Journal of Agricultural Economics & Development

Homepage: https://jead.um.ac.ir

Research Article
Vol. 40, No. 2, Summer 2026, p. 103-127

The Application of Data Mining in Analyzing Factors Affecting the Classification of Technical Efficiency of Wheat Farmers: A Study in Ahar County

J. Vahedi[1], M. Ghahremanzadeh[1*], Gh. Dashti[1], P. Pakrooh[2]

1- Department of Agricultural Economics, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
(*- Corresponding Author Email: Ghahremanzadeh@tabrizu.ac.ir)
2- Marie Sklodowska-Curie Research Fellow at University of Milano-Bicocca, Milan, Italy

Received: 03 May 2025
Revised: 25 November 2025
Accepted: 02 December 2025
Available Online: 02 December 2025

How to cite this article:
Vahedi, J., Ghahremanzadeh, M., Dashti, Gh., & Pakrooh, P. (2026). The application of data mining in analyzing factors affecting the classification of technical efficiency of wheat farmers: A Study in Ahar County. *Journal of Agricultural Economics & Development, 40*(2), 103-127.
https://doi.org/10.22067/jead.2025.93203.1349

Abstract

Ahar County, as one of the main agricultural production areas, produces a considerable share of rainfed wheat in East Azerbaijan Province. Improving technical efficiency (TE) in this region can have a significant impact on farmers' productivity and economic sustainability. Therefore, the present study was conducted with the aim of applying data mining to analyze the efficiency of rainfed wheat farmers in Ahar County. To this end, farmers' TE was calculated using Data Envelopment Analysis (DEA), and those with efficiency scores above the regional average (0.59) were classified as the high-efficiency group, while the rest were categorized as the low-efficiency group. Subsequently, t-tests and chi-square tests were employed to identify variables likely to influence TE. Machine learning algorithms, including logistic regression, support vector machines (SVM), k-nearest neighbors (k-NN), and random forest (RF), were then applied for classification and analysis of TE. The results indicated that logistic regression outperformed the other algorithms. The output of this algorithm revealed that factors such as herbicides, weed control, manure, land rental value, nitrate fertilizers, number of farm plots, pesticides, farmer's age, combined harvesting method, experience, household members with university education, seed procurement from the Agricultural Organization, and residence in rural areas have a positive effect on TE. Conversely, factors including mixed landownership (personal-rental), seed procurement from personal sources, and non-agricultural income exerted a negative influence on efficiency. Based on the findings, it is recommended that farmers receive necessary training on the optimal management of influential agricultural inputs, including herbicides, manure, nitrate fertilizers, and pesticides. Furthermore, policymakers are advised to enhance the motivation of farmers operating on rented lands by providing financial incentives and advisory services. The development of supportive programs for the supply of quality seeds and agricultural inputs through the Agricultural Jihad Organization is also proposed.

Keywords: Data mining, Machine learning, Technical efficiency, Wheat

https://doi.org/10.22067/jead.2025.93203.1349

Introduction

Optimal utilization of production resources in the agricultural sector is of paramount importance for feeding the growing global population. To address the challenges associated with meeting food demand, either the cultivated area must be expanded, or production efficiency must be maximized (Bahiraie *et al.*, 2020). These challenges generally encompass climate change, declining water resources, and limitations on available agricultural land, all of which adversely affect crop yields. Despite the constraints on agricultural production resources and the expanding needs of human societies, determining the efficiency of agricultural producers enables the quantification of the gap between the most efficient producers and others operating under identical technological conditions and with similar resource endowments. Such analysis can help identify weaknesses and strengths within production systems, thereby facilitating the design of more effective policies to enhance efficiency. Consequently, assessing farmers' efficiency can prove highly valuable in evaluating the suite of policies implemented in the agricultural sector (Davodnia *et al.*, 2023).

Numerous researchers have calculated the technical efficiency of wheat production and identified its determinants across various regions of Iran. Khodaverdizadeh *et al*. (2019) employed both Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA) to estimate wheat production efficiency in Urmia County, identifying chemical fertilizers, pesticides, and agricultural machinery as significant factors influencing efficiency. Tanursaz *et al.* (2021), after calculating the technical efficiency of wheat farmers in Dezful County using SFA, applied the two-stage Heckman procedure to evaluate efficiency determinants. They reported an average technical efficiency of 0.78. They found that the adoption of conservation tillage technology, the ratio of conservation tillage machinery per village, farming experience, water consumption, and education level exerted positive and significant effects on efficiency. In contrast, labor input and distance from the village had negative and significant impacts. Ghasemi *et al.* (2023) investigated the technical, environmental, and economic efficiency of rainfed wheat farmers in Ahar County using SFA, calculating an average efficiency score of 0.66. Machinery, animal manure, chemical fertilizers, and pesticides were identified as key factors affecting efficiency. Naruei *et al.* (2024) examined the impact of climate change adaptation strategies on the technical efficiency of wheat farmers in the Sistan region using a modified endogenous stochastic frontier model. The results indicated that changes in land size, conservation tillage practices, and adjustments in planting dates had the strongest effects, while rainfall and the use of bio-fertilizers exhibited the weakest impacts. The average technical efficiency was estimated at 0.82.

Aryan *et al.* (2024) conducted a field experiment in eastern Afghanistan to determine the optimal timing of nitrogen fertilizer application for improving wheat technical efficiency. They found that average technical efficiency could increase by up to 68.6% with optimal nitrogen fertilizer management. Samson *et al*. (2024) employed SFA to assess the technical efficiency of wheat farmers in Jigawa State, Nigeria. Their results indicated that seed quality, chemical fertilizers, labor input, herbicides, and farming experience exerted positive effects on technical efficiency, whereas farmers' age, poor seed quality, inadequate road infrastructure, and limited access to financial resources and inputs had negative impacts. Girma *et al.* (2024) applied SFA to analyze wheat farmers' technical efficiency and identify its determinants in the Oromia region of Ethiopia, reporting an average technical efficiency score of 0.82. They found that education level, frequency of contact with agricultural extension agents, access to credit, farming experience, soil fertility, and participation in training programs contributed to higher efficiency, while non-farm income was associated with reduced efficiency.

Collectively, these studies demonstrate that environmental, economic, and social factors play crucial roles in determining technical efficiency, and identifying these determinants can facilitate improvements in farmers' yield.

Although these studies have employed methods such as stochastic frontier models and Hackman's two-stage procedure to examine factors influencing technical efficiency, such approaches entail certain limitations, including dependence on restrictive statistical assumptions and an inability to uncover nonlinear and complex relationships among variables. These limitations can lead to inaccurate or misleading results, particularly when real-world data contains noise, missing values, and intricate causal relationships. Consequently, the necessity of adopting modern data mining and machine learning techniques becomes evident. Data mining, with its capacity to analyze complex datasets, enables the identification of hidden patterns and relationships among factors affecting technical efficiency. These methods not only provide higher accuracy and reliability in classifying farmers but also allow for more precise targeting of efficiency improvement strategies. Furthermore, data mining techniques offer the advantage of knowledge discovery directly from data without requiring stringent statistical assumptions, while validation techniques (e.g., cross-validation) help prevent overfitting. Nevertheless, research applying data mining and machine learning methods for the systematic classification and analysis of factors influencing farmers' technical efficiency remains scarce in the literature. Therefore, this study, as one of the first endeavors in this domain, seeks to deliver a comprehensive and precise analysis of the determinants of wheat technical efficiency by leveraging advanced data mining approaches.

Ahar County is recognized as one of the major hubs for wheat production in Iran. According to statistics from the East Azerbaijan Agricultural Jihad Organization (2022), in the Iranian calendar year 1400 (2021–2022), Ahar County accounted for approximately 10% of the province's total wheat-cultivated area, of which nearly 50,000 hectares (over 90%) were dedicated to rainfed wheat cultivation. Within the county's 52,000 hectares of arable land, 28,000 hectares are allocated to wheat production. This substantial cultivated area, significantly larger than that devoted to other crops in the region, highlights Ahar County's strategic importance in provincial wheat output. Nevertheless, given the relatively modest wheat yield per hectare in the area, further investigation into technical efficiency is warranted. Researching the economic dimensions of wheat production in Ahar County could therefore yield positive impacts on regional productivity. Moreover, the area's diversity in farming practices and levels of mechanization provides a suitable setting for testing the capabilities of data mining methods in identifying complex patterns influencing technical efficiency. In this context, the present study can serve as a robust framework for analyzing the current situation and identifying pathways to improve efficiency; its findings may assist farmers in making more informed decisions regarding the optimal utilization of production resources.

Based on the foregoing discussion, the primary objective of this study is to classify wheat farmers in Ahar County according to their level of technical efficiency and to systematically identify the most influential factors affecting efficiency by employing modern data mining and machine learning techniques. The fundamental innovation of the study lies in its status as one of the first studies in this domain to substitute traditional econometric and mathematical programming approaches with an analytical framework grounded in intelligent algorithms. Consequently, this research provides policymakers and farmers with a precise and practical framework for targeting efficiency improvement strategies based on real-world data. The study comparatively employs multiple data mining algorithms to deliver the most accurate and reliable model for identifying determinants of technical efficiency. Accordingly, the objective of this research is to apply data mining techniques to analyze the

factors influencing the classification of technical efficiency among wheat farmers in Ahar County.

Materials and Methods

In this study, data mining techniques were employed to classify rainfed wheat-producing farmers in Ahar County based on their technical efficiency and to analyze the factors influencing it. Initially, the technical efficiency of wheat farmers was calculated using Data Envelopment Analysis (DEA) with an output-oriented approach under the assumption of variable returns to scale (VRS). Output was defined as wheat yield (kg), while inputs included area under cultivation (hectares), seed (kg), labor (person-days), and tractor usage (hours). DEA is a non-parametric method for evaluating the efficiency of decision-making units (DMUs) by comparing their yield based on inputs and outputs. This approach enables the simultaneous examination of multiple factors and facilitates the identification of both efficient and inefficient units (Cooper *et al.*, 2007).

Subsequently, wheat farmers were classified into two groups: high-efficiency farmers (those with efficiency scores above the regional average) and low-efficiency farmers (those with efficiency scores below the regional average). Next, considering variables potentially influencing their technical efficiency, the most appropriate machine learning algorithm was identified, one capable of assigning farmers to their respective efficiency class (high or low) with minimal classification error. In other words, for subsequent economic analyses, the algorithm was selected whose predicted classification of farmers into high- or low-efficiency groups showed the highest agreement with their actual efficiency-based classification derived from DEA calculations (For brevity, the phrases "above average" and "below average" will henceforth be omitted when referring to efficiency groups). The higher the algorithm's predictive accuracy, the greater the reliability of the ensuing economic analyses.

Data Mining

Data mining is the process of discovering patterns and hidden information from existing datasets (Shu & Ye, 2023). Within the knowledge discovery process, a dataset is first selected for examination. Subsequently, data cleaning operations, including noise removal, normalization, and standardization, are performed to enhance data quality. The next phase involves applying specific analytical techniques to the processed data, such as statistical tests (e.g., Student's t-test, chi-square test), visualization (to provide clear insights into patterns and relationships among variables), and classification (to identify and organize data into distinct groups based on similarities and shared characteristics). Appropriate algorithms are then selected to perform the intended analytical tasks. Finally, once suitable outputs are generated, the results are interpreted and evaluated (Gullo, 2015; McClendon & Meghanathan, 2015).

Machine Learning

Machine learning is a method by which computer systems learn through examples. A major category of machine learning algorithms is supervised learning, which operates by inferring patterns from labeled training data. In this approach, specific features or a set of features are provided to the algorithm to predict target outcomes. Among machine learning algorithms commonly employed in data mining analyses are classification algorithms (*Ngai et al.*, 2009; Bermudez *et al.*, 2011), whose objective is to construct models capable of predicting future outcomes by assigning database records to predefined classes based on specific criteria. Widely used tools for classification analysis include logistic regression, support vector machines (SVM), k-nearest neighbors (k-NN), and random forest (RF) (Sen *et al.*, 2020), each of which will be elaborated upon in the following sections.

Logistic Regression

Logistic regression is a supervised learning algorithm in which the model is trained on a

labeled dataset comprising input data (features) and corresponding output labels (target class). It is a method employed for binary classification problems to predict the probability of a binary outcome, i.e., two possible classes, such as high-efficiency and low-efficiency farmers. This algorithm models the relationship between a dependent variable (typically binary; in this study, farmers categorized as high- or low-efficiency relative to the regional average) and one or more independent variables (which may encompass various data types, such as labor, cultivated area, and machinery) using a logistic function. The logistic function, also known as the sigmoid function, maps the output to a probability value ranging between 0 and 1 (Shetty *et al.*, 2022). Logistic regression operates based on the logistic (sigmoid) function, which is defined as Equation 1:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

Where z represents a linear combination of the features, as expressed in Equation (2) (James, 2013):

$$z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \quad (2)$$

In Equation (2), α_0 denotes the intercept, α_i represents the regression coefficients associated with the features, and x_i corresponds to the input features (independent variables). The output of the sigmoid function $\sigma(z)$ yields the probability of the dependent variable, for instance, $P(Y = 1|X) = \sigma(z)$. Ultimately, the regression coefficients can be estimated using maximum likelihood estimation, as formulated in Equation (3) (James, 2013).

$$P(Y = 0|X) = 1 - P(Y = 1|X) \quad (3)$$

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning algorithm designed for classification tasks that operates based on the identification of support vectors. This algorithm employs kernel functions and is particularly effective in pattern recognition and the classification of nonlinear data. SVM constructs a hyperplane to separate different classes. The hyperplane is iteratively optimized by the SVM algorithm with the objective of maximizing the margin between classes while minimizing classification errors (Sujatha & Mahalakshmi, 2020).

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a machine learning algorithm that classifies new data points by comparing them with existing data points based on similarity metrics such as Euclidean distance. The algorithm identifies the *k* nearest neighbors (data points) to predict the class of a new observation. The accuracy of this algorithm depends on data quality, and it exhibits robustness against noise. The primary challenge in applying KNN lies in selecting an appropriate number of neighbors (*k*). KNN is widely employed in statistical estimation and pattern recognition based on proximity relationships among objects (Yuvalı *et al.*, 2022).

Random Forest (RF)

Random Forest (RF) is a modern ensemble method based on decision trees, comprising a collection of classification and regression trees. Each tree functions as an independent classifier for learning and prediction. Ultimately, the individual predictions are aggregated, typically through majority voting for classification or averaging for regression, to produce a final ensemble prediction that generally outperforms a single classifier (Xu & Yin, 2021).

Evaluation Metrics

Following the estimation of machine learning models, it is essential to evaluate the performance and predictive power of these algorithms using a set of quantitative metrics to ensure that the model adequately fulfills its intended tasks. The most important evaluation metrics include the confusion matrix, recall, accuracy, precision, F1-score, Cohen's kappa statistic, and the relative operating characteristic (ROC) curve. Each of these metrics reflects distinct aspects of model performance. Furthermore, employing this comprehensive set of metrics facilitates a more precise and nuanced analysis of prediction

outcomes.

Confusion Matrix

The confusion matrix is a tabular layout that enables the visualization of a supervised learning algorithm's performance. Each column of the matrix represents the instances of a predicted class, while each row corresponds to the instances of an actual (true) class. All correct predictions are positioned along the diagonal of the table, allowing errors to be readily identified as non-zero values located off the diagonal. The confusion matrix for a binary classification problem is presented in Table 1 (Vahedi *et al*., 2024).

Table 1- Confusion matrix

		Predicted class	
		Particular class (C_1)	**Different class (C_2)**
Actual class	Particular class (C_1)	True positive (TP)	False negative (FN)
	Different class (C_2)	False positive (FP)	True negative (TN)

In the aforementioned confusion matrix, TP (True Positive) represents the number of samples that truly belong to class C_1 and are correctly predicted as C_1 by the model. TN (True Negative) denotes the number of samples that truly belong to class C_2 and are correctly predicted as C_2. FP (False Positive) refers to samples that do not belong to class C_1 but are erroneously classified as C_1, while FN (False Negative) indicates samples that do not belong to class C_2 but are mistakenly classified as C_2. Based on these definitions, a set of evaluation metrics, formulated in Equations (4) through (10), can be derived from the confusion matrix to assess the classification accuracy of the algorithms.

$$P = Actual\ positive = TP + FN \quad (4)$$

$$P^1 = Predicted\ positive = TP + FP \quad (5)$$

$$N = Actual\ negetive = FP + TN \quad (6)$$

$$N^1 = Predicted\ negetive = FN + TN \quad (7)$$

$$TP\ rate = Sensitivity = TP/P = Recall \quad (8)$$

$$Precision = TP/P^1 \quad (9)$$

$$Accuracy = (TP + TN/P + N) = TP/(P + N) + TN/(P + N) = TP/P \times P/(P + N) + TN/N \times N/(P + N) = Sensitivity \times P/(P + N) + Specificity \times N/(P + N) \quad (10)$$

where, P denotes the actual positive values, P^1 represents the predicted positive values, N indicates the actual negative values, and N^1 refers to the predicted negative values. Recall is defined as the ratio of correctly predicted positive instances to the total number of actual positives in the class, while precision represents the ratio of correctly predicted positives to the total number of predicted positives (e.g., the number of high-efficiency farmers correctly predicted as high-efficiency divided by the total number of farmers predicted as high-efficiency). Finally, Accuracy is the ratio of the total number of correct predictions to the total number of predictions (Hackeling, 2017). In certain cases, Recall and Precision do not increase simultaneously and may fail to enhance overall classification accuracy. To address this limitation, the F1-score is employed alongside the aforementioned metrics. This metric represents the harmonic mean of recall and precision, thereby providing a balanced assessment of classification performance. The F1-score can be calculated as expressed in Equation (11) (Khan *et al*., 2020):

$$F1 - score = 2\frac{(Recall \times Precision)}{(Recall + Precision)} \quad (11)$$

Cohen's Kappa

Cohen's Kappa statistic is commonly employed for binary classification problems. The lowest possible value of this metric is –1, while the highest (optimal) value is +1. Based on the cells of a 2×2 confusion matrix, Cohen's Kappa is defined as Equation (12) (Warrens, 2008):

$$\kappa = \frac{2.(TP.TN - FP.FN)}{(TP + FP).(FP + TN) + (TP + FN).(FN + TN)} \quad (12)$$

Relative Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve is a graphical tool used to evaluate the performance of a binary classification model. This curve illustrates the relationship between two key metrics: sensitivity (true positive rate) and the false positive rate. Accordingly, the area under the curve (AUC), also referred to as the discrimination measure, indicates the model's ability to distinguish between classes, as formulated in Equation (13) (Hordri *et al*., 2018).

$$AUC = \frac{1}{2}.(Sensitivity + Specificity) \quad (13)$$

In this study, 5-fold cross-validation is employed. This procedure divides the dataset into five subsets of equal size. Subsequently, one subset is used for testing while the remaining four subsets are utilized for model training. This process is repeated iteratively until each subset has been used once as the test set (Otunaiya & Muhammad, 2019).

In this research, the grid search method was employed to optimize machine learning models and precisely determine hyperparameters. Hyperparameters are values specified prior to the model training process and exert a considerable influence on the model's final performance. These parameters include values such as the number of trees in a random forest. The grid search method enables a systematic exploration of a predefined hyperparameter space; combined with cross-validation, it facilitates the selection of the optimal hyperparameter combination for the model, thereby achieving enhanced predictive accuracy (Ranjan *et al*., 2019). Data were collected through simple random sampling via face-to-face interviews and structured questionnaires administered to 223 rainfed wheat producers in Ahar County during the Iranian calendar year 1400 (2021–2022). The conceptual framework of the present research is illustrated in Fig. 1.

Results and Discussion

In this study, the technical efficiency of rainfed wheat-producing farmers in Ahar County was initially calculated using Data Envelopment Analysis (DEA), with its frequency distribution reported in Table 2. The results present a clear depiction of efficiency dispersion among farmers. According to these data, a substantial proportion of farmers, 44.4%, fall within the lowest efficiency tier, scoring below 50%. This pattern becomes more pronounced when incorporating farmers in the 51–70% efficiency range, who constitute 25.6% of the sample; collectively, 70% of all sampled farmers, i.e., the majority continue their operations at an efficiency level of 70% or lower. In contrast, the share of farmers achieving higher efficiency levels is considerably limited. Only 13.4% of farmers fall within the 71–90% range, while the highest efficiency tier (above 90%) is attained by merely 16.6% of farmers. The cumulative frequency distribution further corroborates this observation, revealing that 186 farmers, equivalent to 83.4% of the total sample achieved scores of 90% or below. Consequently, it can be inferred that the efficiency distribution among wheat farmers in Ahar County is skewed toward low and medium levels, thereby revealing substantial potential for enhancing productivity and optimizing resource management across a significant portion of this agricultural community.

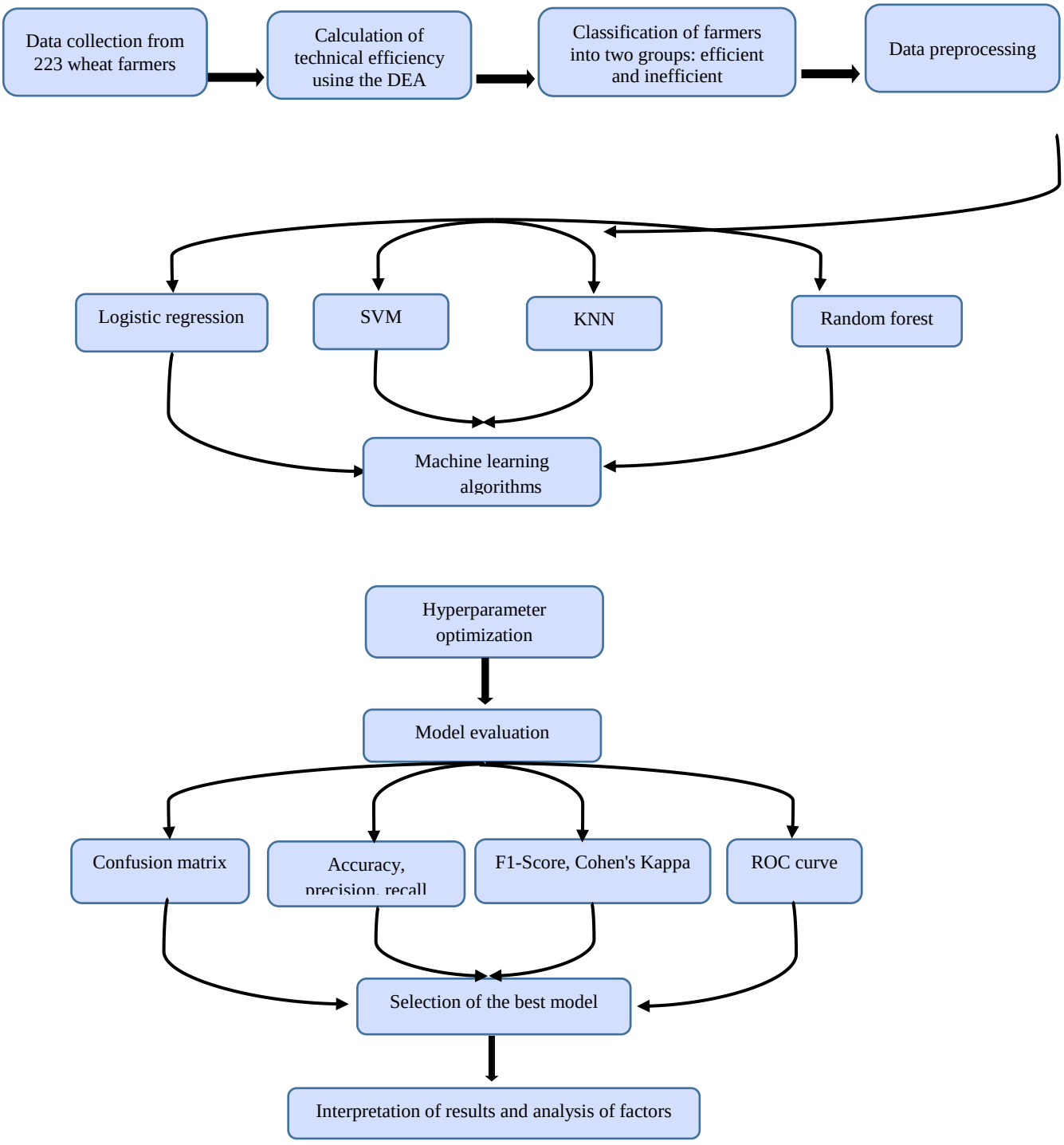

Figure 1- Conceptual framework of study

According to Table 2, the mean technical efficiency of wheat growers in the region equals 0.59, indicating that, on average, farmers utilize only 59% of their production potential. In other words, with their current input levels, farmers could potentially increase output by 41%. The standard deviation of 0.24 further confirms the existence of a relatively pronounced gap in

technical efficiency levels among farmers; while some operate near optimal performance, others face considerable inefficiency. The wide efficiency ranges from 0.22 to 1.00 (a difference of 0.78 units) reflects a substantial yield disparity within the region's rainfed wheat production system. This finding reveals that at least one farmer in the sample utilizes merely 22% of their production potential, whereas others have attained full efficiency (100%). Such a marked disparity clearly underscores the presence of structural deficiencies, unequal access to inputs, gaps in technical and technological knowledge, and variations in farm management quality among farmers. These conditions create an ideal setting for applying data mining techniques to identify the key factors driving this wide efficiency gap and to design differentiated policy interventions tailored to each farmer group.

To scientifically address this extensive heterogeneity and systematically analyze the factors underlying the efficiency gap, it was necessary to classify farmers into distinct and meaningful groups. Accordingly, in the next step and to prepare the data for analysis using machine learning algorithms, farmers with technical efficiency below the regional average were assigned to group zero (low-efficiency farmers), while those with efficiency above the average were placed in group one (high-efficiency farmers). This binary classification formed the basis for implementing classification algorithms in the subsequent phase of the research.

Five types of data have been used in this research: continuous, countable, nominal, dummy, and ordinal. Prior to implementing machine learning models, an analytical examination of these variables was conducted. Table 3 presents the mean and median values of individual and economic characteristics for low-efficiency and high-efficiency farmers in Ahar County. The results indicate statistically significant differences between the two farmer groups across various dimensions, including cultivated area, land rental value, seed quantity,

Table 2- Frequency distribution of technical efficiency of wheat farmers in Ahar county

Technical Efficiency Range	Frequency	Frequency Percentage	Cumulative Frequency
<50	99	44.4	99
51-70	57	25.6	156
71-90	30	13.4	186
>90	37	16.6	223
Mean	0.59		
Std.ev	0.24		
Minimum	0.22		
Maximum	1		

Labor input, tractor usage, nitrate fertilizer consumption, age, and farming experience[1].According to Table 3, the average consumption of all inputs is notably lower among low-efficiency farmers compared to their high-efficiency counterparts. Additionally, high-efficiency farmers are older and have more farming experience relative to inefficient farmers.

The median values further reveal that the cultivated area for 50% of low-efficiency farmers is less than 2 hectares, whereas this figure amounts to 7.3 hectares for high-efficiency farmers. Similarly, the median values of other variables exhibit considerable disparities between the two farmer groups. Moreover, the mean values for high-efficiency farmers closely approximate their respective medians, a pattern also observed among low-efficiency farmers. This proximity between means and medians in both groups may indicate a relatively symmetric distribution of inputs within each category. Overall, it appears that one of the principal factors underlying

1 - Given that the study focuses on rainfed wheat production, wheat output in the region is entirely dependent on precipitation. For this reason, water as an input variable was not considered in the analysis.

efficiency differences among regional farmers is their varying level of access to agricultural inputs.

Table 3- Mean and median of some individual and economic variables in efficient and inefficient farmers

Inputs	Mean		Median	
	Low-efficiency farmers	High-efficiency farmers	Low-efficiency farmers	High-efficiency farmers
Land (hectar)	2.2	7.8	2	7.3
Rental value of lands (million toman)	1.9	5.8	1.2	5
Seed (kg)	481	1707	420	1500
Labor (man-day)	12	33	10	25
Tractor (hour)	22	59	18	50
Nitrate fertilizer	223	455	250	350
Age (year)	48	54	48	56
Experience (year)	32	40	30	40

Fig. 2 illustrates the distribution of continuous variables, including nitrate fertilizer consumption, agricultural land rental value, age, and farming experience. Variables previously utilized in the technical efficiency calculations, namely cultivated area, seed quantity, labor input, and tractor usage, were intentionally excluded from this visualization. According to Fig. 2, nitrate fertilizer consumption exhibits a pronounced peak at 250 kilograms, which may be attributable to the presence of outlier observations. Such outliers can undermine the validity of statistical tests and potentially compromise the reliability of machine learning model outcomes. Therefore, it is essential to identify these observations and mitigate their adverse effects. To this end, Fig. 3 employs box plots for outlier detection. For instance, the box plot for land rental value reveals that most observations fall below 10 million Iranian Tomans, with 8 outliers appearing above this threshold. The nitrate fertilizer variable contains 22 outlier observations, whereas age and farming experience display relatively uniform distributions without significant outliers. Following the identification of outliers in certain variables, this study addresses the issue through data standardization. Specifically, standardization was performed by subtracting the mean from each observation and subsequently dividing the result by the variable's standard deviation.

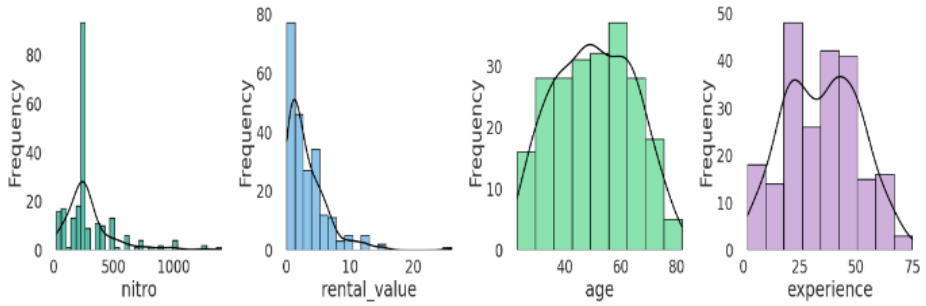

Figure 2- Probability distribution of continuous variables

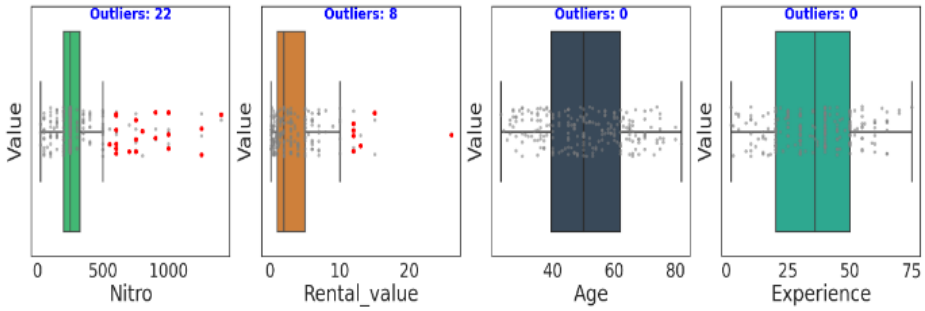

Figure 3- Box plot to identify outliers

Given that one of the primary objectives during the data preprocessing stage is the selection of appropriate variables for the final model, it is essential at this juncture to determine which continuous variables exert a significant influence on the target variable (i.e., classification into high-efficiency versus low-efficiency farmers). To this end, Student's *t*-test was employed to compare the means between the two groups, with results reported in Table 4. According to Table 4, incorporating the continuous variables into machine learning models is statistically justified. For instance, regarding nitrate fertilizer usage, the mean for group zero (low-efficiency farmers) equals 223 kilograms, whereas the mean for group one (high-efficiency farmers) amounts to 445 kilograms. This substantial difference suggests that variations in nitrate fertilizer application may indeed influence the status of the target variable, thereby supporting its inclusion as a predictive feature in the classification models.

Descriptive statistics for countable variables, including the number of farm plots, household members, and household members with university education, are reported in Table 5. The mean number of farm plots equals 11 with a standard deviation of 9, indicating substantial variability in land fragmentation among farmers. The average household size is 5 members with a standard deviation of 2, suggesting relative stability in this variable. Regarding household members with university education, a mean of 0.6 and a standard deviation of 0.9 reveal that only a subset of households includes university-educated members. The minimum and maximum values of these variables further confirm considerable dispersion in both household structure and landholding patterns across the sample.

Table 4- T-test results between continuous and target variables

Variable	t-statistic	P-value	Mean of group 0	Mean of group 1	Result
Nitrate fertilizer (kg)	-6.96	7.04	223	455	Reject
Rental value of lands (million toman)	-7.76	1.33	1.9	5.8	Reject
Age (year)	-2.99	0.003	48	54	Reject
Experience (year)	-3.19	0.001	32	40	Reject

Table 5- Descriptive statistics of countable variables

Variable	Mean	Std.dev	Minimum	Maximum
Farm plots (plot)	11	9	1	40
Familly members (person)	5	2	1	12
Familly members with university education (person)	0.6	0.9	0	4

Subsequently, similar to the continuous variables, the influence of countable variables on the target variable was examined using Student's *t*-test, with results presented in Table 6. The null hypothesis of this test posits no significant difference in the means between the two farmer groups i.e., no significant effect of the countable variable on the target variable.

According to the results in Table 6, the null hypothesis is rejected for the variables "number of farm plots" and "household members with university education," indicating statistically significant differences between high- and low-efficiency farmer groups. Consequently, these two variables were incorporated into the machine learning modeling process. Furthermore, box plot visualizations for these variables revealed the presence of outliers; therefore, standardization was applied to mitigate their potential adverse effects on model performance.

Table 6- T-test results between countable and target variables

Variable	T-statistic	P-value	Mean of group 0	Mean of group 1	Result
Farm plots (plot)	-11.05	0.00	7.20	20.09	Reject
Family members (person)	-1.20	0.22	5.12	5.49	Accept
Family members with university education (person)	-2.65	0.00	0.46	0.84	Reject

Fig. 4 illustrates the distribution of farmers across nominal variables. Regarding residence location, the majority of observations correspond to farmers residing in rural villages. Most farmers source their seeds from their own farms, underscoring the significance of local resources in seed provision. A substantial proportion of farmers cultivate wheat on lands fully owned by themselves. The chart for the harvest method variable clearly indicates farmers' pronounced yield for mechanized harvesting techniques. Furthermore, the majority of farmers employed mowers for wheat harvesting. A primary reason for this choice is the steep slope of certain fields combined with the mower's superior performance on uneven terrain, which enhances harvest efficiency. Additionally, the relatively small size of farms and the impracticality of deploying larger combine harvesters render the mower a more feasible option. The mower's adaptability to diverse field conditions and its more effective operation during harvest, particularly in farms with limited area, enables farmers to improve both productivity and harvest quality.

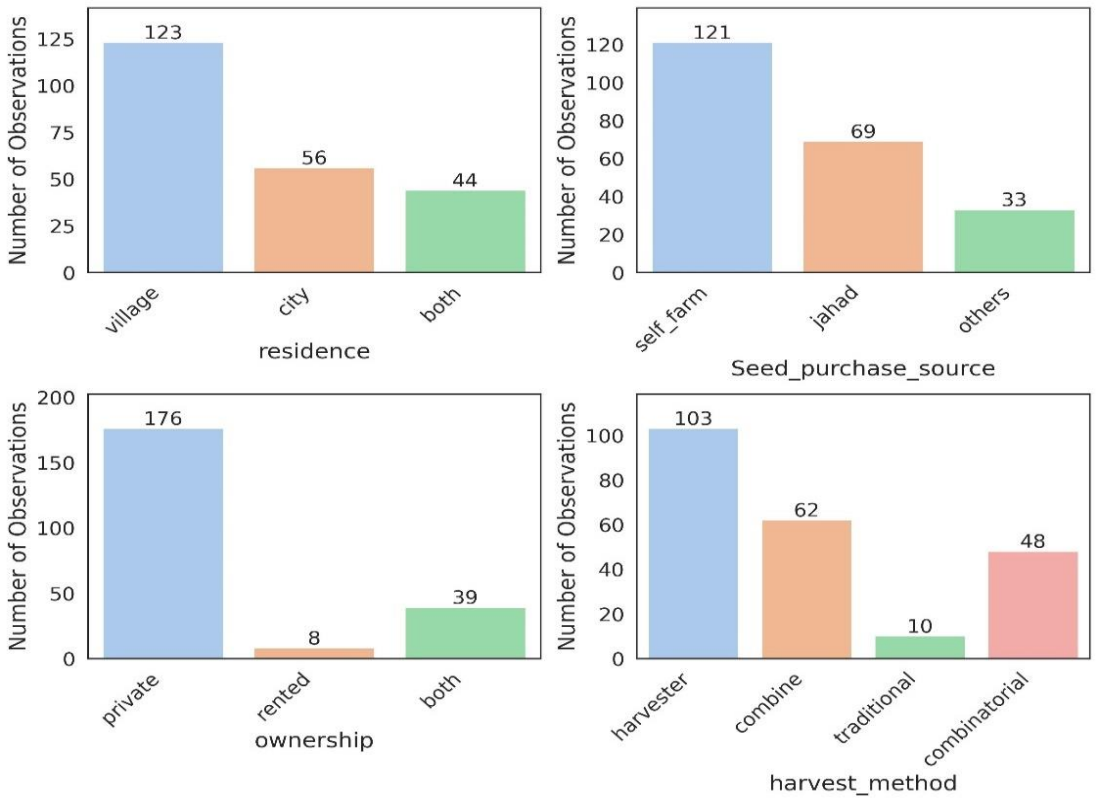

Figure 4- Distribution of farmers based on nominal variables

To incorporate nominal variables into machine learning models, it is necessary to convert them into dummy (binary) variables. Specifically, each category within a nominal variable must be transformed into a separate binary variable. Consequently, three dummy variables were created for each of the following nominal variables: residence location, seed source, and land ownership type. For the harvest method variable, four dummy variables were designed. Additional dummy variables included indicators for performi

ng contour plowing (perpendicular to slope), having off-farm income, and the use (or non-use) of certain variable inputs such as manure, herbicides, and pesticides.

To select dummy variables with a statistically significant influence on the target variable, a chi-square test was conducted, with results presented in Table 7. For instance, the null hypothesis for the rural–urban residence variable was accepted at a *P*-value of 0.07, indicating that residence location does not exert a statistically significant effect on the target variable. Overall, based on Table 7, the null hypothesis of the chi-square test was accepted for the following variables: rural–urban residence, urban residence, seed source (others), seed source (self-produced), land ownership (personal), land ownership (leased), harvest method (combine), harvest method (mower), harvest method (traditional), and contour plowing. In other words, these variables do not demonstrate a statistically significant impact on the target variable and were therefore excluded from the final modeling stage performed by the machine learning algorithms.

Table 7- Results of the Chi-square test between dummy and target variables

Variable	Chi-square statistic	P-value	Result
Residence_both	3.09	0.07	Accept
Residence_city	1.59	0.20	Accept
Residence_village	7.02	0.00	Reject
Seed_purchase_source_jahad	13.52	0.00	Reject
Seed_purchase_source_others	1.76	0.18	Accept
Seed_purchase_source_self_farm	5.39	0.02	Accept
Ownership_both[1]	6.39	0.01	Reject
Ownership_private	3.20	0.07	Accept
Ownership_rented	0.73	0.39	Accept
Harvest_method_combinatorial	5.55	0.01	Reject
Harvest_method_combine	0.79	0.37	Accept
Harvest_method_harvester	0.12	0.72	Accept
Harvest_method_traditional	1.49	0.22	Accept
Vertical_plow	2.10	0.14	Accept
Nonfarm_income	7.45	0.00	Reject
Manure	6.86	0.00	Reject
Herb	15.12	0.00	Reject
Pest	22.00	2.72	Reject

Ordinal variables in this study include education level, timeliness of agricultural operations, input availability, land slope, weed control practices, and the extent of damage caused by the sunn pest (Eurygaster integriceps). Fig. 5 illustrates the distribution of farmers across these ordinal variables. According to the figure, the reluctance of individuals with higher educational attainment to continue farming is clearly evident. Conversely, the majority of farmers carry out agricultural operations in a timely manner, reflecting a strong commitment to adhering to appropriate scheduling in farm management practices. The chart depicting input availability reveals considerable disparities in farmers' access to essential agricultural resources. According to the weed control distribution, the largest number of farmers, 134 observations fall into the "low" category, indicating that a substantial proportion of farmers do not implement effective practical measures for weed management. Finally, the graph illustrating sunn pest damage shows that only 71 surveyed individuals reported no damage from this pest, while the remaining farmers experienced varying degrees of crop loss attributable to sunn pest infestation.

Table 8 presents the results of the chi-square test conducted to select ordinal variables with a significant influence on the target variable's status. According to Table 8, weed control emerges as the sole ordinal variable that will be incorporated into the machine learning models. As previously mentioned in the Materials and Methods section, the present study employed four algorithms: logistic regression, support vector machine (SVM), k-nearest neighbors (KNN), and random forest (RF). In the initial step, considering default hyperparameters and utilizing 5-fold cross-validation, the accuracy of the models was compared, with results reported in Table 9. Here, the average accuracy of each algorithm is displayed alongside accuracies corresponding to different values of *k*. Logistic regression achieved an average accuracy of 88.7%, establishing it as the top-performing algorithm at this stage.

1- In the study area, it is common practice for an individual, in addition to cultivating land they own, to also rent and cultivate several other plots if possible.

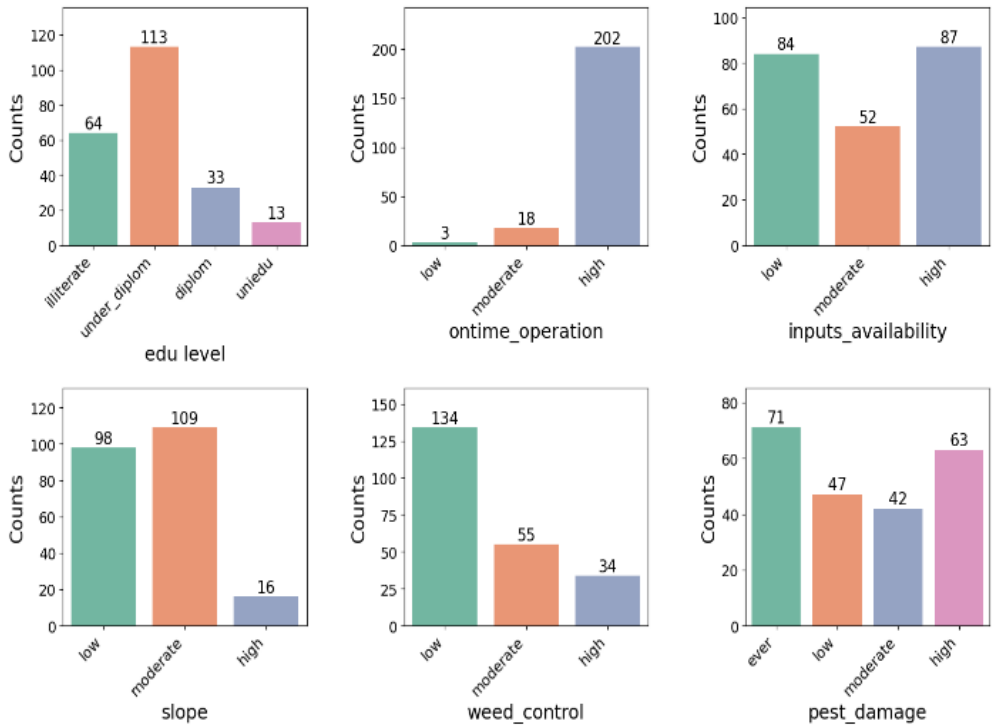

Figure 5- Distribution of farmers based on ordinal variables

Table 8- Chi-square test between ordinal and target variables

Variable	Chi-square statistic	P-value	Result
Edu level	2.20	0.53	Accept
Ontime_operation	20.47	3.58	Accept
Inputs_availability	1.45	0.48	Accept
Slope	4.84	0.08	Accept
Weed_control	5.81	0.05	Reject
Pest_damage	5.77	0.12	Accept

Table 9- The comparison of the algorithms based on accuracy statistics

Algorithm	K=1	K=2	K=3	K=4	K=5	Average accuracy
LogisticRegression	0.844	0.866	0.933	0.909	0.886	88.7%
SVC	0.822	0.866	0.911	0.886	0.886	87.4%
KNN	0.888	0.822	0.888	0.886	0.909	87.7%
Random Forest	0.800	0.866	0.888	0.954	0.909	88.3%

To further validate the employed algorithms, they were re-evaluated using various hyperparameters, and optimal parameters for each algorithm were determined using the GridSearchCV method. The resulting outcomes are presented in Table 10. This table includes the best configurations and the highest accuracy achieved by each algorithm. Overall, Table 10 highlights the importance of proper hyperparameter selection in enhancing algorithm accuracy and the overall performance of machine learning models. Each algorithm yielded different results with its specific settings, which can aid in selecting the most suitable algorithm. The Logistic Regression algorithm, with hyperparameter C=1 and maximum iterations (max_iter=10000), achieved an accuracy of 88.7%, indicating superior performance compared to the other algorithms.

Table 10- GridSearchCV results for determining optimal hyperparameters

Algorithm	Highest accuracy	The best hyperparameters
LogisticRegression (max-iter=10000)	0.887	C=1
SVC	0.886	C=5, kernel=sigmoid
KNN	0.879	N_beighbors=5
RandomForestClassifier (random-state=0)	0.883	N_estimators=50

Furthermore, to substantiate the superiority of the selected algorithm, the confusion matrix results are presented in Fig. 6. As observed, the Logistic Regression algorithm correctly classified 26 individuals within the low-efficiency farmer group as low-efficiency farmers, misclassifying only 2 individuals into the high-efficiency farmer group. With regard to high-efficiency farmers, the algorithm accurately predicted 15 individuals as high-efficiency farmers while erroneously assigning 2 individuals to the low-efficiency group. The Support Vector Machine (SVM) algorithm misclassified only one individual in the low-efficiency farmer category, whereas the K-Nearest Neighbors (KNN) algorithm correctly assigned all 28 individuals to their true low-efficiency class. However, concerning the high-efficiency farmer group, both algorithms exhibited suboptimal performance, each misclassifying 6 individuals. Finally, the Random Forest algorithm produced one misclassification in the low-efficiency group and four misclassifications in the high-efficiency group. Overall, analysis of the confusion matrix corroborates the relative superiority of the Logistic Regression algorithm compared to the other evaluated classifiers.

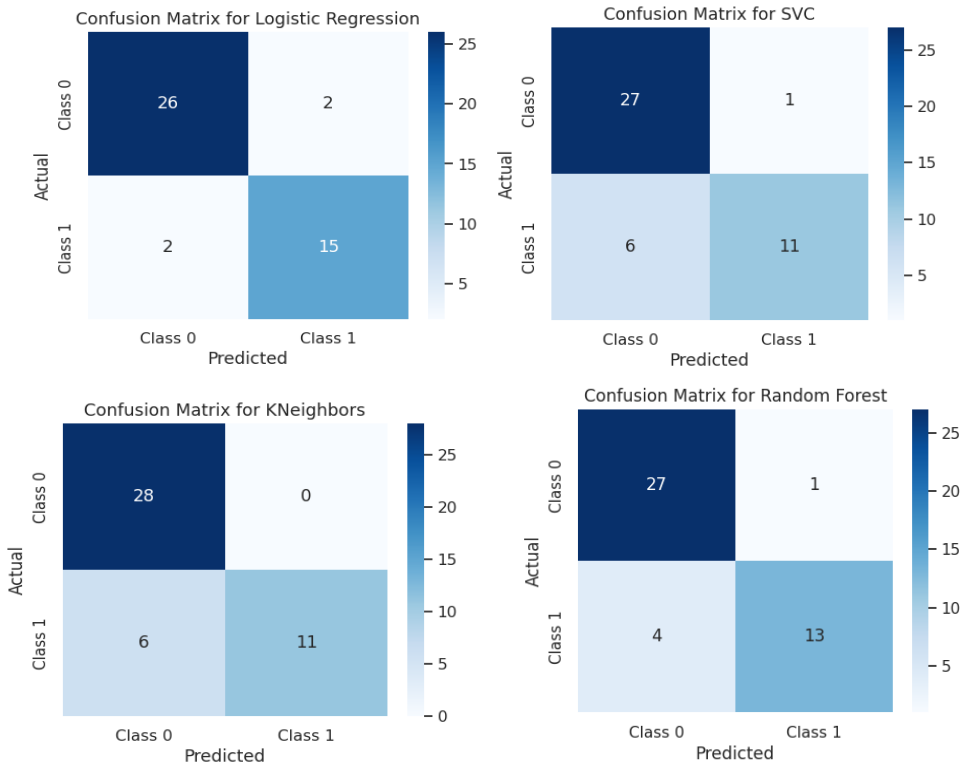

Figure 6- Confusion matrices from the research algorithms

Table 11 presents a comparative analysis of the prediction performance of various machine learning algorithms, reporting distinct metrics for each algorithm across the two classes of

low-efficiency and high-efficiency farmers. To this end, the evaluation metrics, including precision, recall, F1-score, accuracy, and Cohen's Kappa, were computed based on the confusion matrix. According to the results in Table 11, the Logistic Regression algorithm demonstrates superior predictive performance overall compared to the other algorithms under consideration.

Table 11- Comparing the predictability of machine learning algorithms

Algorithm	Class	Precision	Recall	F1-score	Accuracy	Cohen's Kappa
LR	Non-efficient	0.93	0.93	0.93	0.91	0.81
	Efficient	0.88	0.88	0.88		
SVC	Non-efficient	0.82	0.96	0.89	0.84	0.64
	Efficient	0.92	0.65	0.76		
KNeighbors	Non-efficient	0.82	1.00	0.90	0.86	0.69
	Efficient	1.00	0.65	0.79		
RF	Non-efficient	0.87	0.96	0.92	0.88	0.75
	Efficient	0.93	0.76	0.84		

Additionally, to compare the predictive power of the algorithms, ROC curves can be plotted as illustrated in Fig. 7. The horizontal axis of the plot represents the false positive rate (FPR), while the vertical axis denotes the true positive rate (TPR). According to Fig. 7, an algorithm whose ROC curve approaches the coordinate (1,1) more closely and exhibits a higher area under the curve (AUC) possesses superior predictive capability. The Logistic Regression algorithm, with an AUC of 0.98, demonstrates the best performance in identifying positive cases, thereby outperforming the other algorithms. The Support Vector Machine (SVM) algorithm, with an AUC of 0.95, ranks second, indicating high accuracy, albeit marginally lower than that of Logistic Regression.

The K-Nearest Neighbors and Random Forest algorithms follow in subsequent ranks, with AUC values of 0.91 and 0.93, respectively. Finally, the presence of a 45-degree diagonal line representing random chance underscores that algorithms exhibiting curves above this baseline demonstrate superior capability in identifying positive cases. In this context, Logistic Regression exhibits notably superior performance, with its ROC curve positioned distinctly above the diagonal reference line.

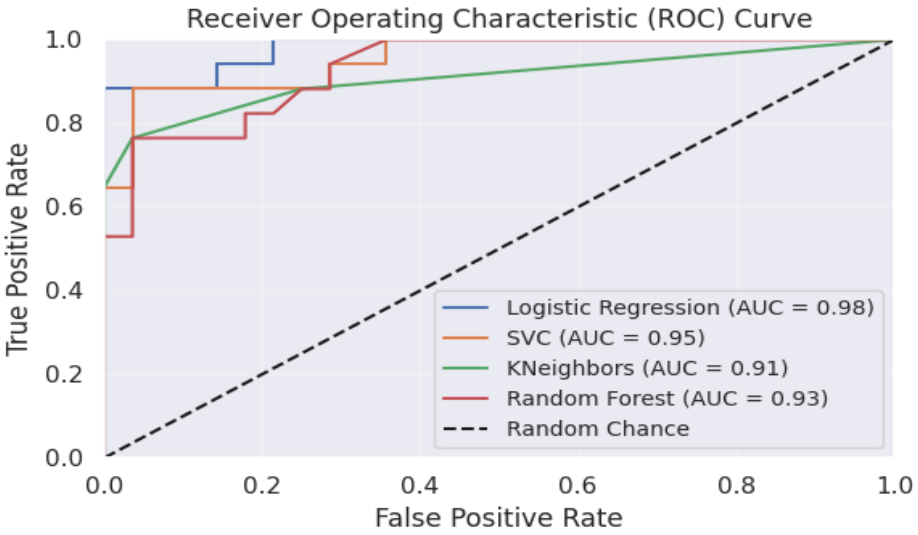

Figure 7- Comparison of the performance of machine learning algorithms using the ROC curve

Table 12 employs the Logistic Regression algorithm to examine the factors influencing the

technical efficiency of rain-fed wheat producers in Ahar County. The table comprises two principal columns presenting the regression coefficients and corresponding odds ratios for each explanatory variable. According to the results, variables including herbicide application, weed control practices, manure usage, rental value of wheat cropland, nitrate fertilizer application, number of farm plots, pesticide use, farmer age, combined harvesting method, farming experience, presence of household members with university education, seed procurement from the Agricultural Jihad Organization, and rural residence exert a positive and statistically significant effect on technical efficiency levels. Conversely, mixed ownership status of cropland (personal-rental combination), seed procurement from personal sources, and possession of non-farm income demonstrate negative effects on technical efficiency.

Given that the estimated coefficients in this logistic regression model lack direct intuitive interpretation, odds ratios were computed to facilitate meaningful inference. For instance, the odds ratio of 2.927 for herbicide application indicates that farmers who utilize herbicides are approximately three times more likely to achieve improvements in technical efficiency compared to those who do not. Optimal herbicide application at various crop growth stages effectively suppresses weed competition, thereby reducing resource rivalry with wheat plants and enhancing both yield quantity and grain quality. Effective weed management enables wheat plants to more efficiently absorb rainfall and soil nutrients, ultimately contributing to increased productivity and improved technical efficiency. The application of manure, as a natural source of essential nutrients enhances soil structure and fertility, supporting the nutritional requirements of wheat throughout its growth cycle. A higher rental value of wheat cropland may reflect superior soil quality and greater production potential, while also enabling the adoption of modern agronomic and managerial practices that positively influence technical efficiency. Nitrate fertilizer application contributes to elevated soil nitrogen levels, promoting robust plant growth and higher grain yields. Furthermore, farmers cultivating multiple land plots can better mitigate risks associated with adverse weather conditions or pest outbreaks by diversifying spatial exposure and implementing more flexible, optimized strategies for planting and harvesting operations.

The application of pesticides enables farmers to mitigate pest damage to crops, thereby enhancing agricultural productivity. Nevertheless, to ensure long-term sustainability, adopting environmentally conscious approaches, such as Integrated Pest Management (IPM), is essential to minimize adverse ecological impacts while maintaining effective pest control.

Farmer age can serve as a proxy for accumulated agricultural experience in cultivation and farm management practices. Older farmers typically possess greater experiential knowledge in managing agricultural operations and selecting effective agronomic techniques, which may contribute to enhanced technical efficiency. Agricultural experience enables farmers to make more informed decisions regarding optimal farming practices and to utilize resources more efficiently. Experienced farmers tend to adopt superior strategies for addressing agronomic challenges such as adapting to variable climatic conditions, optimizing input allocation, and implementing timely interventions, thereby fostering improvements in technical efficiency. This accumulated practical knowledge, developed over years of engagement with farming systems, represents a valuable intangible asset that positively influences production outcomes and operational performance.

The adoption of combined harvesting methods can enhance efficiency throughout the crop harvesting process. Such approaches enable farmers to optimize the utilization of time and resources, resulting in greater crop recovery and elevated technical efficiency. The presence of university-educated household members within farming families can contribute to the advancement of agricultural

management knowledge and skills. These individuals, leveraging their formal education and expertise, can assist in refining farming strategies and decision-making processes, thereby fostering improvements in technical efficiency. Procurement of seeds from reputable agricultural institutions, such as the Agricultural Jihad Organization, facilitates access to high-quality, pest-resistant seed varieties. This practice directly contributes to enhanced crop performance and yield stability through improved genetic traits and reduced vulnerability to biotic stresses. Residence in rural areas serves as a significant enabler for the timely execution of agronomic operations. Proximity to farmland allows farmers to respond promptly to critical cultivation windows, while also facilitating greater accessibility to local resources, extension services, and emerging agricultural technologies, all of which collectively support the enhancement of technical efficiency in farming systems.

Regarding factors exerting negative influences on technical efficiency, mixed land ownership (personal and rental) warrants particular attention. Under this arrangement, farmers may exhibit diminished motivation to invest in long-term land improvement or optimize resource utilization. This disincentive can lead to reduced adoption of modern cultivation practices and insufficient attention to proper land stewardship, ultimately exerting an adverse effect on technical efficiency. Seed procurement from personal or informal local sources often results in the use of low-quality seeds that are more susceptible to pests and diseases. Such suboptimal genetic material typically yields lower productivity and compromises crop resilience, thereby contributing to diminished output and reduced technical efficiency. Possession of non-farm income sources may divert farmers' focus away from agricultural activities, potentially reducing their commitment to farming operations and limiting investment in agronomic improvements. Farmers with supplementary income streams may, due to reduced reliance on agricultural earnings, demonstrate lower motivation to optimize production processes or adopt innovative farming techniques. This diminished engagement can consequently lead to suboptimal resource allocation and a decline in technical efficiency within wheat production systems.

Table 12- Factors affecting the technical efficiency of farmers producing rainfed wheat in Ahar County

Feature	Coefficient	Odds ratio	Feature	Coefficient	Odds ratio
herb	**1.074**	**2.927**	**Harvest_method_combinatorial**	**0.156**	**1.169**
Weed_control	0.895	2.448	Experience	0.142	1.153
Manure	0.827	2.286	Fam_mem_edu	0.096	1.101
Rental_value	0.505	1.658	Seed_purchase_score_jahad	0.088	1.092
Nitro	0.440	1.554	Residence_village	0.034	1.035
Farm_plots	0.380	1.462	Ownership_both	-0.251	0.777
Pest	0.263	1.301	Seed_purchase_source_self_farm	-0.306	0.736
Age	0.162	1.176	Nonfarm_income	-0.856	0.424

Fig. 8 presents a ranking of the relative influence exerted by each explanatory variable within the logistic regression model developed to analyze the technical efficiency of rain-fed wheat producers. The horizontal axis represents the magnitude and direction of the estimated coefficients, with positive values indicating factors that enhance efficiency, and negative values denoting those that diminish it. The vertical axis enumerates the various determinants of farmer efficiency. The length of each bar extending horizontally from the central axis reflects the strength of a given variable's impact on technical efficiency: longer bars to the right signify stronger positive influences, whereas longer bars to the left indicate more pronounced negative effects. As illustrated in the figure, herbicide application demonstrates the strongest positive effect on technical efficiency, while rural residence exhibits the weakest positive influence among the favorable factors. Conversely, mixed land

ownership status (combining personal and rental plots) exerts the most substantial negative impact, whereas the presence of non-farm income sources displays the least detrimental, though still negative effect on technical efficiency.

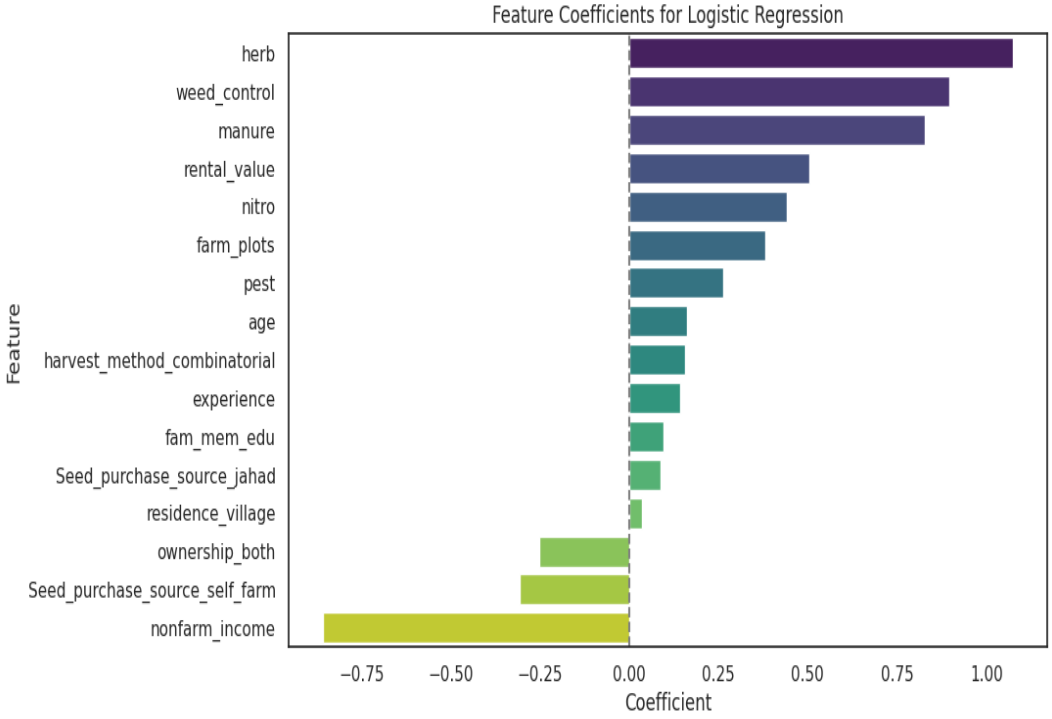

Figure 8- Examining the importance of variables that affect farmers' technical efficiency

Conclusion

This study was conducted with the objective of applying data mining techniques to analyze factors influencing the classification of rain-fed wheat producers in Ahar County based on their technical efficiency. The principal innovation of this research lay in the implementation and comparison of various machine learning algorithms to address an agricultural economics problem. The study demonstrated that integrating Data Envelopment Analysis (DEA) with machine learning algorithms can provide a powerful analytical framework for efficiency-related issues in the agricultural sector.

Subsequently, to predict farmers' efficiency and analyze its determinants, statistical methods were first employed to identify variables exhibiting significant associations with the target variable. Following the identification of these variables, machine learning algorithms, including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Random Forest, were utilized to predict farmers' technical efficiency. Comparison of these algorithms through various performance evaluation metrics revealed that each method possesses distinct strengths and limitations in analyzing agricultural data. Model estimation and examination of results derived from the confusion matrix, accuracy, and error evaluation metrics, and ROC curves indicated that the Logistic Regression algorithm demonstrates superior predictive capability for farmers' efficiency compared to the other algorithms. This advantage primarily stems from its simpler structure, higher interpretability, and favorable performance with datasets of limited sample size.

Results obtained from the Logistic Regression algorithm indicated that the technical efficiency of rain-fed wheat producers in Ahar County is influenced by two principal groups of factors: those exerting positive effects

and those exerting negative effects. Variables with positive impacts include herbicide application, weed control practices, manure usage, rental value of wheat cropland, nitrate fertilizer application, number of farm plots, pesticide use, farmer age, combined harvesting method, farming experience, presence of household members with university education, seed procurement from the Agricultural Jihad Organization, and rural residence. The second group comprises variables with negative effects, namely mixed ownership status of cropland (personal-rental combination), seed procurement from personal sources, and possession of non-farm income.

Findings showed that both positive and negative influences of various factors on the technical efficiency of rain-fed wheat producers in Ahar County, and given the efficacy of data mining techniques in identifying these determinants, several comprehensive and actionable management recommendations can be proposed. In particular, the optimal application of herbicides, pesticides, and appropriate fertilizers such as manure and nitrate fertilizer can significantly enhance farmers' technical efficiency. This finding is consistent with the results reported by Aryan *et al*. (2024). Therefore, it is recommended that educational and extension programs be implemented for farmers to enhance their knowledge and promote the optimal utilization of these agricultural inputs. In this context, data mining-based decision support systems can further assist farmers in selecting the optimal input combinations. The obtained results are consistent with the findings of Tanursaz *et al*. (2021) and Ghasemi *et al*. (2023). Additionally, emphasis should be placed on procuring seeds from certified and reputable sources, such as the Agricultural Jihad Organization, to improve sowing quality and ultimately enhance the productivity of rain-fed wheat farms. This finding also aligns with the results reported by Samson *et al*. (2024). On the other hand, factors such as mixed land tenure arrangements (combining personal ownership with rental plots) and the procurement of seeds from informal personal sources may place farmers in the lower efficiency frontier. Policymakers are therefore advised to enhance these farmers' motivation to improve technical efficiency through financial incentives, such as input and equipment purchase discounts targeted specifically at cultivators operating on leased agricultural farms. Enhancing access to certified seeds through the establishment of seed distribution centers in rural areas presents a viable opportunity for farmers to conveniently procure high-quality seeds at reasonable prices. Furthermore, fostering effective collaborations with seed-producing companies can guarantee both the quality and continuous supply of these seeds, ultimately contributing to enhanced agricultural productivity. In conclusion, this study clearly demonstrates that data mining approaches can serve as a powerful analytical tool in agricultural policymaking. By generating profound insights from existing datasets, such approaches can significantly improve decision-making processes across various managerial levels. For future research, it is recommended to employ more sophisticated deep learning algorithms as well as clustering techniques to enable a more precise analysis of efficiency patterns.

References

1. Aryan, S., Gulab, G., Hashemi, T., Habibi, S., Kakar, K., Habibi, N., & Zerak, A. (2024). Pre-spike emergence nitrogen fertilizer application as a strategy to improve floret fertility and production efficiency in wheat. *Field Crops Research*, *319*, 109623. https://doi.org/10.1016/j.fcr.2024.109623
2. Bahiraie, A., Hamedi, R., Alinia, H., & Sanayei, Y. (2020). Modeling the efficiency of banks by cover data method and Genetic programming. *Islamic Economics and Banking*, *9*(31), 69-98. (In Persian). http://mieaoi.ir/article-1-960-en.html
3. Bermudez, R.S., Manalang, J.O., Gerardo, B.D., & Tanguilig, B.T. (2011). Predicting faculty performance using regression model in data mining. In 2011 Ninth International Conference on Software Engineering Research, Management and Applications: 68-72. IEEE. https://doi.org/10.1109/SERA.2011.29

4. Cooper, W.W., Seiford, L.M., & Tone, K. (2007). *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*, 2, 489. New York: springer. https://doi.org/10.1007/978-0-387-45283-8
5. Davoudniya, Z., Hashemibonab, S., & Molaei, M. (2023). Investigating environmental and technical efficiency in grape production in Bijar County with input-oriented approach. *Agricultural Science and Sustainable Production*, *33*(3), 289-302. (In Persian). https://doi.org/10.22034/saps.2022.50319.2826
6. East Azerbaijan Agricultural Jihad Organization. (2022). Agricultural statistics yearbook 2021. https://www.eaj.ir/
7. Ghasemi, E., Dashti, G., & Vahedi, J. (2023). Technical efficiency, environmental efficiency and economic losses of rainfed wheat production in Ahar County. *Agricultural Economics*, *17*(1), 1-20. (In Persian). https://doi.org/10.22034/IAES.2022.1971035.1954
8. Girma, M., Mehare, A., & Ketema, M. (2024). Wheat crop producers' technical efficiency and its determinants in Oromia region of Ethiopia: evidence from West Shewa zone. *Cogent Social Sciences*, *10*(1), 2329791. https://doi.org/10.1080/23311886.2024.2329791
9. Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia*, *62*, 18-22. https://doi.org/10.1016/j.phpro.2015.02.005
10. Hackeling, G. (2017). *Mastering Machine Learning with Scikit-learn*. Packt Publishing, Birmingham.
11. Hordri, N.F., Yuhaniz, S.S., Azmi, N.F.M., & Shamsuddin, S.M. (2018). Handling class imbalance in credit card fraud using resampling methods. *International Journal Advanced. Computer Science Apply*, *9*(11), 390-396. https://doi.org/10.14569/IJACSA.2018.091155
12. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical earning; Springer Texts in Statistics; Springer New York: New York, NY; Vol. 103. https://doi.org/10.1007/978-1-4614-7138-7
13. Khan, B., Naseem, R., Muhammad, F., Abbas, G., & Kim, S. (2020). An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. *IEEE Access*, *8*, 55012-55022. https://doi.org/10.1109/ACCESS.2020.2981689
14. Khodaverdizadeh, M., Mohammadi, M., & Miri, D. (2019). Estimation of technical efficiency of wheat production with emphasis on sustainable agriculture in Urmia county. *Journal of Agricultural Science and Sustainable Development*, *29*(4), 233-245. (In Persian)
15. McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1), 1-12. https://doi.org/10.5121/mlaij.2015.2101
16. Ngai, E.W., Xiu, L., & Chau, D.C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021
17. Naruei, H., Ahmadpour Borazjani, M., Salarpour, M., Keikha, A., & Esfanjari Kenari, R. (2024). Impact of adopting strategies to cope with climate change on the technical efficiency of wheat farmers in Sistan Region-Iran. *Journal of Agricultural Economics and Development*, *38*(2), 195-208. (In Persian). https://doi.org/10.22067/jead.2024.87331.1259
18. Otunaiya, K.A., & Muhammad, G. (2019). Performance of datamining techniques in the prediction of chronic kidney disease. *Computer Science and Information Technology*, *7*(2), 48-53. https://doi.org/10.13189/csit.2019.070203
19. Ranjan, G.S.K., Verma, A.K., & Radhika, S. (2019). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. *5th international conference for convergence in technology (I2CT)* (pp. 1-5). IEEE. . https://doi.org/10.1109/I2CT45611.2019.9033691
20. Samson, G.L., Haruna, U., & Idi, S. (2024). Analysis of wheat production efficiencies in Hadejia and Ringim local government areas of Jigawa state, Nigeria. *Nigerian Journal of Agriculture and Agricultural Technology*, *4*(2), 203-213. https://doi.org/10.59331/njaat.v4i2.705
21. Sen, P.C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph*, 99-111. Springer Singapore. https://doi.org/10.1007/978-981-13-7403-6-11
22. Shetty, S.H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised machine learning: algorithms and applications. *Fundamentals and methods of machine and deep learning: algorithms, tools and applications*, 1-16. https://doi.org/10.1002/9781119821908.ch1

23. Shu, X., & Ye, Y. (2023). Knowledge discovery: Methods from data mining and machine learning. *Social Science Research, 110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817
24. Sujatha, P., & Mahalakshmi, K. (2020). Performance evaluation of supervised machine learning algorithms in prediction of heart disease. *International conference for innovation in technology (INOCON)* IEEE, 1-7. https://doi.org/10.1109/INOCON50539.2020.9298354
25. Tanursaz, A., Bakhshoodeh, M., & Azarm, H. (2021). The effects of conservation tillage on technical efficiency of wheat growers in Dezful county. *Journal of Agricultural Science and Sustainable Production, 31*(1), 331-348. (In Persian). https://doi.org/10.22034/saps.2021.12819
26. Vahedi, J., Niazifar, M., Ghahremanzadeh, M., Taghizadeh, A., Abachi, S., Palangi, V., & Lackner, M. (2024). Predicting livestock farmers' attitudes towards improved sheep breeds in Ahar city through data mining methods. *World, 5*(4), 848-864. https://doi.org/10.3390/world5040044
27. Warrens, M.J. (2008). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification, 25*, 177-183. https://doi.org/10.1007/s00357-008-9023-7
28. Xu, Q., & Yin, J. (2021). Application of random forest algorithm in physical education. *Scientific Programming, 2021*(1), 1996904. https://doi.org/10.1155/2021/1996904
29. Yuvalı, M., Yaman, B., & Tosun, Ö. (2022). Classification comparison of machine learning algorithms using two independent CAD datasets. *Mathematics, 10*(3), 311. https://doi.org/10.3390/math10030311

نشریه اقتصاد و توسعه کشاورزی
https://jead.um.ac.ir

مقاله پژوهشی
جلد ۴۰، شماره ۲، تابستان ۱۴۰۵، ص. ۱۲۷-۱۰۳

کاربرد داده‌کاوی در تحلیل عوامل مؤثر بر طبقه‌بندی کارایی فنی گندم‌کاران: مطالعه‌ای در شهرستان اهر

جبرئیل واحدی[۱]- محمد قهرمان‌زاده[۱*]- قادر دشتی[۱]- پریسا پاکروح[۲]

تاریخ دریافت: ۱۴۰۴/۰۲/۱۳
تاریخ پذیرش: ۱۴۰۴/۰۹/۱۱

چکیده

شهرستان اهر به‌عنوان یکی از مناطق اصلی تولید محصولات کشاورزی، بخش قابل‌توجهی از گندم دیم استان آذربایجان شرقی را تولید می‌کند. بهبود کارایی فنی در این منطقه می‌تواند تأثیر قابل‌توجهی بر بهره‌وری و پایداری اقتصادی کشاورزان داشته باشد. بنابراین مطالعه حاضر با هدف کاربرد داده‌کاوی در تحلیل کارایی کشاورزان گندم دیم در شهرستان اهر انجام شد. بدین منظور کارایی فنی کشاورزان با استفاده از روش تحلیل پوششی داده‌ها محاسبه گردید و کشاورزانی که کارایی بالاتر از میانگین منطقه (یعنی ۰/۵۹) داشتند، به‌عنوان گروه کشاورزان با کارایی بالاتر از میانگین و بقیه به‌عنوان گروه کشاورزان با کارایی پایین‌تر از میانگین طبقه‌بندی شدند. سپس، با استفاده از آزمون‌های تی و خی‌دو، متغیرهایی که احتمال می‌رفت بر کارایی اثرگذار باشند، شناسایی شدند و برای طبقه‌بندی و تحلیل کارایی فنی از الگوریتم‌های یادگیری ماشین شامل رگرسیون لجستیک، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه و جنگل تصادفی استفاده گردید. نتایج نشان داد که رگرسیون لجستیک عملکرد بهتری نسبت به سایر الگوریتم‌ها دارد. خروجی این الگوریتم حاکی از آن بود که عواملی مانند علف‌کش، کنترل علف‌های هرز، کود حیوانی، ارزش اجاره‌ای زمین‌ها، کود نیتراته، تعداد قطعه زمین‌ها، آفت‌کش، سن کشاورز، روش برداشت ترکیبی، تجربه، اعضای خانوار با تحصیلات دانشگاهی، تأمین بذر از جهاد کشاورزی و سکونت در روستا دارای تأثیر مثبت بر کارایی فنی می‌باشند. عواملی مانند، مالکیت زمین‌ها به‌صورت شخصی-اجاره‌ای، تأمین بذر از منابع شخصی و درآمد غیرکشاورزی نیز تأثیر منفی بر کارایی دارند. بر اساس نتایج، توصیه می‌شود کشاورزان آموزش‌های لازم در خصوص مدیریت بهینه مصرف نهاده‌های کشاورزی اثرگذار شامل علف‌کش، کود حیوانی، کود نیتراته و آفت‌کش را دریافت کنند. همچنین، سیاست‌گذاران با ارائه مشوق‌های مالی و خدمات مشاوره‌ای، انگیزه بهبود کارایی را در کشاورزانی که روی زمین‌های اجاره‌ای فعالیت می‌کنند، افزایش دهند. توسعه برنامه‌های حمایتی برای تأمین بذر و نهاده‌های کشاورزی با کیفیت از طریق جهاد کشاورزی نیز پیشنهاد می‌شود.

واژه‌های کلیدی: داده‌کاوی، کارایی فنی، گندم، یادگیری ماشین

۱- گروه اقتصاد کشاورزی، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران
(*- نویسنده مسئول: Email: Ghahremanzadeh@tabrizu.ac.ir)
۲- محقق فلوشیپ ماری کوری، دانشگاه میلان، میلان، ایتالیا

https://doi.org/10.22067/jead.2025.93203.1349