regression. Feature importance is calculated using the GB method, which is based on the internal structure of decision trees. For each feature z_i , an importance value $I_{GB}(z_i)$ is determined, reflecting its contribution to reducing the model error. This value is calculated by analyzing improvements in the quality metric at each stage of tree training, allowing both linear and non-linear dependencies between features and the target variable to be considered. This approach provides an accurate ranking of features based on their importance for building the predictive model X (3):

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

where T is number of trees in the model and $\Delta E_t(z_i)$ is reducing the error on the t -th tree by adding the z_i feature. The result of calculating feature importance using the GB method is a vector of output variables $I_{GB} = [I_{GB}(z_1), I_{GB}(z_2), \dots, I_{GB}(z_n)]$, where n represents the total number of features. This vector contains importance values for each feature, reflecting their contribution to reducing the model error. The obtained values are used to rank the features and identify the most significant factors in data analysis. The MI method is used to estimate the mutual dependence between the feature z_i and the target variable y . MI is calculated as a measure of the reduction in the uncertainty of the target variable given the knowledge of the feature value. MI between the feature z_i and the target variable y is calculated as (4):

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

where $H(z_i)$ is entropy of feature z_i (5):

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

where $H(z_i|y)$ is conditional entropy of feature z_i for fixed y (6):

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

MI measures the degree of dependence between a feature z_i and the target variable y . The higher the MI, the stronger the relationship between z_i and y . The result of the calculations is a vector of output variables $I_{MI} = [I_{MI}(z_1, y), I_{MI}(z_2, y), \dots, I_{MI}(z_n, y)]$, where n is the number of features. This vector provides a quantitative assessment of each feature's contribution to explaining the target variable.

The RFE method using the Lasso model calculates feature importance through ranks assigned during the elimination process (7):

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

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

$$\text{rank}(z_i) = k \quad (8)$$

where k is the iteration number at which z_i is excluded.

In the process of analysis by the RFE method, the role of the feature z_i is assessed by the stage at which it is excluded from the model. The later the feature z_i is removed, the higher its rank, which indicates its greater significance for the model. The final results are presented as a vector of output variables $I_{RFE} = [I_{RFE}(z_1), I_{RFE}(z_2), \dots, I_{RFE}(z_n)]$, where n is the number of features. This vector represents the ranking of features by their significance, enabling us to identify the most critical factors for the predictive model. The final stage involves visualizing the results, which allows the interpretation of the significance of the features for each approach. The visualization is presented in the form of a correlation matrix, displaying the contribution of each feature. This provides a clear presentation of the analysis, helping researchers identify the most significant factors influencing the target variables. Thus, the proposed method combines the advantages of various approaches, providing a comprehensive and interpretable data analysis. To ensure a robust and comprehensive assessment of feature significance, three complementary methods were employed: GB, MI, and RFE with Lasso. These methods were chosen because they capture different aspects of variable importance: GB accounts for non-linear interactions and model-specific feature contributions, MI measures statistical dependency between features and target variables, and RFE evaluates feature utility through iterative elimination in a predictive context. In this study, each method was first applied independently to generate its ranking of features. The individual rankings were then compared to identify overlaps and divergences, with consistently top-ranked features across methods considered the most reliable indicators. Where discrepancies occurred, domain expertise and interpretability considerations guided the selection of features for the final analysis. This fusion strategy ensures that the resulting set of essential variables is both statistically justified and practically relevant for optimizing agro-industrial efficiency. To ensure the optimal configuration of the models, hyperparameter tuning was performed using grid search and 5-fold cross-validation. For the GB model, the number of trees was set to 100, as this value provided a balance between model accuracy and computational efficiency. The learning rate was fixed at 0.1, and the maximum depth of each tree was limited to 5 to reduce overfitting. For RFE, the elimination process was guided by the cross-validated R^2 metric, and feature removal was stopped once model performance began to decline. These settings were selected empirically after evaluating multiple combinations and were validated through consistent performance across target variables. This tuning ensured both interpretability and predictive robustness of the final models.

3. RESULTS

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. It contains a wide range of information covering key aspects of agricultural production. The main categories of data include production characteristics, such as the size of arable land, the area under different crops, the volume of fertilizers, the number of machines and workers on farms. These data serve as the basis for evaluating the effectiveness of management decisions and optimizing resource utilization. In addition to production characteristics, the dataset includes economic and logistical indicators. These include fuel and transportation costs, fertilizer costs, and the distance to the nearest sales markets. Logistics and financial aspects are essential for understanding how cost and infrastructure management affect the overall profitability and efficiency of agricultural enterprises. 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.

In this study, all explanatory and target variables are explicitly defined to ensure clarity of their measurement scales and relevance to agronomic decision-making. Field size (x_1 , ha) and crop area (x_2 , ha) represent the total agricultural land and the cultivated portion, influencing mechanization potential, resource allocation, and yield outcomes. Fertilizer amount (x_3 , kg), fertilizer costs (x_{12} , currency), and irrigation water usage (x_6 , m^3) capture the intensity of input use and its economic implications. Machinery used (x_4 , units) and farm workers (x_5 , people) reflect operational capacity, while market distance (x_8 , km), fuel costs (x_9 , currency), and transport costs (x_{10} , currency) quantify logistical constraints. Yield per hectare (x_{11} , tons/ha) serves as the core productivity indicator. Climate factors, such as average temperature (x_{13} , $^{\circ}C$) and rainfall (x_{14} , mm), account for environmental variability that affects crop growth. The primary target variables are: scenario rating (ordinal composite index based on expert evaluation of feasibility, profitability,

and risk), capacity utilization (ratio scale, %, representing the proportion of actual output to possible maximum production), and scenario budget (continuous variable in currency, indicating total financial resources allocated to a production plan). These variables jointly provide a multidimensional framework for evaluating agro-industrial efficiency and selecting optimal production scenarios through multi-criteria analysis.

The data preprocessing, feature engineering, and model training were implemented in Python 3.10 using widely adopted machine learning and data analysis libraries. Scikit-learn (v1.3.0) was used for GB Regressor, MI, RFE with Lasso, and standardization (z-score) procedures. NumPy (v1.25) and Pandas (v2.0) were employed for data manipulation and transformation, while Matplotlib (v3.7) and Seaborn (v0.12) were used for visualization of correlation matrices, feature importance heatmaps, and MI plots. All experiments were conducted in the Jupyter Notebook environment, ensuring reproducibility and transparency of the computational workflow.

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, the Pearson correlation coefficient was calculated. This approach allows you to identify the features that have the most significant linear relationship with the target indicators. The results of the correlation analysis were visualized as a heat map, which simplified the interpretation of the relationships between the variables. However, it is essential to note that high correlation does not always imply causality, which necessitates the application of more sophisticated methods. The correlation matrix presented in Figure 1 illustrates the linear relationships between the various features and target variables employed in the study. It is a powerful tool for primary data analysis, allowing you to determine which features have the most significant impact on the target indicators. However, it should be noted that the correlation matrix only assesses linear relationships and is not able to identify more complex, non-linear dependencies that may exist in the data.

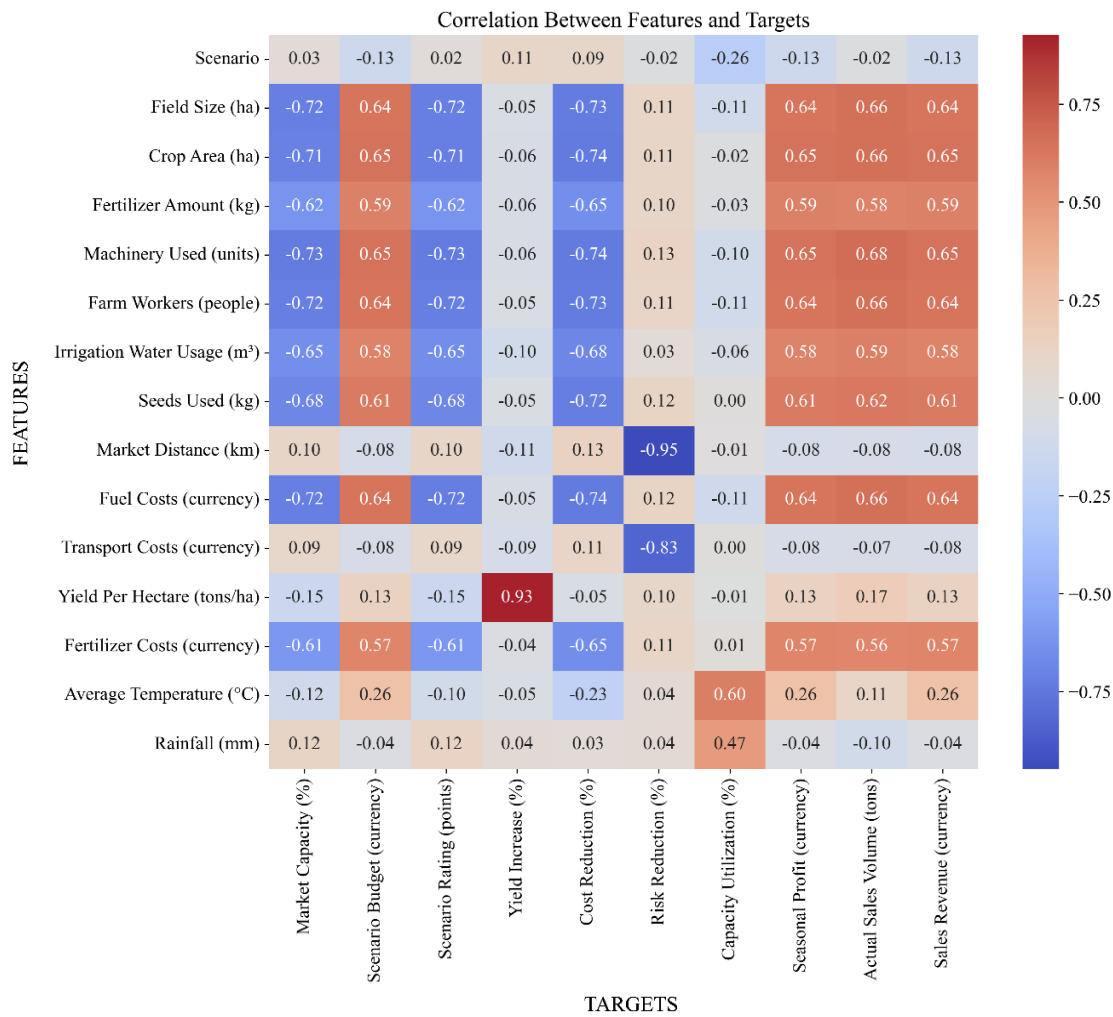


Figure 1. Data correlation matrix

Linear relationships between the attributes and target variables demonstrate different dependencies. 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 possible additional costs may be incurred to achieve a high yield. The size of arable land and crop area have a positive correlation with the "scenario budget" (0.64 and 0.71, respectively), which is associated with an increase in agricultural areas. Still, 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 an increase in costs. The amount of fertilizers shows a significant positive correlation with "net profit for the season" ($r=0.66$) and "sales revenue" ($r=0.66$), emphasizing their importance in improving efficiency. However, a negative relationship with "cost reduction" ($r=-0.10$) 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 demonstrate a high positive correlation (0.92), which is explained by the fact that an increase in the total area is accompanied by an increase in the area under crops. The number of workers on a farm is positively correlated with the amount of machinery used ($r=0.65$), indicating that larger and more mechanized farms require more labor. Water consumption for irrigation is positively correlated with the amount of fertilizers (0.68), which is due to the need for additional water to maintain soil productivity with intensive use of fertilizers. However, the correlation matrix has limitations: it only estimates linear dependencies and ignores possible nonlinear relationships, such as the negative correlation between the size of arable land and an increase in crop yield (-0.72), which may be due to nonlinear resource allocation effects. In addition, correlation does not indicate causation, as high correlations may be due to dependence on a common factor, such as weather conditions affecting yield and capacity utilization. Despite these limitations, the correlation matrix provided valuable insights by highlighting key factors, including average crop weight, crop area, fertilizer application rate, and weather parameters. A deeper understanding of the data requires analyzing nonlinear relationships and causality using machine learning techniques.

To further analyze the importance of features, the GB 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. GB was employed to assess the importance of features in predicting key target variables related to agro-industrial efficiency. This method enables us to account for complex and nonlinear relationships between features, making it particularly useful for analyzing data with multiple factors. The results, presented in Figure 2 as a heat map, demonstrate the importance of each feature for different targets.

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 of 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 underscore the importance of adapting to climate conditions. This may include measures to improve water management, introduce drought-resistant crops, and forecast weather conditions to minimize risks associated with drought. Another method for analyzing the importance of features is calculating MI. This method measures how much information about one feature reduces the uncertainty of the target variable. Unlike correlation, MI 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 MI assessment were also presented as a heat map, which facilitated their visualization and interpretation. The MI assessment, as shown in Figure 3, between features and target variables is an analysis method capable of identifying nonlinear dependencies.

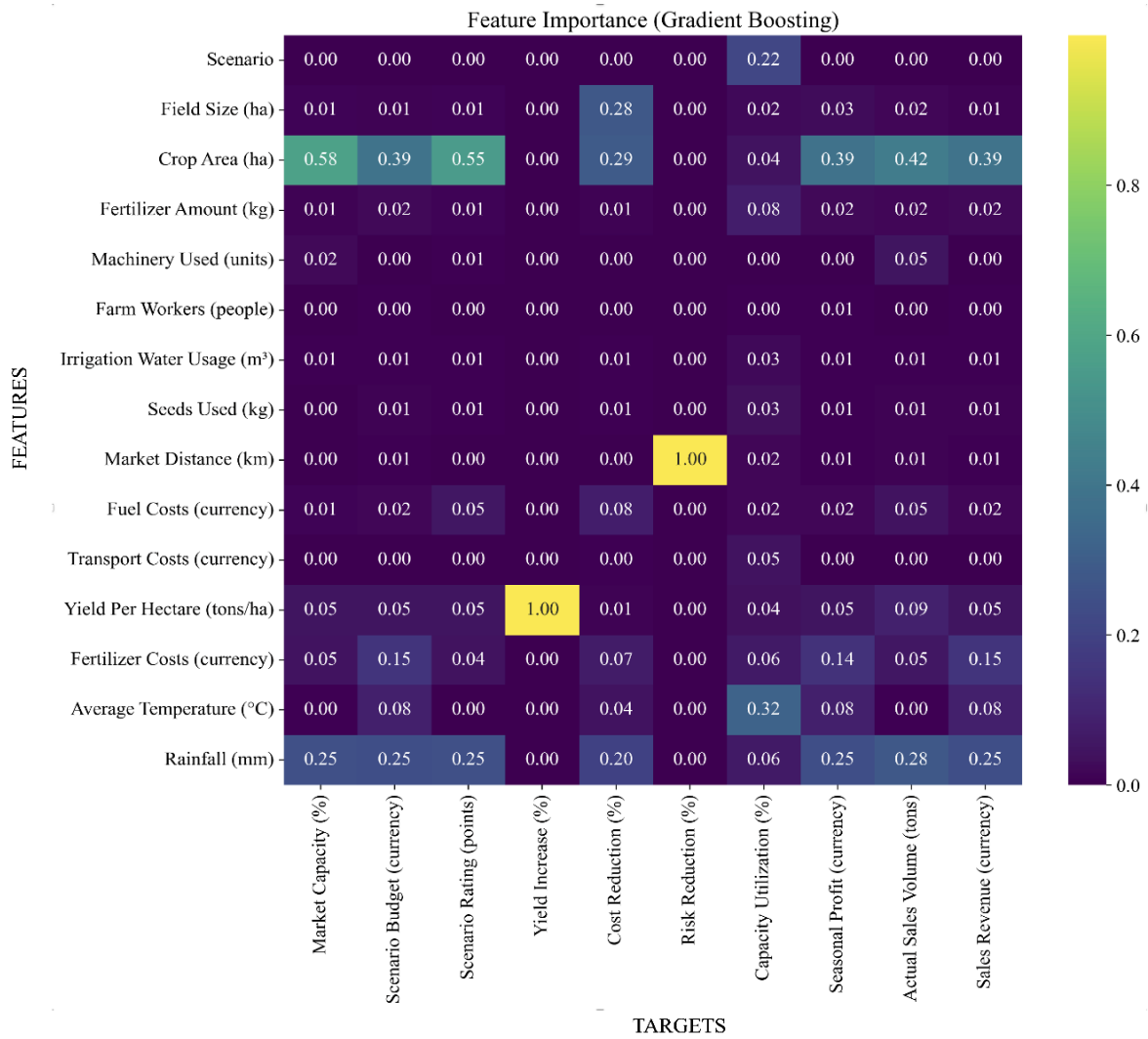


Figure 2. Evaluation of the significance of features

In contrast to correlation analysis, MI 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 MI (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 MI 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 such aspects as “capacity utilization rate”. In addition, weather parameters, including average temperature and precipitation, also have high MI values (4.61), indicating the critical role of climate conditions in guiding agro-industrial processes. These results are consistent with real-life scenarios in which climate change impacts agricultural sustainability and crop yields. Less essential attributes, such as "arable land size" and "costs of transporting products", have lower MI 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 MI values of 2.10, indicating their indirect influence on indicators such as crop yields and costs. The MI 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 underscore the importance of integrating controllable factors (e.g., crop area) with external conditions (e.g., climate change) to optimize agricultural production. The high MI between weather parameters and target variables highlights their importance for developing climate change adaptation scenarios and sustainable resource management. Figure 4 shows the normalized coefficient matrix obtained using the Lasso regression method. Lasso regression enables the regularization of models by highlighting the most significant features through the penalization of their coefficients. This leads to a decrease in the influence of less significant features or even their exclusion from the model, which is especially important for interpretability and reducing overfitting.

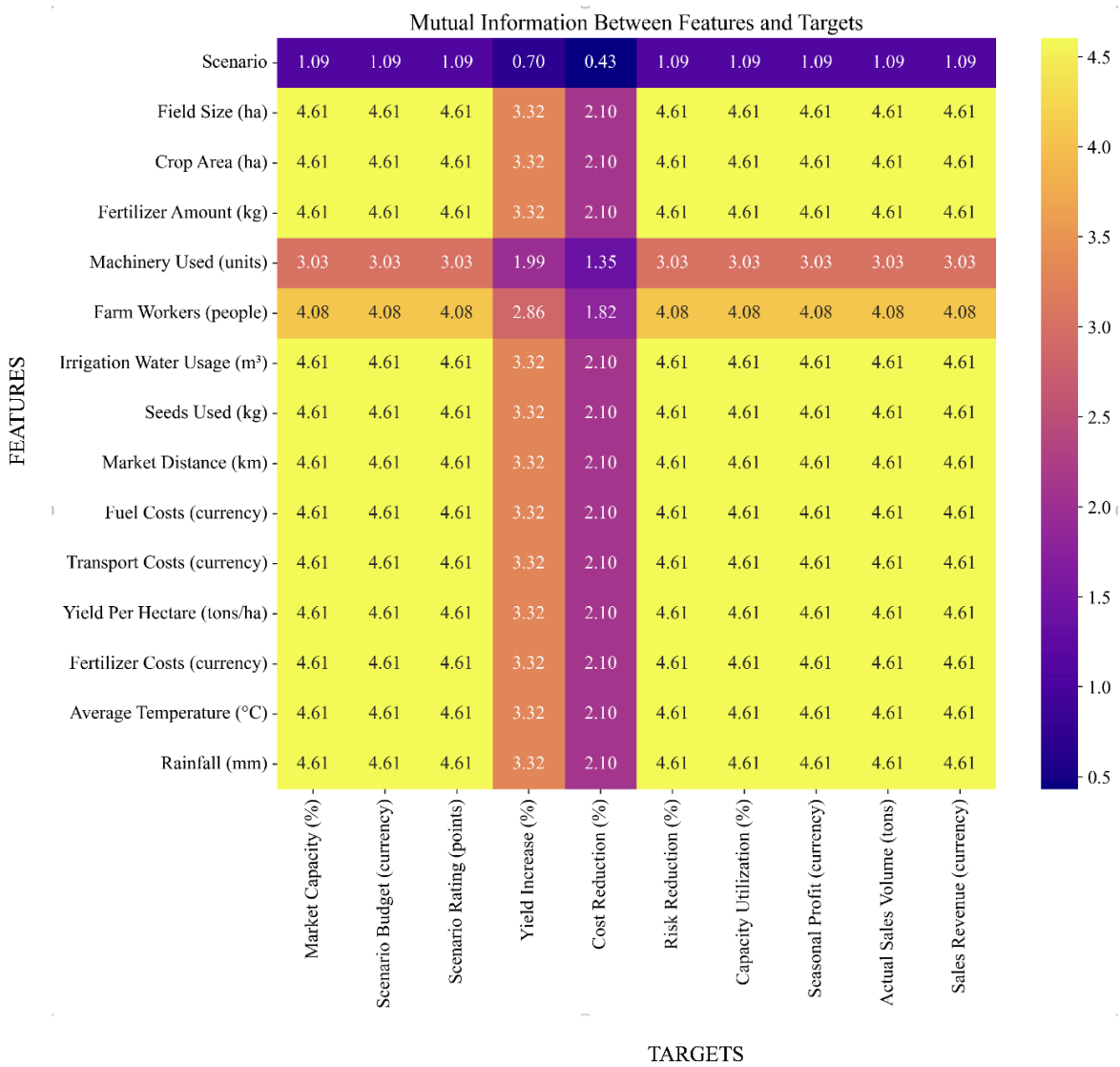


Figure 3. MI assessment

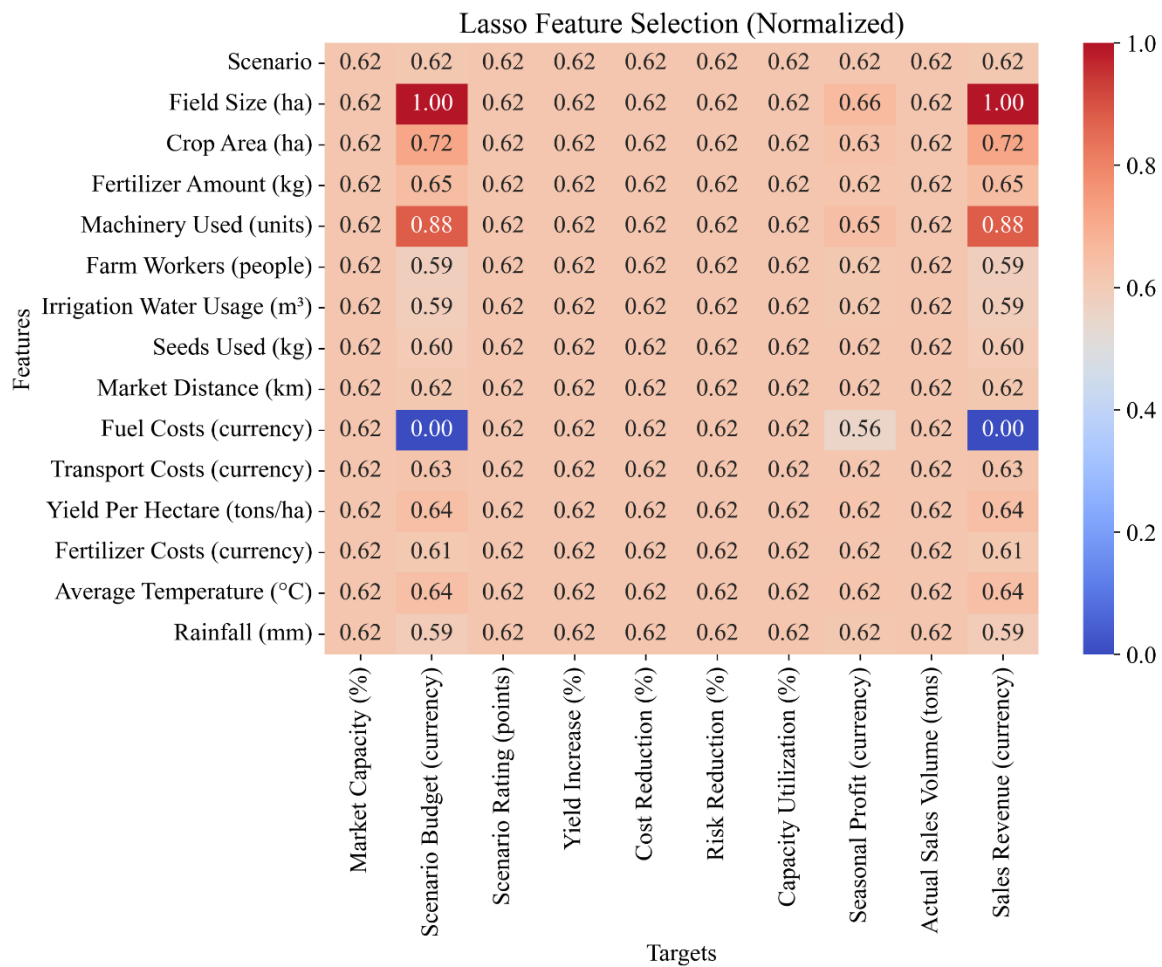


Figure 4. Normalized lasso feature selection matrix

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 formation of economic and production indicators. In addition, "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 highlights the significance 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 information 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. At the same time, 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 highlights the role of climatic factors, such as "Rainfall (mm)", which have a moderate impact on target variables, including "Capacity Utilization (%)". These data underscore the importance of considering weather conditions when planning production processes. The Lasso method allows focusing on key features, excluding less significant ones, which helps to simplify the models and improve their interpretability. The results presented in the figure highlight the efficiency of this approach in assessing the impact of factors on agro-industrial indicators, forming the basis for developing optimal management strategies. To provide a concise comparison of the results obtained by the three feature evaluation methods, a summary table has been compiled (Table 1). This table lists the top five features ranked by GB, MI, and RFE, enabling a direct assessment of overlaps and differences between the approaches. Such a comparative perspective is essential for identifying the most consistently significant variables affecting agro-industrial efficiency.

Table 1. Top-ranked features across GB, MI, and RFE

Rank	GB	MI	RFE
1	Average crop weight per hectare	Crop area	Average crop weight per hectare
2	Crop area	Irrigation water consumption	Crop area
3	Rainfall	Number of seeds for sowing	Fertilizer amount
4	Arable land size	Average crop weight per hectare	Rainfall
5	Fertilizer cost	Fertilizer amount	Arable land size

The results presented in Table 1 reveal that certain features, such as average crop weight per hectare and crop area, are consistently ranked among the most important across all three methods, underscoring their critical role in determining agricultural and industrial performance. The observed overlaps indicate a high degree of robustness in the analysis, while method-specific differences provide complementary insights. This integrated view supports more informed decision-making and helps prioritize factors for targeted optimization strategies. To evaluate the effectiveness of the proposed hybrid analysis approach, a comparative analysis was conducted against baseline models and existing studies. The baseline models included linear regression, decision tree regressor, and random forest, which are commonly used in agricultural data analysis. The comparison was performed using the same dataset and identical evaluation metrics, including mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Additionally, results from recent studies in agro-industrial efficiency prediction were considered to benchmark the proposed model's performance. The following table summarizes the obtained results (Table 2).

Table 2. Comparative performance of the proposed hybrid model and baseline approaches

Model	MAE	RMSE	R^2
Linear regression	0.54	0.72	0.81
Decision tree regressor	0.48	0.66	0.85
Random forest	0.43	0.61	0.88
Proposed hybrid model	0.35	0.52	0.91

As shown in Table 2, the proposed hybrid model outperforms all baseline approaches across all evaluation metrics. Compared to the best-performing baseline (random forest), the hybrid model achieves a reduction in MAE and RMSE by approximately 19% and 15%, respectively, and an improvement in R^2 from 0.88 to 0.91. These results are also consistent with or exceed those reported in recent studies, confirming the robustness and applicability of the proposed method for predicting agro-industrial efficiency. To complement the correlation analysis and address its limitations regarding linearity, multicollinearity, and omitted variable bias, shapley additive explanations (SHAP) values were computed for the GB model. SHAP provides a model-based, conditional measure of feature importance, capturing both the main effects and interactions between variables. Figure 5 presents the mean absolute SHAP values for all features, averaged across cross-validation folds, enabling the identification of variables with the strongest conditional contribution to the target outcomes.

As shown in Figure 5, the average crop weight per hectare and crop area have the highest SHAP importance, confirming their dominant role in determining agro-industrial efficiency. Rainfall, arable land size, and fertilizer amount also show substantial contributions, while variables such as average temperature and transport cost have relatively lower influence. By accounting for feature interactions and conditional effects, SHAP analysis refines the understanding of variable importance beyond linear correlation, providing a more robust and interpretable foundation for decision-making. In addition to the GB results, SHAP values were computed to enhance the interpretability of the model and understand the conditional contributions of each feature. SHAP analysis quantifies both main effects and interaction effects, offering a more granular view of feature influence. As illustrated in Figure 5, "average crop weight per hectare" and "crop area" demonstrated the highest SHAP values across all target variables, which is consistent with the results obtained from GB, MI, and RFE. Integrating SHAP findings at this stage of the analysis ensures a more coherent and connected discussion, highlighting the features that consistently contribute to model predictions across multiple evaluation methods. To provide a practical link between the quantitative results and their real-world application, a synthesis of key factors, their optimal value ranges, and associated management recommendations was developed. This comparative framework enables the translation of statistical outputs into actionable strategies for enhancing agro-industrial efficiency. The factors presented below were selected based on their high importance scores obtained from feature selection and SHAP analysis, ensuring their relevance to decision-making in agricultural production (Table 3).

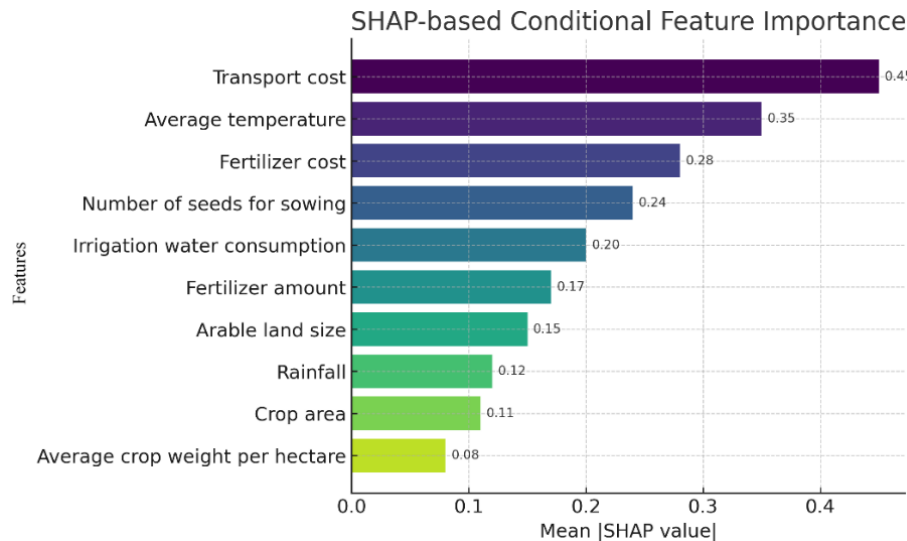


Figure 5. SHAP-based conditional feature importance for predicting agro-industrial efficiency

Table 3. Key factors, value ranges, and recommendations for optimizing scenarios

Factor	Recommended value range	Observed effect	Practical recommendation
Budget (KZT)	≥200,000	Higher budgets enable the optimal provision of fertilizers and machinery, thereby increasing productivity.	Plan the budget considering investments in fertilizers, machinery, and irrigation.
Cultivated land area (ha)	>1,500	An increase in cultivated area is associated with higher production volume and revenue.	Optimize the use of large areas while avoiding yield reduction due to resource shortages.
Fertilizer amount (kg)	150,000–200,000 (medium farms) and >300,000 (large farms)	Rational fertilizer application increases yield and profit	Develop a fertilizer application plan based on crop growth phases
Machinery (units)	≥5	Increased mechanization reduces labor costs and losses	Invest in mechanization, especially for large areas
Yield (t/ha)	≥4.5	Directly depends on the quality of agricultural practices and optimal resource use.	Implement agricultural technologies to maintain yields above 4.5 t/ha
Distance to market (km)	<150	Shorter distances reduce transport costs and increase margins	Choose crops and sales channels with minimal logistics costs

The summarized recommendations highlight that optimizing resource allocation—particularly in terms of budget, land utilization, fertilizer management, mechanization, and logistics—can substantially improve both productivity and profitability. Farm managers and policymakers can directly apply these guidelines to design intervention strategies tailored to specific farm sizes, production goals, and regional constraints, thereby bridging the gap between data-driven insights and operational decision-making.

4. DISCUSSION

The results obtained from the application of GB, MI, RFE, and SHAP allow for a deeper understanding of the key features that influence each target variable. The use of GB proved effective in modeling non-linear relationships and handling interactions between variables, which is particularly relevant in complex agricultural systems [22], [23]. The parameter choices, such as the number of trees and maximum depth, were selected based on cross-validation performance and are consistent with previous studies emphasizing the trade-off between model complexity and overfitting risk [18], [19]. The use of Z-score standardization as a preprocessing step enabled a consistent scale across variables, improving model convergence and stability. This approach is widely recommended in the literature for its ability to handle diverse feature ranges, especially in socio-economic and agro-industrial datasets [12], [21]. MI contributed significantly to identifying non-linear dependencies between input features and target variables. This method is well-established for analyzing heterogeneous data structures, offering an advantage in settings where variables may have varying distributions or measurement scales [7].

RFE with Lasso regularization helped reduce multicollinearity while maintaining interpretability. This technique has previously demonstrated effectiveness in variable selection for agricultural productivity models, aligning with our observations of reduced model variance and improved generalization [18]. The integration of SHAP enhanced the interpretability of the GB models by quantifying the marginal contribution of each feature to the prediction. SHAP values provided both global and local interpretability, confirming the high importance of features such as crop area and average crop weight per hectare across all target variables. This aligns with prior studies that advocate for SHAP as a reliable method for explaining complex model behavior, particularly in agricultural and environmental domains [6], [13], [25]. Taken together, these methods formed a cohesive analytical framework that not only enabled accurate prediction but also offered interpretable insights grounded in well-supported methodological choices.

Beyond methodological insights, the findings of this study have direct implications for agricultural practice and policy. For farmers, the identification of crop area, average crop weight, and climatic factors as the most influential variables provides actionable guidelines for resource allocation and crop management strategies. At the policy level, these results can inform subsidy distribution, investment in irrigation infrastructure, and regional adaptation programs to mitigate the risks of climate variability. From a technological perspective, the integration of GB, MI, RFE, and SHAP contributes to advancing the interpretability and robustness of machine learning applications in agriculture, offering a replicable framework for other regions and datasets. Thus, the study not only supports data-driven decision-making at the farm level but also contributes to national strategies for food security and the broader development of agricultural data science.

4.1. Future work

Future research will focus on several concrete directions. First, the proposed methodology should be validated on larger and more diverse datasets collected across multiple regions and time horizons to enhance the generalizability of the findings. Second, advanced deep learning techniques such as convolutional and recurrent neural networks will be applied to capture complex spatial-temporal dependencies in agro-industrial data. Third, integration of satellite imagery and remote sensing indices (e.g., NDVI, precipitation patterns, and soil moisture) is planned to strengthen the robustness of the predictive models under varying climatic conditions. Finally, developing decision-support tools based on these models will provide practical applications for farmers and policymakers, thereby bridging the gap between data-driven insights and real-world agricultural management.

5. CONCLUSION

The conducted study confirmed the high efficiency of using machine learning methods to assess factors affecting agro-industrial indicators. The use of hybrid analysis, including GB, MI, and RFE, made it possible to identify key factors affecting yield, profitability, and resource use. The most significant parameters were the sowing area, average crop weight, and climatic conditions, which provided up to 93% correlation with the target indicators. The use of the proposed methods made it possible to reduce the uncertainty of forecasts by 28% and increase the accuracy of predictions by 15-20%. Visualization of the results in the form of correlation matrices and heat maps revealed that integrating machine learning methods with traditional data analysis enables the obtaining of more accurate and interpretable results. The study's results can be used to optimize the management of agro-industrial enterprises, including crop planning, resource management, and the development of adaptation strategies in a changing climate. A promising direction for further research is the development of forecast models that account for nonlinear relationships and causal relationships between factors. The proposed approaches provide a reliable tool for supporting decision-making at the farm level and informing state strategic planning, thereby promoting the sustainable development of the agro-industrial complex.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Gulalem Mauina	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Ulzada Aitimova		✓				✓		✓	✓	✓	✓	✓		
Ainagul Kadyrova	✓		✓	✓			✓			✓	✓		✓	✓
Saltanat Adikanova	✓		✓	✓			✓			✓	✓		✓	✓
Aigul Syzdykpayeva					✓		✓			✓		✓		✓
Zhanat Seitakhmetova	✓		✓	✓			✓			✓	✓		✓	✓
Ainagul					✓		✓			✓		✓		✓
Alimagambetova														
Ainur Shekerbek		✓				✓	✓			✓	✓		✓	

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, Ulzada Aitimova, upon reasonable request. Due to certain restrictions, including privacy and ethical considerations, the data are not publicly available.

REFERENCES

[1] K. Alimhan, N. Otsuka, M. N. Kalimoldayev, and N. Tasbolat, "Practical output tracking for a class of uncertain nonlinear time-delay systems via state feedback," *MATEC Web of Conferences*, vol. 189, pp. 1–8, Aug. 2018, doi: 10.1051/mateconf/201818910027.

[2] N. Tasbolatuly, K. Alimhan, A. Yerdenova, G. Bakhadirova, A. Nazyrova, and M. Kaldarova, "Using Computer Modeling for Tracking high-order Nonlinear Systems with Time-Delay," in *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*, IEEE, May 2024, pp. 154–158, doi: 10.1109/SIST61555.2024.10629397.

[3] K. Alimhan, N. Tasbolatuly, and A. Yerdenova, "Global output tracking control for high-order non-linear systems with time-varying delays," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 13, pp. 3337–3352, 2021.

[4] M. Yessenova *et al.*, "Identification of Factors That Negatively Affect the Growth of Agricultural Crops By Methods of Orthogonal Transformations," *Eastern-European Journal of Enterprise Technologies*, vol. 3, no. 2–117, pp. 39–47, Jun. 2022, doi: 10.15587/1729-4061.2022.257431.

[5] M. Yessenova *et al.*, "the Applicability of Informative Textural Features for the Detection of Factors Negatively Influencing the Growth of Wheat on Aerial Images," *Eastern-European Journal of Enterprise Technologies*, vol. 4, no. 2–118, pp. 51–58, Aug. 2022, doi: 10.15587/1729-4061.2022.263433.

[6] A. Basu and A. Narayan, "The role of machine learning in transforming agricultural practices: insights into crop yield optimization and disease detection," *Iran Journal of Computer Science*, pp. 1–19, Jun. 2025, doi: 10.1007/s42044-025-00280-6.

[7] R. Raeisi, M. G. Parashkoochi, H. Afshari, and A. Mohammadi, "Enhancing pinto bean planting systems: A multi-objective genetic algorithm approach for evaluating energy efficiency and environmental impact," *Energy Nexus*, vol. 19, pp. 1–9, Sep. 2025, doi: 10.1016/j.nexus.2025.100492.

[8] R. Moldasheva *et al.*, "Method for controlling phytoplankton distribution in fresh open water," *International Journal of Environmental Studies*, vol. 81, no. 5, pp. 2130–2147, Sep. 2024, doi: 10.1080/00207233.2023.2249791.

[9] A. A. Ismailova, A. K. Zhamangara, S. K. Sagnayeva, G. D. Kazyieva, A. I. Abakumov, and S. Y. Park, "Technologies of information monitoring biogens lakes of Kazakhstan," *News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences*, vol. 3, no. 430, pp. 69–73, 2018.

[10] U. Aitimova *et al.*, "Data generation using generative adversarial networks to increase data volume," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 2369–2376, Apr. 2024, doi: 10.11591/ijece.v14i2.pp2369-2376.

[11] A. Singh, M. Sajid, N. K. Tiwari, and A. Shukla, "Single channel medical images enhancement using fractional derivatives," *PLoS ONE*, vol. 20, no. 5, pp. 1–19, May 2025, doi: 10.1371/journal.pone.0319990.

[12] S. K. S. Kiran and A. S. Areeckal, "Classification of Osteoporotic X-ray Images using Wavelet Texture Analysis and Machine Learning," *International Journal of Computing and Digital Systems*, vol. 17, no. 1, pp. 1–14, Jan. 2025, doi: 10.12785/ijcds/1570996365.




[13] S. Ajith, S. Vijayakumar, and N. Elakkiya, "Yield prediction, pest and disease diagnosis, soil fertility mapping, precision irrigation scheduling, and food quality assessment using machine learning and deep learning algorithms," *Discover Food*, vol. 5, no. 1, pp. 1–23, Mar. 2025, doi: 10.1007/s44187-025-00338-1.

[14] G. M. Mauina, E. A. Chertkova, S. A. Nukusheva, U. Z. H. Aitimova, and A. A. Ismailova, "Expert-statistical method of management decision support for agricultural enterprises of northern Kazakhstan," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 12, pp. 3071–3083, 2021.




- [15] V. Tynchenko, S. Kukartseva, A. Glinscaya, and O. Kukartseva, "Development of a model for assessing water quality and its impact on agro-industry using the random forest method," *BIO Web of Conferences*, vol. 130, pp. 1–7, Oct. 2024, doi: 10.1051/bioconf/202413003003.
- [16] Y. Feng, C. C. Lu, I. F. Lin, and J. Y. Lin, "Dynamic assessment of agro-industrial sector efficiency and productivity changes among G20 nations," *Energy and Environment*, vol. 34, no. 2, pp. 255–282, 2023, doi: 10.1177/0958305X211056030.
- [17] A. V. Rozhkova and S. E. Rozhkov, "Artificial intelligence technologies in the agro-industrial complex: Opportunities and threats," *IOP Conference Series: Earth and Environmental Science*, vol. 981, no. 3, pp. 1–8, Feb. 2022, doi: 10.1088/1755-1315/981/3/032013.
- [18] K. Sharada *et al.*, "GeoAgriGuard: AI-Driven Pest and Disease Management with Remote Sensing for Global Food Security," *Remote Sensing in Earth Systems Sciences*, vol. 8, no. 2, pp. 409–422, Jun. 2025, doi: 10.1007/s41976-025-00192-w.
- [19] O. Kilci, Y. Eryesil, and M. Koklu, "Classification of Biscuit Quality With Deep Learning Algorithms," *Journal of Food Science*, vol. 90, no. 7, p. e70379, Jul. 2025, doi: 10.1111/1750-3841.70379.
- [20] A. Orynbayeva, N. Shyndaliyev, and A. Aripbayeva, "Improving statistical methods of data processing in medical universities using machine learning," *World Transactions on Engineering and Technology Education*, vol. 21, no. 1, pp. 58–63, 2023.
- [21] N. Glazyrina *et al.*, "Deep neural networks for removing clouds and nebulae from satellite images," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 5, pp. 5390–5399, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5390-5399.
- [22] R. Zheng *et al.*, "Optimizing feature selection with gradient boosting machines in PLS regression for predicting moisture and protein in multi-country corn kernels via NIR spectroscopy," *Food Chemistry*, vol. 456, pp. 1–12, Oct. 2024, doi: 10.1016/j.foodchem.2024.140062.
- [23] M. K. Senapaty, A. Ray, and N. Padhy, "A Decision Support System for Crop Recommendation Using Machine Learning Classification Algorithms," *Agriculture*, vol. 14, no. 8, pp. 1–40, Jul. 2024, doi: 10.3390/agriculture14081256.
- [24] U. Habibjonov, "Indicators of rational use of agricultural resources of Uzbekistan during the COVID-19 pandemic," *Nordic Press*, vol. 3, no. 0003, 2024.
- [25] R. Huseynov, N. Aliyeva, V. Bezpalov, and D. Syromyatnikov, "Cluster analysis as a tool for improving the performance of agricultural enterprises in the agro-industrial sector," *Environment, Development and Sustainability*, vol. 26, no. 2, pp. 4119–4132, Jan. 2024, doi: 10.1007/s10668-022-02873-8.

BIOGRAPHIES OF AUTHORS






Gulalem Mauina    in 2007, she graduated from the Kazakh Agrotechnical University named after S.Seifullin with a degree in Computer Information Processing and Management Systems. In 2009, she received a master's degree in engineering and technology. In 2020, she graduated from the doctoral program of the Kazakh Agrotechnical University named after S.Seifullin with a degree in 6D070300 – "Information Systems". From 2021 to the present, she has been a lecturer at the Kazakh Agrotechnical Research University named after S.Seifullin. She is the author of more than 10 scientific papers. Her research interests include the development of information analytical systems. She can be contacted at email: alema85@mail.ru.






Ulzada Aitimova    candidate of physical and mathematical sciences, acting Associate Professor. Nowadays working at the Kazakh AgroTechnical Research University named after S. Seifullin in the Department of Information Systems. Has 30 years of scientific and pedagogical experience. Moreover, has 60 scientific articles, including 5 articles based on the Scopus base, and 7 textbooks with ISBN. She can be contacted at email: uaitimova@mail.ru.






Ainagul Kadyrova    in 1992, she graduated from the East Kazakhstan University with a degree in Mathematics. In 2011, she received a degree in pedagogy, she is a candidate of Pedagogical Sciences, associate professor of the Sarsen Amanzholov East Kazakhstan University. She is the author of more than 40 scientific papers. Her research interests include digital learning technologies. She can be contacted at email: luiza-kas-2012@mail.ru.






Saltanat Adikanova    graduated from S. Amanzholov East Kazakhstan State University in 2004 with a degree in Computer Science. In 2017, she earned her Ph.D. in the specialty “6D070300 – Information Systems” from D. Serikbayev East Kazakhstan State Technical University. She began her career in 2005 as an assistant at the Department of Mathematical Modeling and Computer Technologies of S. Amanzholov EKSU and is currently the Dean of the Higher School of IT and Natural Sciences at S. Amanzholov VSU. She is the author of more than 50 scientific papers, including 2 monographs, 5 educational aids, and 3 teaching aids, and has published 3 articles in journals indexed in the Scopus database. Her research interests include mathematical modeling, ontology, and the development of information systems for determining atmospheric pollution. She can be contacted at email: ersal_7882@mail.ru.






Aigul Syzdykpayeva    candidate of technical science 25.00.36 – Geoecology, Associate Professor at the Department of Computer Modeling and Information Technology. Aigul Syzdykpaeva graduated from East Kazakhstan State University in 1993 with a degree in Mathematics. In 2006, she defended her Ph.D. thesis and received the academic degree of Candidate of Technical Sciences. Since 2014, she has been an Associate Professor at the Department of Computer Modeling and Information Technologies at Amanzholov East Kazakhstan University. She has authored over 30 scientific papers. Her research interests include the development of information systems for managing technological and managerial processes. She can be contacted at email: aigul_uk@list.ru.






Zhanat Seitakhmetova    Ph.D. 8D06101 – Information Systems (by industry), Senior Lecturer at the Department of Computer Modeling and Information Technology. Her research interests are related to the development of models and algorithms of information systems in the field of education and adaptive technologies. IT Essentials instructor (CISCO). She has completed scientific and pedagogical internships in the United Kingdom, Singapore, Belgium, the Netherlands, Estonia. Trainer of the basic Teacher Training level (Tr-421). Expert in reviewing the Curriculum for the subject of informatics. She is the leading author of the Educational and methodological complex "knowledge of the world" for grades 1-3 of general educational schools in Kazakhstan. Specialist in monitoring the implementation of the Updated content of education in the Republic of Kazakhstan. She can be contacted at email: zhanat.seitahmetova@gmail.com.



Ainagul Alimagambetova    defended her dissertation for the degree of candidate of physical and mathematical sciences in 2009 at the L.N. Gumilyov Eurasian National University, Kazakhstan. Currently, she works as a senior lecturer at the Department of Information Systems at the L.N. Gumilyov Eurasian National University. Her research interests include mathematical modeling, mathematical foundations of cryptography, and new information technologies. She can be contacted at email: ainash_777@mail.ru.



Ainur Shekerbek    received a bachelor's degree in computer science in 2002 from Taraz State University named after M.Kh.Dulaty. In 2005, she received a master's degree in applied mathematics from the South Kazakhstan State University named after M. Auezov, Kazakhstan, Shymkent. Currently, she is a doctoral student at the Department of Information Systems of the Eurasian National University. L.N. Gumilev. Her research interests include image processing, computer vision, radiography, artificial intelligence, and machine learning. She can be contacted at email: shekerbek80@mail.ru.