Received: 2 May 2022 | Revised: 1 August 2022 | Accepted: 11 September 2022

DOI: 10.1002/agr.21773

RESEARCH ARTICLE

Agribusiness An International Journal WILEY

Forecasting drinking milk price based on economic, social, and environmental factors using machine learning algorithms

Abdulkadir Atalan

Department of Industrial Engineering, Gaziantep Islam Science and Technology University, Gaziantep, Turkey

Correspondence
Abdulkadir Atalan, Department of Industrial Engineering, Gaziantep Islam Science and Technology University, Gaziantep, Turkey.
Email: abdulkadiratalan@gmail.com

Abstract

The study aimed to describe and test machine learning (ML)-based algorithms to evaluate the unit price of drinking milk. The algorithms were applied to the data collected over 8 years in 2014 and 2021 related to the price of drinking milk in Turkey. The economic, social, and environmental factors that have an impact on the unit price of drinking milk were evaluated. Five ML algorithms, including random forest, gradient boosting, support vector machine (SVM), neural network, and AdaBoost algorithms, were utilized to predict the drinking milk unit price. ML also applied hyperparameter tuning with nested cross-validation to calculate the prediction accuracy for each algorithm. The results show that the random forest algorithm based on the features of the ML algorithms has the best performance, with the accuracy of 99.30% for training and 98.10% for testing the dataset. The average accuracy of gradient boosting, SVM, neural network, and AdaBoost are obtained as 97.30%, 96.15%, 95.65%, and 96.05%, respectively. Random forest performed best as the target variable with the lowest deviation values of mean squared error (MSE) (0.004), root mean square error (RMSE) (0.060), and mean

Abbreviations: AI, artificial intelligence; ANN, artificial neural networks; ANN-ABO, artificial neural network-artificial butterfly optimization; ANN-GA, artificial neural network-genetic algorithm; CART, classification and regression trees; EPT, transmission elasticity coefficient; GBM, gradient boosting machine; LASSO, least absolute shrinkage and selection operator; MA, model averaging; MAE, mean absolute error; ML, machine learning; MLP, multilayer perceptron; MSE, mean squared error; PLSDA, partial least squares discriminant analysis; PLSR, partial least squares regression; R, coefficient of correlation; RBF, radial basis function; RF, random forest; RMSE, root mean square error; RR, ridge regression; SVM, support vector machine; TUIK, Turkish Statistical Institute.

absolute error (MAE) (0.029) in the training and MSE (0.009), RMSE (0.096), and MA (0.055) in the testing dataset. This study presents an interesting perspective with practical potential to adopt ML methods in the dairy industry. The developed ML algorithms can provide dairy investors and policymakers with important decision-support information. [EconLit Citations: C13, C53, L66, C88].

KEYWORDS
factors, machine learning algorithms, milk price, prediction

1 | INTRODUCTION

Milk is an essential food item for human consumption in terms of health and economic dimension regarding business. Milk consumption and production are directly or indirectly related to many parameters. Researchers emphasize that many parameters should be considered since these parameters are considered in terms of social, environmental, and economic aspects. In most studies, environmental factors affecting milk production and consumption have been measured (Capper et al., 2009; Kirschbaum et al., 2017). In addition to these factors, social (demographic structure) and economic factors are discussed in this study.

In terms of social factors, the amount of milk consumption varies according to age groups, gender, and life expectancy (Costa Leite et al., 2017). From an economic point of view, milk producers need to have the most appropriate infrastructure to maintain their current production levels with the increase in milk production costs, and milk unit price is of crucial importance, especially in milk production due to milk consumption. While Turkey can buy approximately 1071 L of milk with minimum wage in 2014, a minimum wage earner can only buy 1071 L of milk in the last period of 2021. With the decrease in milk purchasing power of individuals, a reduction in milk production level is inevitable. Other parameters affecting milk production are the price increases in gasoline used for transportation and animal feed products. The unit electricity fee used in the conversion process from raw milk to drinking milk has increased due to high energy demand in recent years (Dalala et al., 2022). In short, both economic and social factors have a potential impact on the unit price of milk.

Many studies have emphasized that environmental factors play an important role in milk production, quality, and components in milk. Rhone et al. calculated that farm milk yield and average milk yield per cow are higher in winter ($p < 0.05$) and lower in summer and rainy seasons (Rhone et al., 2008). In a study dealing with a cow species, correlation values between environmental factors and milk production were calculated, and researchers emphasized that environmental factors seriously affect milk production in this study (Mylostyvyi & Chernenko, 2019). Another study examined the relationship between temperature and humidity and milk production characteristics and concluded a 16% decrease in milk production at high temperatures (July and August), but the humidity factor had little effect on milk production. This study also determined that the nutritive rate of protein and fat contents in milk at high temperatures decreased (Zhu et al., 2020). One study highlighted that milk production is related to overall environmental factors as well as barn environmental conditions and the production performance of individual cows. This study suggested that more sustainable milk production is achieved by ensuring optimum living conditions for animals, improvements in milk quality, and production efficiency (Lovarelli et al., 2020). In this study, environmental factors such as temperature and water are discussed, with the thought that these factors will impact milk production and, therefore, milk prices. Machine learning (ML) algorithms, one of the most innovative

methods of today, were utilized to statistically analyze the factors considered and obtain the target factor's estimation results in this study.

The vertical price transmission mechanism between producer and consumer prices is the most important factor in the product unit price. Fernández-Amador et al. (2010) empirically evaluated the vertical price transmission mechanism between producer and consumer prices in dairy products. In another study, approximately 6 years of data were observed in China using the Market Chain Cooperation Model for the milk price transmission mechanism (Xiaoxia et al., 2013). Another study used an error correction model to examine the milk price transmission mechanism in the Greek milk market using monthly data between 1998 and 2014 (Reziti, 2014). Regression models stand out with the effect of many parameters on the milk unit price, especially the time series. In this study, regression algorithms, which are the basis of ML algorithms, were preferred in the milk unit price formation mechanism.

ML models are types of artificial intelligence (AI) that discover patterns specific to the data and query multiple datasets to determine the relationships of parameters (Myszczynska et al., 2020). ML is an AI method that focuses on building systems that learn or improve performance based on observed data. ML models are known as the process of training and predicting a certain amount of data by learning a dataset in detail (Ongsulee, 2017). In particular, ML models play a significant role in estimating complex datasets by analyzing them quickly and efficiently (Jordan & Mitchell, 2015). Researchers have done many studies on milk and generally provided information about milk using statistical methods. Information about the methods and data types used by the researchers for milk is given in Table 1, and the difference between the proposed study and other studies is revealed. ML algorithms were used to estimate the target variable and measure the effect of the factors mentioned in milk production, concentrating on milk quantity, milk quality, milk authenticity, and milk composition in all of the studies listed in Table 1. However, this study aims to examine the factors affecting the milk unit price and estimate the future milk unit price using ML algorithms.

The most important feature that distinguishes this study from other studies is that ML models are not directly related to a food economy but are related to food sciences. Frizzarin et al. (2021) employed ML methods to predict cow milk quality characteristics from available milk spectra. Farah et al. used the combined differential scanning calorimetry from ML models to detect raw bovine milk using random forest (RF), gradient boosting machine, and multilayer sensor algorithms. The RF algorithm provided the accuracy of the estimation data of this study, and the estimation accuracy was calculated at a rate of approximately 88.5%. (Farah et al., 2021). One of the studies aimed to develop a spectroscopic sensor system using the neural network algorithms used to design ML-capable multispectral spectroscopic sensors, sample preparation, spectral data collection, their processing, and analysis for IoT Application milk adulteration identifications. The neural network algorithm provided the best estimation results (accuracy rate of 0.927) in this study (Sowmya & Ponnusamy, 2021). Mu et al. (2020) preferred ML algorithms to perform milk source identification and milk quality estimation and reached the best estimation results with the RF algorithm. Another study used the time series cross-validation model to predict the next month's daily milk yield, composition, and milking frequency (Ji et al., 2022). Studies that prefer the ML algorithm for modeling milk productivity and quality have obtained high-accuracy prediction data with artificial neural networks (ANN) (87%) and ANN-artificial butterfly optimization algorithms. Shine et al. (2018) used ML algorithms to estimate the amount of milk production by considering energy and environmental parameters (Swarup Kumar et al., 2022).

The paper proposes that the ML perspective should be used to contribute to the development of data analysis in the food economy. There is no specific ML model for calculating the predictive values of the data. Input and output variables run in ML models to obtain estimated data in this study. Measurement criteria values were calculated to compare the performance of the results of the AI models used. For this reason, it is possible to reach the forecast data by using multiple forecasting models, and the model results with the best performance are preferred. In this study, five different ML models named RF, gradient boosting, support vector machine (SVM), neural network, and AdaBoost algorithms were run to estimate the price of drinking milk. The performance measurement values of these algorithms were compared for the present paper.

TABLE 1 Related studies on drinking milk using machine learning techniques

Purpose	Method	Data	Best model and accuracy of prediction	References
Predicting cow milk quality traits	PLSR, RR, LASSO, NN, MA, PLSDA, random forest, boosting decision trees, support vector machine	Protein composition and technological trait	Support vector machine (dividing traits into two classes) % of accuracy for each factor	Frizzarin et al. (2021)
Determining the milk authenticity	Random forest, gradient boosting machine, MLP	The temperature of boiling-crystallization-melting, enthalpy of boiling-crystallization-melting, enthalpy of melting-boiling-crystallization,	Random forest, 100% (recognition) 88.5%, (prediction)	Farah et al. (2021)
An IoT application of adulteration identification on milk	Decision tree, Naive Bayes, linear discriminant analysis, support vector machine, and neural network	Generation numbers, population, chromosome size, crossover probability, mutation, selection method	Neural network, 92.7%	Sowmya and Ponnusamy (2021)
Milk Source Identification and Milk Quality Estimation	Logistic regression, support vector machine, random forest,	Gas sensor information	Rrandom forest, 93.99% for milk fat; 93.01% for milk protein	Mu et al. (2020)
Predicting the next month's daily milk yield, milk composition, and milking frequency	Time series cross-validation	Environment conditions, individual animal's behaviors, health, productivity, and milk quality	Time-series cross-validation, 80%	Ji et al. (2022)
Modeling Milk Productivity and Quality	Artificial neural networks, Bayesian Regularization training	Cow data and daily environmental parameters	Artificial neural networks 87% for model 1, 86% for model 2	Fuentes et al. (2020)
Modeling of Milk Yield	Random Forest	Environmental conditions	Random forest 79.26% (median accuracy of the predictions)	Bovo et al. (2021)
predicting on-farm direct water and electricity consumption on pasture-based dairy farms	CART decision tree, random forest ensemble, artificial neural network, support vector machine	Milk yield, dairy cows, electricity water	Support vector machine, 54% (electricity prediction accuracy)	Shine et al. (2018)

(Continues)

TABLE 1 (Continued)

Purpose	Method	Data	Best model and accuracy of prediction	References
Quality Assessment and Grading of Mil	Neural Network	pH, temperature, odor, color, taste, fat, turbidity	Neural network not available	Swarup Kumar et al. (2022)
Prediction of Milk Production	Support vector machine, ANN-ABO, ANN-GA	The health condition of cows, feed intake capacity and expected relative milk yield	ANN-ABO	Suseendran and Duraisamy (2021)
Prediction of the unit price of drinking milk	Random forest, gradient boosting, support vector machine, neural network, and adaBoost	Economic, social, environmental parameters	Random forest, 99.30% for training and 98.10% for testing the dataset	The proposed study

Abbreviations: ANN-ABO, artificial neural network-artificial butterfly optimization; ANN-GA, artificial neural network-genetic algorithm; CART, classification and regression trees; LASSO, least absolute shrinkage and selection operator; MA, model averaging; MLP, multilayer perceptron; NN, neural network; PLSDA, partial least squares discriminant analysis; PLSR, partial least squares regression; RR, ridge regression.

The novelty of this study is to use ML algorithms to determine the factors affecting the price of drinking milk, predict the price of milk, and present a case study on Turkey. The present research consists of five main parts. A literature review about the subject of the study and the methods used were included in the first and the second part of the study. Following the introduction part of the study, detailed information about milk production and consumption in Turkey was provided. The explanation of ML models was discussed in the paper's third part. In this section, detailed information about the study data was also provided. The numerical results of the methods used were examined in the fourth part of the study. The consequences of the emergence of this study and the results of the recommendations were involved in the last section of the study.

2 | DRINKING MILK PRODUCTION AND CONSUMPTION IN TURKEY

Milk production and consumption have an immediate product life cycle regarding food. With the development of technology, although unopened packaged dairy products last for a long time, the shelf life of dairy products is short because their consumption is rapid. In such a situation, it is natural that there will be fluctuations in the country's production of dairy products. Figure 1 shows the amount of drinking milk produced in Turkey and the amount of milk consumption per capita (per 100 people). The amount of milk consumption per capita was calculated by the ratio of the amount of milk production to the population. However, not all of the drinking milk produced is consumed by humans. According to one study, 45% of the milk consumed only at school breakfasts is thrown away (Blondin et al., 2017). For this reason, in this study, the amount of milk production was taken into account, not the amount of milk consumed. Despite the population change, there is a close relationship between milk production and milk consumption in Turkey. The correlation value between the amount of milk production and the amount of milk consumption per capita was calculated as 0.9796.

Milk production and consumption are expected to increase in the coming years, and even by 2050, it is likely that the consumption of milk and dairy products will increase by 20% (Alexandratos & Bruinsma, 2012). The link between milk production and consumption is difficult to explain in terms of population. Because although the highest milk production in Turkey took place in the last months of 2017 and the first months of 2018, Turkey had the highest population in 2021. In the 8-year data period discussed in the study, milk production in 2011–2012 was

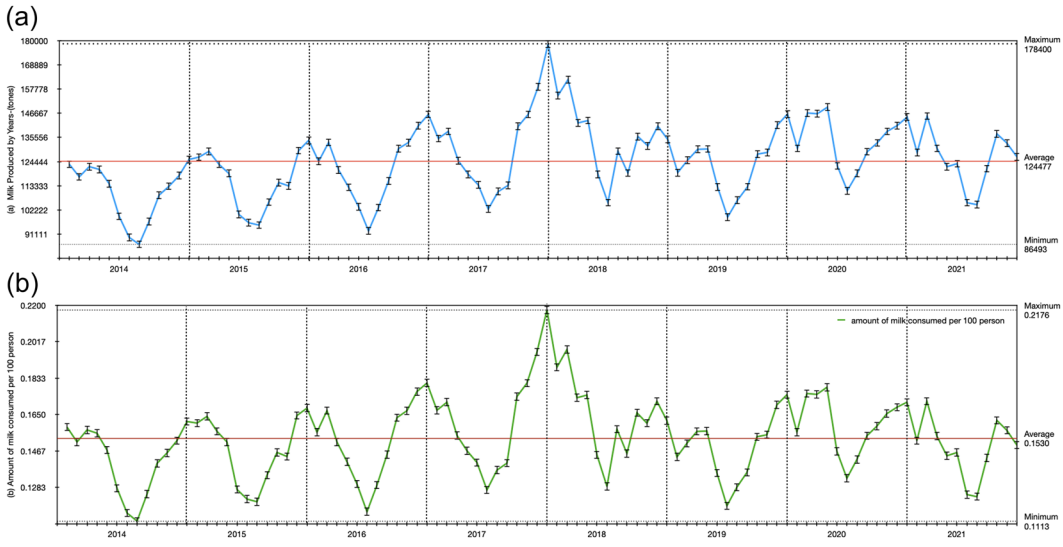

FIGURE 1 (a) Milk production amount; (b) milk consumption per capita

below the average milk production. Turkey produces an average of 124.5 thousand drinking milk for 8 years. During this period, the amount of milk per capita was calculated as 0.15 L per 100 people. This rate only determines the amount of drinking milk and does not include other dairy products (fresh milk). Milk production generally increases and decreases according to seasonal periods. The amount of milk production in summer seasons is less than in winter seasons in Turkey.

Milk consumption contains rich nutritional values for human health. Milk is an essential nutrient considered a nutrient-dense beverage that contains all the nutrients necessary for a healthy life (Chalupa-Krebzdak et al., 2018). Drinking milk has a great deal of powerful components such as proteins, minerals, fat, and vitamins (Mourad et al., 2014). Regarding food, the average fat rate of cow milk collected by commercial dairy enterprises was 3.3%, and the protein ratio was 3.2% on average in Turkey (TUIK, 2021).

3 | MATERIAL AND METHODS

The methodology for estimating the price of drinking milk is explained in this section of the study. As summarized in Figure 2, the input parameters for the model are preprocessed, and then the data is divided into learning and testing rations. Learning data related to drinking milk was used to create prediction models with five different ML algorithms. ML algorithms were performed with orange data mining software. Orange program is free and open-source software based on phyton. The models are then used to predict the testing data, and their performance is compared on four different metrics. Ultimately, the data of the algorithm that provides the best prediction result from the ML models for the study are provided in this study.

3.1 | Data preparation

A complex dataset is required because a multidimensional approach is needed to determine the unit price of drinking milk. Possible input parameters for drinking milk price estimation were selected from the data used in this

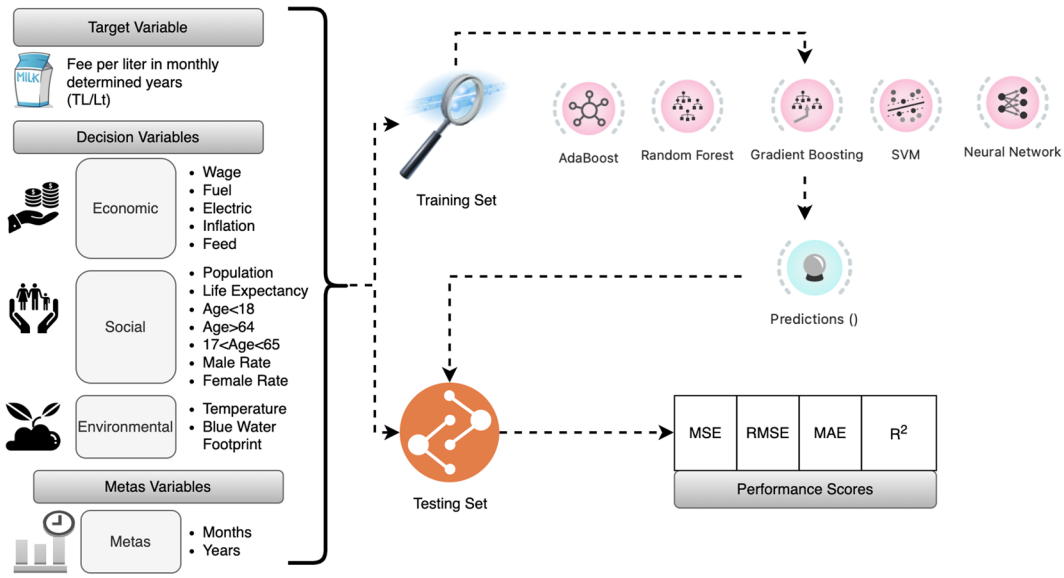

FIGURE 2 Flow chart of the proposed study

study. First, drinking milk price data were obtained from the statistical reports published on the official website of the Turkish national milk council (https://ulusalsutkonseyi.org.tr/) from the monthly data for 2014 and 2021. The factors affecting milk production are discussed under three main headings: economic, social, and environmental. In addition, the target information on the important parameters related to the unit price of drinking milk was obtained from the official website of the Turkish Statistical Institute (https://tuik.gov.tr/).

The minimum wage, fuel prices, which have an important place in milk transportation, electricity unit prices especially used in pasteurized milk production, annual and monthly inflation rates of the country, and finally, the feed prices used without feeding animals were taken into account as indexed economic indicators that affect milk prices. Naturally, the social factors that affect milk consumption in a country are the population of the country, life expectancy, age ranges (discussed in three different categories), and gender differences. According to the literature review, temperature and blue water footprint data for pasteurized milk are effective in the amount of milk production. According to a study by Drastig et al. (2010) small, medium, and large dairy farms in Sweden required 1.3 L, 1.05 L, and 1.2 L of water for 1 L of milk production, respectively. The amount of water used in this study was calculated by proportioning the amount of milk produced. Since the regions where dairy farms are located in Turkey are many and different, the unit price of water is not measured. In this study, these environmental factors have been taken into account, considering that they will also have an effect on milk prices. The time indicators in the dataset of the study were defined as metavariables, and these data were used only to compare the forecast data. This study carried out detailed research by considering 17 (including time indicators) input factors that affect milk unit price. Descriptive statistics data were summarized by using statistical analysis to identify the data of the variables. These data are presented in Table 2.

Correlation data were obtained by using the correlation algorithm of the orange software to show whether there is a statistical correlation between the variables. The image of the algorithm created to measure the correlation value is shown in Figure 3.

The correlation value between the variables varies between −1.00 and +1.00. Correlation values are categorized by being in a certain limit range. Generally, the correlation coefficient is defined as a "weak," "moderate," and "strong" relationship in the literature. Considering in more detail, according to Schober et al. (2018), if these values are between 0.00 and 0.10 (or −0.10) there is no correlation (negligible), between 0.10 (or −0.10) and 0.39 (or −0.39) there is a weak correlation, between 0.40 (or −0.40) and 0.69 (or −0.69) there is a moderate correlation, between 0.70 (or −0.70) and 0.89 (or −0.89) there is a strong correlation, and between 0.90 (or −0.90) and 1.00 (or −1.00) there is a very strong correlation. The correlation values obtained by considering all the combinations between the output variable milk unit price and the input variables are given in Table 3.

There is a connection between the variables when the correlation values of all variables are examined in detail. Only environmental factors have weak correlation values on milk price. Correlation values indicating moderate, strong, and very strong correlations between the remaining variables were calculated.

The most fundamental factor in the formation of the price mechanism is the supply-demand relationship. It is observed that the supply-demand balance in milk production for the country whose data are analyzed has increased over time. Because, according to the economic theory, any unexpected negative events that may occur in the formation of the price mechanism and the deterioration of the efficient distribution of economic resources are caused. Due to the short shelf life of products belonging to the dairy industry, any adverse situation in the market will adversely affect the price mechanism. In this study, the milk market in the pilot country was evaluated under normal conditions.

All possible levels of a product chain must be evaluated before analysis of the milk price mechanism. The price transmission elasticity coefficient (EPT) is the primary indicator to assess the quality and quantity of the price transmission intensity of a product. EPT is that the price change in the upstream market influences the price change in the downstream demand. By defining the upper and lower markets as i and j, the milk unit price transfer coefficient (EPT_{ij}) is determined based on the relationship between these two markets. The formula was formed as below (McCorriston et al., 2001):

$$EPT_i = \frac{\delta p_j}{p_j} \Big/ \frac{\delta p_i}{p_i}. \quad (1)$$

TABLE 2 Descriptive statistics values of output and input variables

Variables	Size	Mean	SE mean	Standard deviation	Minimum	Maximum	Skewness	Kurtosis
Output variable								
Drinking milk price (TL/Lt)	96	1.7089	0.0726	0.7112	1.0000	4.500	1.34	1.60
Input variables								
Economic factors								
Feed (TL/kg)	96	1.3058	0.0628	0.615	0.7200	2.860	1.23	0.55
Electric (MW/price)	96	405.40	18.900	185.1	234.30	813.5	0.83	−0.72
Fuel (TL/Lt)	96	5.8290	0.1220	1.197	4.1290	9.760	0.57	−0.25
Wage (TL/month)	96	1728.5	58.400	571.9	1071.0	2825.9	0.74	−0.78
Inflation (year)	96	12.351	0.5260	5.153	6.5700	36.080	1.59	3.7
Inflation (month)	96	1.1620	0.1640	1.606	−1.4400	13.580	5.26	38.4
Social factors								
Population (million)	96	81.304	0.2390	2.337	77.696	84.680	−0.12	−1.30
Age < 18 (% of the total population)	96	28.022	0.1030	1.012	26.080	29.400	−0.52	−0.60
Age > 64 (% of the total population)	96	8.7625	0.0598	0.586	8.0000	9.7000	0.34	−1.31
Age > 17	96	63.215	0.0466	0.456	62.600	64.220	0.96	0.61
Age < 65 (% of the total population)								
Female (% of the total population)	96	0.4984	0.0001	0.003	0.4982	0.4990	1.27	0.12
Male (% of the total population)	96	0.5016	0.0001	0.003	0.5010	0.5018	−1.27	0.12
Life expectancy (years at birth)	96	74.576	0.1460	1.432	73.050	77.990	1.51	1.49
Environmental factors								
Temperature (°C)	96	12.451	0.7860	7.706	2.1000	26.000	0.49	−1.21
Blue water footprint (tones/Lt) 10^5	96	0.0692	0.0010	0.009	0.0486	0.0991	0.19	0.28

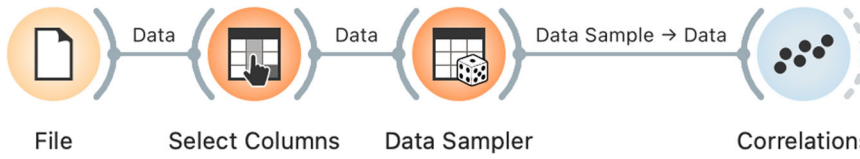

FIGURE 3 The algorithm created to measure the correlation value

TABLE 3 Correlation values between variables according to the algorithm

Feature 1	Feature 2	Correlation of coefficient
Price (TL/Lt)	Wage (TL/month)	0.980
Age < 17 (% of the total population)	Price (TL/Lt)	−0.977
Age > 65 (% of total population)	Price (TL/Lt)	0.977
Life expectancy (years at birth)	Price (TL/Lt)	0.976
Population (million)	Price (TL/Lt)	0.976
Feed price (kg/TL)	Price (TL/Lt)	0.969
Age 18–64 (% of the total population)	Price (TL/Lt)	0.949
Electric (MW/price)	Price (TL/Lt)	0.947
Fuel (TL/Lt)	Price (TL/Lt)	0.901
Inflation year (%)	Price (TL/Lt)	0.802
Male (%)	Price (TL/Lt)	−0.775
Female (%)	Price (TL/Lt)	0.775
Price (TL/Lt)	Inflation month (%)	0.363
Blue water footprint (tones*L) per 10^5	Price (TL/Lt)	0.363
Price (TL/Lt)	Temperature (°C)	0.026

The (EPT) value of drinking milk price in Turkey was calculated as 0.12878 according to the data time interval and the first-to-last production quantities. With this result, there is an inelastic situation in the milk market (|EPT| < 1 ⇒ demand is inelastic). In other words, milk price increases do not decrease the quantity demanded. Peltzman (2000) argues that the price increase of different factors is reflected in the cost instantly, while the price decrease does not change at the same rate (Peltzman, 2000). Often, regression models are used to analyze changes to prove whether positive or negative cost changes are better transferred to the product unit price. The power of the regression of the time terms, which is the basis of the ML algorithms and also considered in this study, the regression equation represented by the annual-monthly unit price differences, in which the significant positive and negative price differences are followed, was formed as follows (Hušek, 2007; Revoredo-Giha et al., 2004);

$$\Delta P_{jt} = \left\{ A^{+} + \sum_{l=1}^{k} B_l^{+} x \Delta P_{it}^{+};\quad A^{-} + \sum_{l=1}^{k} B_l^{-} x \Delta P_{it}^{-} \right\}. \tag{2}$$

Generally, the time delay parameter is not considered in the milk price formation mechanism. In particular, due to the nature of milk and dairy products, the density coefficient of the time delay is neglected since the durability of milk and dairy products is short. In this study, only metavariables were defined to express the values of time factors in the data used.

For the milk price formation mechanism, the following formula is considered for the transmission analysis of the milk unit price with a linear vector error correction technique with more than one input variable (Fernández-Amador et al., 2010):

$$\Delta P_t = \left\{ \beta_0 + \alpha\theta' + P_{t-1} + \sum_{j=1}^{l} \Gamma_j \Delta P_{t-j} + \in_t, \in_t \sim N(0, \Sigma) \right\} \quad (3)$$

where P_t is a vector composed of the customer price of milk price, β_0 is the constant value of the equation, $\theta' P_{t-1}$ defines the relationship between the price levels. The price dynamics are formed through Γ_j parameter matrices based on price variability up to the l^{th} lag. ML algorithm approaches are used to estimate the price difference between real-time markets for the same node in estimating the price difference between real-time markets. ML algorithms were used in this study by taking these equations into account, while the equations mentioned earlier are at the base of the milk unit price formation mechanism. Generally, while price transfer takes place from producer to consumer prices, since only milk producer prices are taken into account in this study, price transfer is not carried out using ML algorithms. The data were processed in ML algorithms in their raw form. In this paper, the suitability of multivariate linear and nonlinear regression-based ML algorithms was also measured in estimating algorithm performances.

4 | ML ALGORITHMS

Today, ML algorithms are the most preferred method to generate prediction data in problems with complex and large-scale datasets. In this study, ML algorithms were used to estimate the unit price of drinking milk, taking into account the data on the factors affecting the drinking milk price in Turkey. Different ML algorithms such as RF, gradient boosting, SVM, neural network, and AdaBoost in orange software were applied in this study. Each model tries to predict the output variable with different algorithms. Orange software provided under the GNU General Public License (https://www.gnu.org/licenses/gpl-3.0.html), an open-source and free computer program that performs data visualization, ML, and data mining, has been used in this study. Forecast data of five different ML models discussed in this software were calculated with a single interface. The program interface image created for

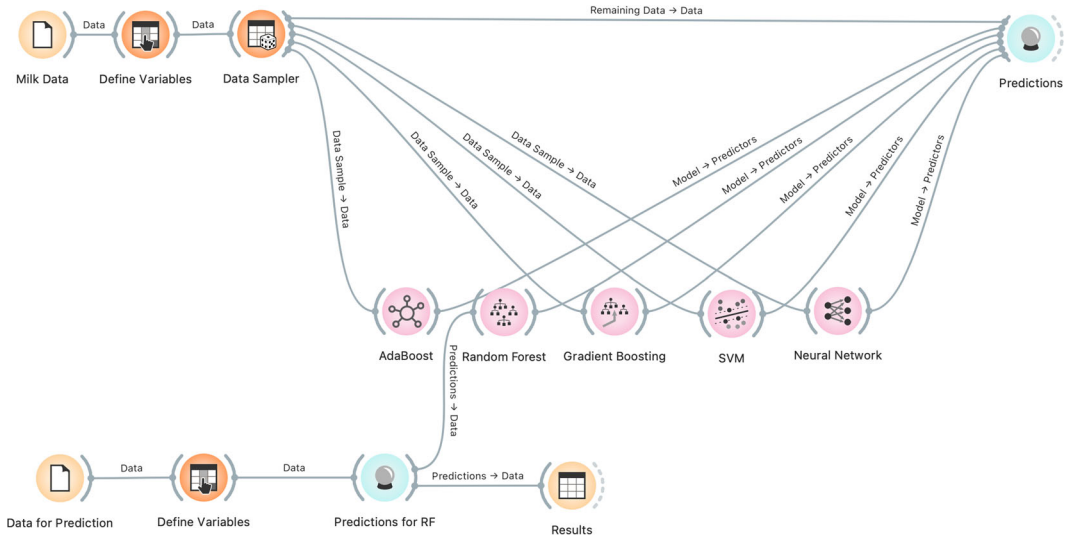

FIGURE 4 The algorithm of machine learning models

this study is presented in Figure 4. The basic definitions, working principles, and pseudo-codes of model algorithms of ML models used in the method of this study are briefly given in subsections.

4.1 | RF

The RF algorithm is an ML algorithm based on a general-purpose classification and regression. This model provides the results by combining a few randomly selected decision trees from the generated decision trees and summing the averages of the obtained prediction values. RF gives outstanding results, especially in studies where the number of variables is more than the number of samples. Researchers apply the RF model because it is versatile in large and highly variable problems. The pseudo-code of the RF algorithm is shown below; in the algorithm, the symbol s represents the partition number; c, the number of the decision tree of the RF; k, the ratio of positive and negative states, λ the sampling threshold, and δ the nearest neighbor search radius (Lin et al., 2017).

Algorithm 1. Random Forest algorithm

Input: Training data T, parameters {λ, δ, k, s, c}
Output: Model with evaluation
1. Ensemble-RF (T, λ, δ, k, s, c)
2. for i < −1 to s **do**
3. (train, test) ← randomSplit(T, λ)
4. split ← bootstrap(train, δ, k)
5. model ← RandomForest.train(split, c)
6. score ← evaluate(model, test)
7. out[i] ← (model, score)
8. end for
9. return out

4.2 | Gradient Boosting algorithm

The Gradient Boosting algorithm has been proposed for the classification of ML problems. The primary approach of this model is to present a statistical approach by connecting the concepts of loss functions in the created algorithm. Thus, an advanced algorithm is created by increasing the prediction accuracy. Gradient boosting is also defined as an optimization algorithm that aims to create an additive model by minimizing the loss function values. This algorithm reaches the maximum number of iterations until the missing function values decrease. Thus, the previous decision tree incrementally adds to the next decision tree to obtain the best estimation results. The essential pseudo-code of this algorithm is created as follows (Touzani et al., 2018):

Algorithm 1. Gradient Boosting algorithm function

1. The depth of the decision trees d, the number of repetitions K, the learning rate α, and the subsample fraction η.
2. Initialization: set the residual $r_0 = y$ and $\hat{f} = 0$. The mean value of y as an initial prediction of $\hat{f}$
3. For $k = 1, 2, K$, do the following:
   a subsample $\{y_i, x_i\}^{N'}$ from the training set, with N' is the number of data points corresponding to the fraction η
   b $\{y_i, x_i\}^{N'}$ apt a decision tree $\hat{f}^k$ of depth d to the residual r_{k-1}
   c $\hat{f}$ by adding the decision tree to the model $\hat{f}(x) \leftarrow \hat{f}(x) + \alpha\hat{f}^k(x)$
   d the residual $r_k \leftarrow r_{k-1} - \alpha\hat{f}^k(x)$
4. end For

4.3 | SVM

SVM is an ML algorithm that works with the classification method by labeling objects. This algorithm provides a decision boundary between classes (usually two classes) that allows estimating the vectors formed by tagging objects with one or more features (Huang, 2018). However, SVMs are among the most suitable ML techniques for binary classification in non-parametric statistical applications. In parametric applications, the suitability of the estimation function to linear and margin maximization is tested. Then the SVM is converted to nonlinear and nonparametric (Auria & Moro, 2008). Created the necessary pseudo-code for training the SVM is shown below (Harimoorthy & Thangavelu, 2021):

Algorithm 1. SVM algorithm function

```
Input: D = [X,Y]; X(array of input with m features),
       Y(array of class labels)Y = array(C)//Class label
Output: Find the performance of the system
function train_svm(X, Y, number_of_runs)
initialize: learning_rate=Math.random();
for learning_ rate innumber_of_runs
        error = 0;
for i in X
if (Y[i] * (X[i] * w))<1 then
update: w = w + learning_rate * ((X[i] * Y[i]) * (−2 * (1/number_of_runs) * w)
else
update: w = w + learing_rate * (−2 * (1/number_of_runs) * w)
end if
end
end
```

This supervised ML algorithm classifies data into positive and negative classes with maximum distance or margin. This distance can be expressed by Equations (1) and (2) (Burges, 1998):

$$x_i w + b \geq +1 \text{forall} y_i = +1, \tag{4}$$

$$x_i w + b \leq -1 \text{forall} y_i = -1, \tag{5}$$

where, x and y are defined as training data, w denotes the weight of vectors, and b represents the bias value. Expressing the hyperplane as $w^T x = 0$, in SVM, the best hyperplane should be perceived as a considerable distance between the margins. So, to increase the margin between the hyperplane $\|w\|$ Equation (3) is taken into account:

$$min_{w,b} = \frac{1}{2} \|w\|^2. \tag{6}$$

4.4 | Neural Network Algorithm

Neural Network neurons connect input variables to output or target variables so that each neuron is attached to the synapse. The neuron information generated by the input variables at each synapse is generated by a weighting method. These data are calculated by iterative or iterative learning. In this method, for the training set to be compatible, excessive repetition should be avoided; otherwise, the errors of the test set will increase (see Figure 5). For this reason, in this method, a generalized model is developed to obtain the forecast data with a low margin of error (Jung & Kim, 2016).

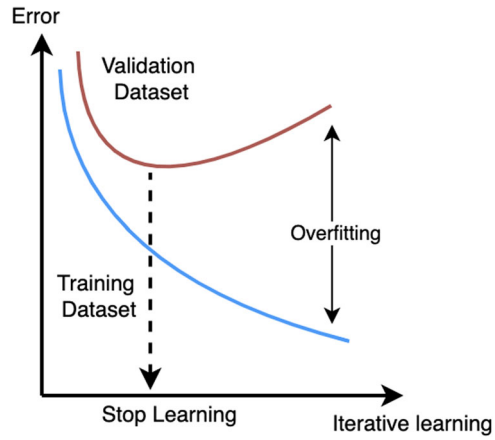

FIGURE 5 Learning curve expressing the effect of over-repetition on the Neural Network algorithm (Jung & Kim, 2016)

4.5 | AdaBoost algorithm

AdaBoost is an ML algorithm formulated by Yoav Freund and Robert Schapire in 1995 (Freund et al., 1999). The AdaBoost algorithm works with the classifier approach, as in other ML models, and works with the logic of adjusting the misclassified samples. The AB model creates weighting values for the input parameters with the sense of repetition and using the weak or faulty classifier over and over as in the Neural Network and AdaBoost models (Sprenger et al., 2017). In short, AdaBoost provides a useful ML model by strengthening the weak classifier and increasing the prediction of the weak classification algorithm. This model creates a second model that tries to correct the errors of the first model by developing a prototype model from the values of the training dataset. Models are continued to be created and added to each other until the errors are minimized (Khan et al., 2020). Formed the fundamental pseudo-code for training the SVM is as below (Khan et al., 2020):

Algorithm 1. AdaBoost algorithm function

```
ADABOOSTING(samples, A, H) returns a weighted-majority hypothesis
  inputs: samples, set of N labeled samples (x1, y1), … ,(xN,yN)
    A, an algorithm
    H, the hypotheses in the ensemble
  other variables: a, a vector of N example weights, initially 1 / N
      b, a vector of H hypotheses
      z, a vector of H hypothesis weights
  for i = 1 to H do
    b[i] ← L(samples, a)
    error ← 0
    for j = 1 to N do
      if b[i](xj) ≠ yj then error ← error + a[j]
    for j = 1 to N do
      if b[i](xj) = yj then a[j] ← a[j] • error / (1 − error)
    a ← NORMALIZE(a)
    c[i] ← log(1 − error) / error
  return MAJORITY(b, c)
```

5 | ALGORITHMS PERFORMANCE INDICATOR

Estimating the drinking milk price per unit was provided using five different ML algorithms, and some performance criteria need to be calculated to compare these models in the present study. These criteria are included in ML as root mean square error (RMSE), mean absolute error (MAE), mean squared error (MSE), and the coefficient of correlation (R), respectively. The models use these criteria to detect the linear connection between the variables defined in R^2 (coefficient of determination). MAPE calculates the amount of error in the forecast data as a percentage. Returns the standard deviation of the RMSE estimation errors, an indicator of variability. The following equations show the formula for each of these performance criteria:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2, \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|, \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n}}, \tag{9}$$

$$R^2 = \sum_{i=1}^{n}\left[\frac{y_i - \tilde{y}_i}{y_i - \bar{y}_i}\right]^2, \tag{10}$$

where y_i, $\overline{y}_i$, and $\tilde{y}_i$ represent the actual value in the dataset, the average value of the dataset, and the predictive values generated by ML models, respectively. n is the number of samples in a dataset. MSE, MAE, and RMSE are the performance measurement tools that minimize empirical predictive values in ML models. An ML model with a high R^2 value and low RMSE, MSE, and MAE values are considered to outperform other models. In ML models, the option of determining the variables that will give the best result for which model by applying feature selection is widely used. The best prediction results were obtained in this study using the model's unique feature selection option. The following formulas are used for feature selection (Cai et al., 2018; Zhang et al., 2020):

$$F(g_i, T) = \int p(g_i, T) ln\left(\frac{p(v_i, T)}{p(v_i, T)}\right) dg_i dT \tag{11}$$

$$oT_f = \frac{1}{|\varphi|}\sum_{v_i \in \varphi} F(v_i, T) - \frac{1}{|E|^2}\sum_{v_i v_{ij} \in \varphi} F(v_i, v_j), \tag{12}$$

where, F denotes the feature selection information and φ represents the milk price. v_i signifies the milk price in a given time. T symbolizes the decision variable variety. $F(v_i, v_j)$ represents the interactive information between v_i and v_j. oT_f characterizes the number and type of decision variable to determine optimal feature selection.

6 | RESULTS AND DISCUSSION

This study is based on ML modeling using economic, social, and environmental data as inputs. Considering the 17 independent variables that affect the drinking milk price, developing a more robust and less error-prone estimation model is inevitable due to the inadequacy of the classical approaches used for estimation methods to ensure that the deviations are at a minimum level. Recently, ML algorithms have been used for estimation processes, despite the possibility of complexity and high divergence. These algorithms are widely used in many fields such as energy, health, economy, agriculture, and education. The main reason these algorithms are preferred is that statistical approaches are ignored. They learn how the data of the input parameters change and thus provide the data for the

output parameter. For this reason, five different ML algorithms, which are the most preferred, were used in this study.

The multidimensional scaling (MDS) projection of the target variable onto a plane suitable for certain distances between data points is given in Figure 6. The particle in the target variable needs either a dataset or a distance matrix. Multidimensional scaling in ML algorithms finds the low-dimensional projection of the points and tries to fit the distances between the points into a plane. The main reason for this in practice is to obtain a perfect fit since the distances of high-dimensional data are not Euclidean. The proximity of the points representing the data indicates a force separating the points. Thus, the interval between the maximum impact (or minimum impact) on the data is considered.

6.1 | Comparison of ML algorithms

The main theme of this study is to estimate the unit price of drinking milk using datasets of economic, social, and environmental factors that affect the unit price of drinking milk. ML algorithms were used to calculate the forecast data. The models considered for calculating predictive values of ML algorithms must have some features. The features of the algorithms are given in Table 4. These properties are defined as the number of trees, hidden layer values, the number of neurons in the hidden layers, the number of hidden nodes in the test data, and the loss function degrees.

These algorithms were divided into test and training datasets, and prediction data were obtained. ML algorithms were trained and tested by dividing the training and test data with a ratio of 85%/15% on the condition of being random. The performance data showing the error margins and accuracy rates of these algorithms on the prediction data were calculated and compared. The performance values of the algorithms are discussed in detail in Table 5.

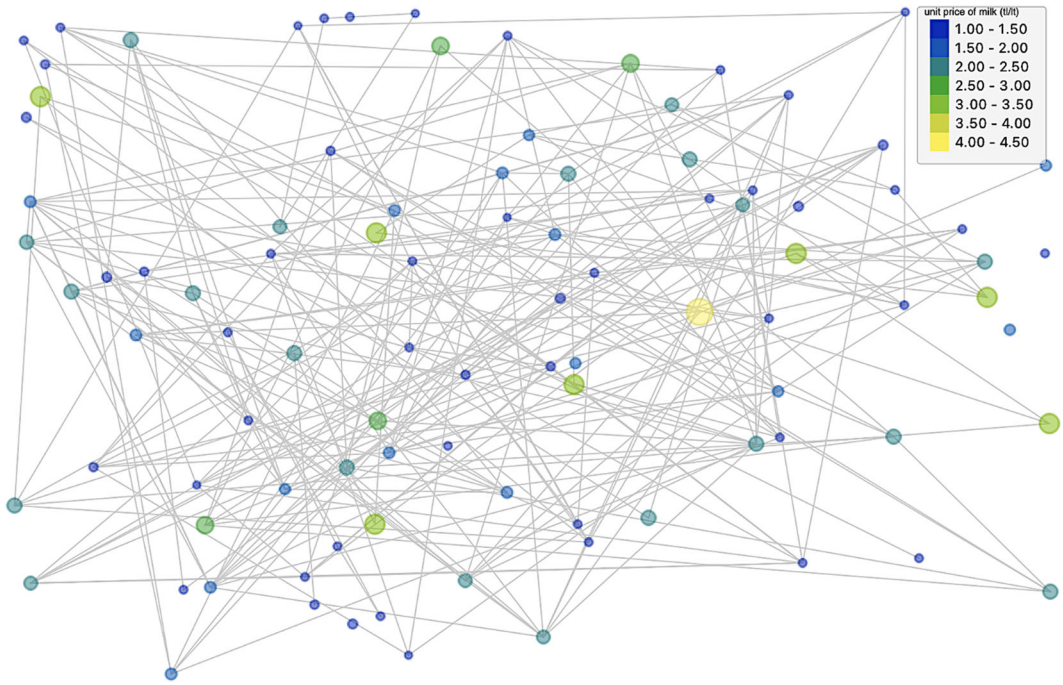

FIGURE 6 Reflection of the target variable on a suitable plane at certain distances between data points

TABLE 4 Brief of initial hyper-features for each machine-learning algorithm

Algorithms	Hyper-features
Random Forest	Number of trees: 10–2000
	Number of attributes at each split:5
Gradient Boosting	Method: Gradient Boosting (scikit-learner)
	Number of trees: 100
	Learning rate: 0.100
	Replicable training: yes
	Limit depth of individual trees:3
	Not split subsets smaller than: 2
	Fraction of training instances: 1.00
SVM	SVM Type: SVM
	Cost (C): 1.00
	Regression loss epsilon (ε): 0.10
	Kernel: RBF[a], g: auto
	Optimization parameters
	Numerical tolerance: 0.0010
	Iteration limit: 100
Neural Network	Neurons in hidden layers:100
	Activation: Relu[b]
	Solver: Adam
	Regularization, α = 0.0001
	Maximal number of iterations: 200
	Replicable training: yes
AdaBoost	Base estimator: tree
	Number of estimators: 50
	Learning Rate: 1.00000
	Classification algorithm: SAMME.R[c]
	Regression loss function: Linear

Abbreviation: SVM, support vector machine.

[a]C of the Radial Basis Function (RBF) kernel SVM.

[b]Rectified Linear Unit.

[c]Type of algorithm that uses probability estimates to update the additive model.

As a result of statistical data of other models, RF, gradient boosting, SVM, neural network, and AdaBoost correlation coefficient values in the testing-training dataset were computed as 0.993–0.981, 0.972–0.974, 0.946–0.977, 0.972–0.941, and 0.948–0.973, respectively. The RF algorithm has data of consistent estimation results with high correlation values among the five models. The RSME values of RF, gradient boosting, SVM, neural network, and AdaBoost were calculated as 0.060–0.096, 0.118–0.111, 0.165–0.105, 0.118–0.167, and

TABLE 5 Value of measurement performances of ML models for drinking milk prices

Algorithm	Training dataset				Testing dataset			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
Random forest	0.004	0.060	0.029	0.993	0.009	0.096	0.055	0.981
Gradient boosting	0.014	0.118	0.095	0.972	0.012	0.111	0.058	0.974
SVM	0.027	0.165	0.078	0.946	0.011	0.105	0.078	0.977
Neural network	0.014	0.118	0.095	0.972	0.028	0.167	0.128	0.941
AdaBoost	0.026	0.163	0.038	0.948	0.013	0.114	0.046	0.973

Note: Consider MSE, RMSE, and ME as %.

Abbreviations: MAE, mean absolute error; ML, machine learning; *R*, coefficient of correlation; RMSE, root mean square error; SVM, support vector machine.

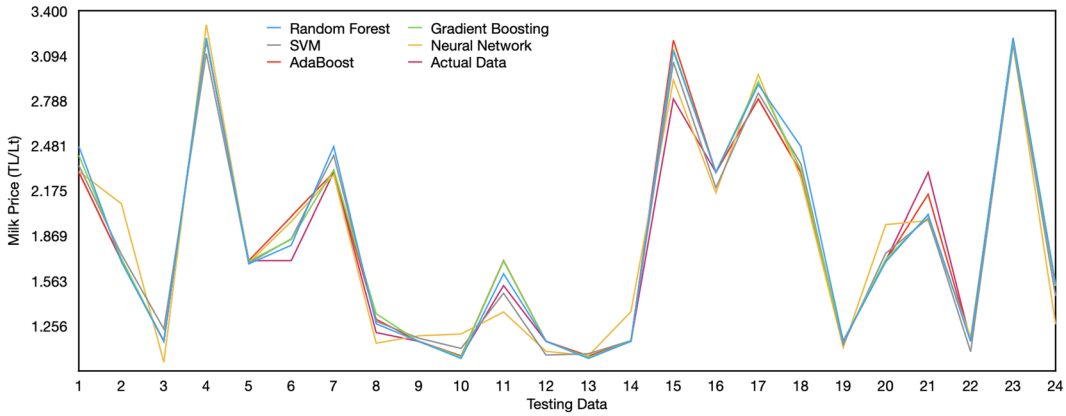

FIGURE 7 Comparison of estimated data of machine learning algorithms

0.163–0.114, respectively. Likewise, the RF has low errors compared to the RMSE values of the models. The RE values of RF, gradient boosting, SVM, neural network, and AdaBoost were estimated as 0.029–0.055, 0.095–0.058, 0.078–0.078, 0.095–0.128, and 0.038–0.046, respectively. The MSE values of RF, gradient boosting, SVM, neural network, and AdaBoost were computed as 0.004–0.009, 0.014–0.012, 0.027–0.011, 0.014–0.028, and 0.026–0.013, respectively. According to the values of the performance measurement criteria, RF provides the best estimate of the most response variable, while SVM provides the worst estimate data.

The estimated data of the target variable according to the deviation values and accuracy rates of the ML algorithms are given in Figure 7. The estimated values of the ML algorithms discussed in this study are very close to each other. The price of drinking milk information has been calculated by taking into account the deviations that occur in line with the measurement performance values of the ML models. These values are visualized in Appendix A part of the study.

The RF algorithm was run with different tree sizes between 10 and 2000 to measure the effect on the tree size estimation data of the test dataset with the RF algorithm. Figure 8 displays the relationship between the prediction accuracy of the RF algorithm and the number of trees. The RF algorithm, which fluctuates up to a certain point with the change of the tree bot, works without affecting the prediction accuracy coefficient after a certain number of trees.

Figure 9 represents the training, testing, prediction, and actual data of the ML model working according to the RF algorithm. Test and training data were determined according to completely random logic.

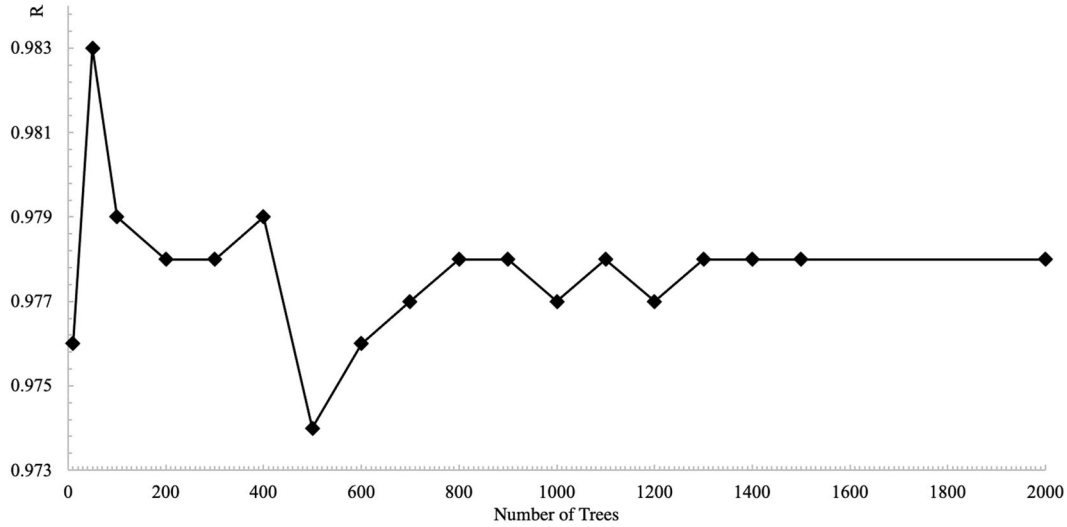

FIGURE 8 R^2 values by the number of trees in random forest

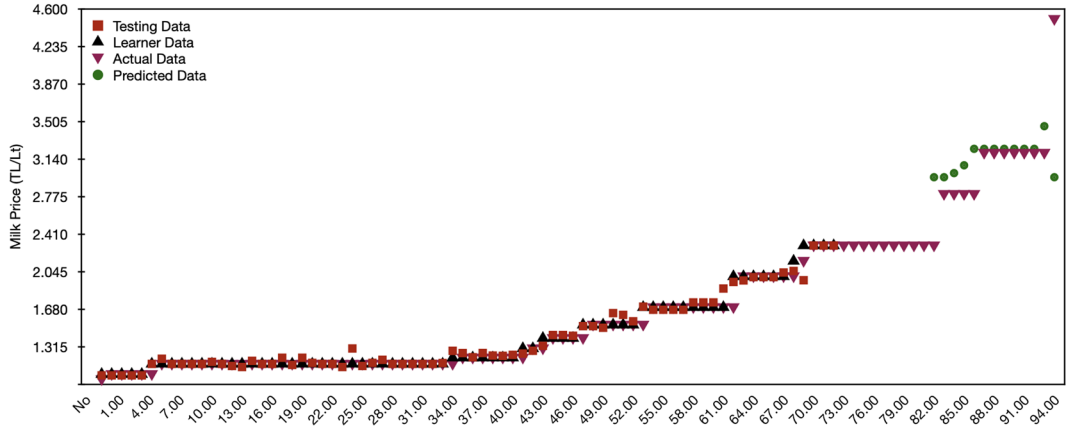

FIGURE 9 Comparison of the predicted performance values of the random forest model with actual, testing, and learner values for drinking milk prices

Figure 10 demonstrates the actual and estimated values for the monthly drinking milk price per liter. The fact is that there is a very close relationship between the forecast data and the actual data for 2021. The very good performance of the RF algorithm is explained by the fact that the prediction data of the last period is very close to the real data, the error rates of the RF algorithm are low, and the reliability degree is very high. The fact that the forecast value for the last month is far from the actual data can be explained by several reasons, such as the negative impact of the COVID-19 epidemic on production and the adverse developments in the country's economy.

This study has some limitations. First of all, only the amount of milk obtained from cows is included in the drinking milk considered in this study. The amount of milk and dairy products belonging to other animal species is not mentioned. Second, the data of this study are discussed in the data covering the COVID-19 epidemic period. Figure 1 shows that the amount of milk production in Turkey has increased from the end of 2019, which is considered a pandemic period, to the last period in which data are used. Another limit is that due to the high inflation data for the last period of the country whose data are taken into account, the milk price suddenly increased

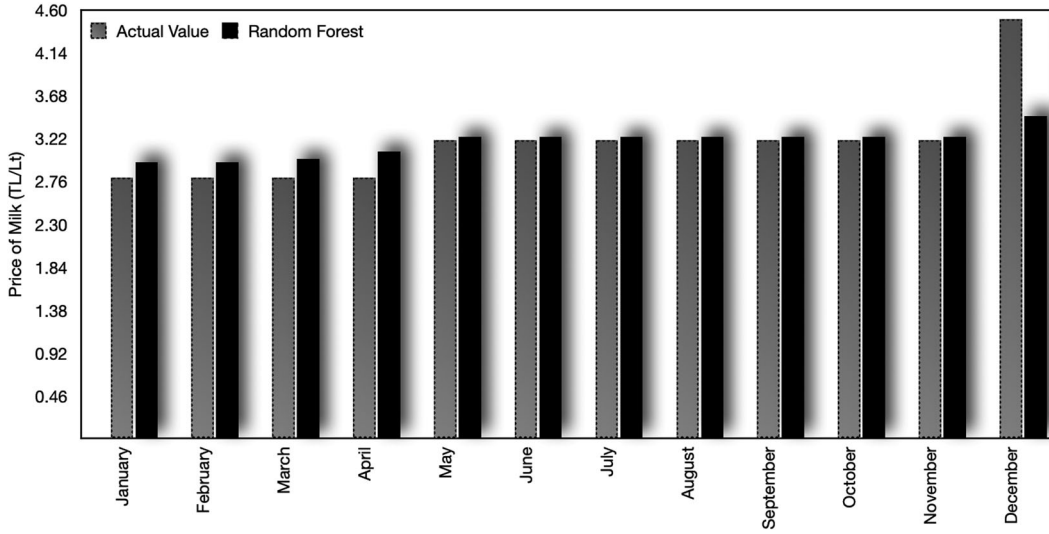

FIGURE 10 Comparison of the actual and predicted values of the random forest algorithm

by 40.62% (3.20 TL/Lt in October 2021, 4.50 TL/Lt in December 2021). It should be emphasized that this increase slightly affects the margin of error on the forecast data. However, since this high inflation data coincides with the last period of the study data, the last period of inflation data was ignored in this study. But for future studies, the presence of such severe fluctuations will cause uncertainty in the estimation data. Unexpected shock events (war, natural disaster, etc.) that are thought to have an impact on the drinking milk price have not been taken into account. The fact that access to numerical data of such events or having short-term data has a negative effect on the numerical results obtained. Finally, the study did not include the meat consumption factor because sufficient meat consumption data could not be obtained. The effect of this factor on the price of drinking milk could not be measured. In addition, no connection was established between milk quality and milk unit price, and data were taken into account according to a uniform quality approach. This study can be expanded by using the data of other countries, or this study can be handled in a narrower scope by determining specific locations (such as cities or regions) and using data from those locations to compare the results of this study.

7 | CONCLUSION AND FUTURE WORKS

The ML algorithms developed in this study can be applied to predict and evaluate the unit price of drinking milk automatically. Depending on the inputs of the algorithms, ML modeling techniques can be applied to a dairy farm in any country, region, or city. Especially for AI applications and developed ML algorithms, economic, social, and environmental factors that are effective on milk unit price are required. This article demonstrated a practical application of ML using detailed information for the benefit of countries' milk production investments to increase milk productivity. The results obtained in the study appear to be relevant for three main reasons:

1. The size of the training dataset adopted from the data collected for ML algorithms is suitable for the purpose of the algorithms;
2. Since the default input variables in the study are effective on the target variable, the input variable selections seem suitable for the study;

3. Among the ML algorithms, the RF algorithm is quite robust and reliable according to the data type;
4. The unit price of drinking milk is ensured to be confirmed based on results from RF, which can be a reliable and applicable tool for evaluating future scenarios.

The recent increase in the unit price of drinking milk may cause concern in the dairy industry. Most of the sources of this increase can be shown as fluctuations in economic factors such as inflation, gasoline price, and feed price. As a result, some factors affecting the unit price of drinking milk need to be regulated. The sustainability of the dairy sector can be sustained by developing economic support, especially for farmers to increase milk yield, improve milk production environments according to environmental conditions, and provide other needed immediate assistance.

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Abdulkadir Atalan http://orcid.org/0000-0003-0924-3685

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/agr.21773.

REFERENCES

Alexandratos, N., & Bruinsma, J. (2012). *World Agriculture towards 2030/2050: The 2012 revision*. ESA Working paper No. 12-03. FAO.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics*, *15*(1), 41–51.

Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis. *SSRN Electronic Journal*, DIW Berlin Discussion (Paper No. 811, 1-18). http://doi.org/10.2139/ssrn.1424949

Blondin, S. A., Cash, S. B., Goldberg, J. P., Griffin, T. S., & Economos, C. D. (2017). Nutritional, economic, and environmental costs of milk waste in a classroom school breakfast program. *American Journal of Public Health*, *107*(4), 590–592.

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals*, *11*(5), 1305.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79.

Capper, J. L., Cady, R. A., & Bauman, D. E. (2009). The environmental impact of dairy production: 1944 compared with 20071. *Journal of Animal Science*, *87*(6), 2160–2167.

Chalupa-Krebzdak, S., Long, C. J., & Bohrer, B. M. (2018). Nutrient density and nutritional value of milk and plant-based milk alternatives. *International Dairy Journal*, *87*, 84–92.

Costa Leite, J., Keating, E., Pestana, D., Fernandes, V. C., Maia, M., Norberto, S., Pinto, E., Moreira-Rosário, A., Sintra, D., Moreira, B., Costa, A., Silva, S., Costa, V., Martins, I., Mendes, F. C., Queirós, P., Peixoto, B., Caldas, J. C., Guerra, A., … Calhau, C. (2017). Iodine status and iodised salt consumption in Portuguese school-aged children: The iogeneration study. *Nutrients*, *9*(5), 458.

Dalala, Z., Al-Omari, M., Al-Addous, M., Bdour, M., Al-Khasawneh, Y., & Alkasrawi, M. (2022). Increased renewable energy penetration in national electrical grids constraints and solutions. *Energy*, *246*, 123361.

Drastig, K., Prochnow, A., Kraatz, S., Klauss, H., & Plöchl, M. (2010). Water footprint analysis for the assessment of milk production in Brandenburg (Germany). *Advances in Geosciences*, *27*, 65–70.

Farah, J. S., Cavalcanti, R. N., Guimarães, J. T., Balthazar, C. F., Coimbra, P. T., Pimentel, T. C., Esmerino, E. A., Carmela, M., Duarte, K. H., Freitas, M. Q., Granato, D., Neto, R. P. C., Tavares, M. I. B., Calado, V., Silva, M. C., & Cruz, A. G. (2021). Differential scanning calorimetry coupled with machine learning technique: An effective approach to determine the milk authenticity. *Food Control*, *121*, 107585.

Fernández-Amador, O., Baumgartner, J., & Crespo-Cuaresma, J. (2010). *Milking the prices: The role of asymmetries in the price transmission mechanism for milk products in Austria*. University of Innsbruck, Department of Public Finance.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal of Japanese Society For Artificial Intelligence*, *14*(771–780), 1612.

Frizzarin, M., Gormley, I. C., Berry, D. P., Murphy, T. B., Casa, A., Lynch, A., & McParland, S. (2021). Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *Journal of Dairy Science*, *104*(7), 7438–7447.

Fuentes, S., Viejo, C. G., Cullen, B., Tongson, E., Chauhan, S. S., & Dunshea, F. R. (2020). Artificial intelligence applied to a robotic dairy farm to model milk productivity and quality based on cow data and daily environmental parameters. *Sensors*, *20*(10), 2975.

Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing*, *12*(3), 3715–3723.

Hušek, R. (2007). *Ekonometrická Analýza* (1st ed.). Agricultural Development Economics Division (ESA).

Ji, B., Banhazi, T., Phillips, C. J. C., Wang, C., & Li, B. (2022). A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. *Biosystems Engineering*, *216*, 186–197.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Jung, S.-K., & Kim, T.-W. (2016). New approach for the diagnosis of extractions with neural network machine learning. *American Journal of Orthodontics and Dentofacial Orthopedics*, *149*(1), 127–133.

Khan, F., Ahamed, J., Kadry, S., & Ramasamy, L. K. (2020) Detecting malicious URLs using binary classification through adaboost algorithm. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(1), 997.

Kirschbaum, M. U. F., Schipper, L. A., Mudge, P. L., Rutledge, S., Puche, N. J. B., & Campbell, D. I. (2017). The Trade-Offs between milk production and soil organic carbon storage in dairy systems under different management and environmental factors. *Science of the Total Environment*, *577*, 61–72.

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access*, *5*, 16568–16575.

Lovarelli, D., Finzi, A., Mattachini, G., & Riva, E. (2020). A survey of dairy cattle behavior in different barns in Northern Italy. *Animals*, *10*(4), 713.

McCorriston, S., Morgan, C. W., & Rayner, A. J. (2001). Price transmission: The interaction between market power and returns to scale. *European Review of Agricultural Economics*, *28*(2), 143–159.

Mourad, G., Bettache, G., & Samir, M. (2014). Composition and nutritional value of raw milk. *Issues in Biological Sciences and Pharmaceutical Research*, *2*, 115–122.

Mu, F., Gu, Y., Zhang, J., & Zhang, L. (2020). Milk source identification and milk quality estimation using an electronic nose and machine learning techniques. *Sensors*, *20*(15), 4238.

Mylostyvyi, R., & Chernenko, O. (2019). Correlations between environmental factors and milk production of Holstein cows. *Data*, *4*(3):103.

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*(8), 440–456.

Ongsulee, P. (2017). Artificial Intelligence, Machine Learning and Deep Learning. In *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)* (pp. 1–6). IEEE.

Peltzman, S (2000). Prices rise faster than they fall. *Journal of Political Economy*, *108*(3), 466–502.

Revoredo-Giha, C., Nadolnyak, D. A., & Fletcher, S. M. (2004). Explaining price transmission asymmetry in the US peanut marketing chain. 2004 Annual meeting, American Agricultural Economics Association, 1-18.

Reziti, I. (2014). Price transmission analysis in the Greek milk market. *SPOUDAI—Journal of Economics and Business*, *64*(4), 75–86.

Rhone, J. A., Koonawootrittriron, S., & Elzo, M. A. (2008). Factors affecting milk yield, milk fat, bacterial score, and bulk tank somatic cell count of dairy farms in the central region of Thailand. *Tropical Animal Health and Production*, *40*(2), 147–153.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768.

Shine, P., Murphy, M. D., Upton, J., & Scully, T. (2018). Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms. *Computers and Electronics in Agriculture*, *150*, 74–87.

Sowmya, N., & Ponnusamy, V. (2021). Development of spectroscopic sensor system for an IoT application of adulteration identification on milk using machine learning. *IEEE Access*, *9*, 53979–53995.

Sprenger, M., Schemm, S., Oechslin, R., & Jenkner, J. (2017). Nowcasting foehn wind events using the AdaBoost machine learning algorithm. *Weather and Forecasting*, *32*(3), 1079–1099.

Suseendran, G., & Duraisamy, B. (2021). Predication of dairy milk production using machine learning techniques. In Lung Peng, S., Yuan Hsieh S., Suseendran, G., & Duraisamy B., *Intelligent computing and innovation on data science* (pp. 579–588). Springer.

Swarup Kumar, J. N. V. R., Indira, D. N. V. S. L. S., Srinivas, K., & Satish Kumar, M. N. (2022). Quality Assessment and Grading of Milk Using Sensors and Neural networks. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1772–1776). IEEE.

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543.

TUIK. (2021). Milk and Dairy Products Production, December 2021. Retrieved from Februay 14, 2022. https://data.tuik.gov.tr/Bulten/Index?p=Milk-and-Milk-Products-December-2021-45747

Xiaoxia, D., Zhemin, L. I., Shiwei, X. U., & Ganqiong, L. I. (2013). Study on price transmission mechanism between corn, beanpulp and milk in China based on MCM approach. *Journal of Systems Science and Mathematical Sciences*, *33*(1), 55–66.

Zhang, Z.-M., Tan, J.-X., Wang, F., Dao, F.-Y., Zhang, Z.-Y., & Lin, H. (2020). Early diagnosis of hepatocellular carcinoma using machine learning method. *Frontiers in Bioengineering and Biotechnology*, *8*, 254.

Zhu, X., Wen, J., & Wang, J. (2020). Effect of environmental temperature and humidity on milk production and milk composition of guanzhong dairy goats. *Veterinary and Animal Science*, *9*, 100121.

AUTHOR BIOGRAPHY

Abdulkadir Atalan holds his PhD degree in Industrial Engineering from Marmara University, MSc degree in Healthcare Systems and Industrial Engineering from Lehigh University, and BSc degree in Industrial Engineering from Selçuk University. Dr. Atalan's research expertise and interests include machine learning, discrete event simulation, statistics, and optimization.

How to cite this article: Atalan, A. (2023). Forecasting drinking milk price based on economic, social, and environmental factors using machine learning algorithms. *Agribusiness*, 39, 214–241. https://doi.org/10.1002/agr.21773

APPENDIX A

Figures A1-A5.

The Appendix below contains the Boxplot figures containing the deviation statistics for estimating the unit price data of invisible drinking milk. Boxplot is preferred to show the distributions of the estimated values of ML algorithms. Boxplot has been applied to discover anomalies such as duplicate values, outliers, and the like and check for any new data. The deviation values for the five models (random forest, gradient boosting, SVM, neural network, and adaBoost) were generated based on the mean estimation data. The thin blue line in the figures represents the standard deviation of the algorithms. The blue area represents the values between the first and third quartiles in the dataset used in the algorithms. Median values are expressed as a yellow color.

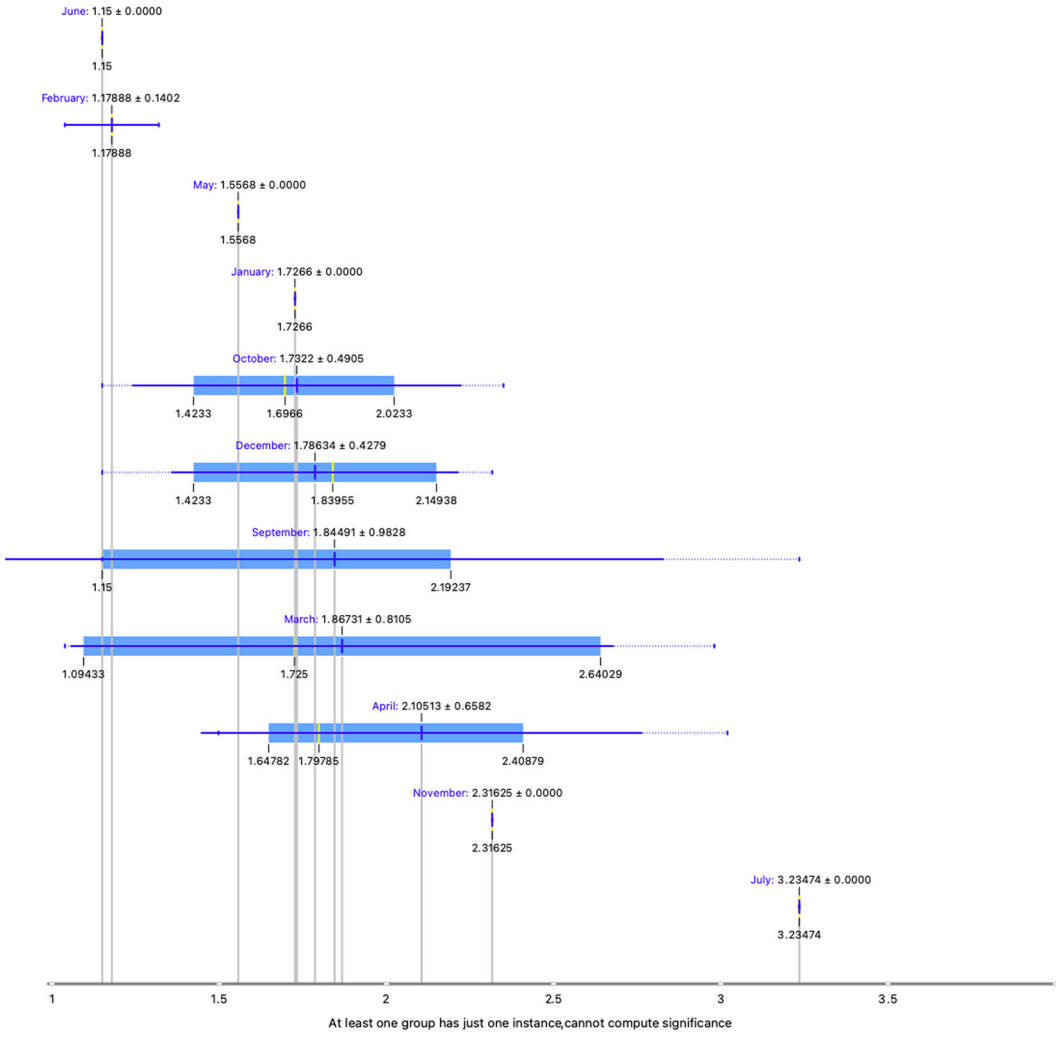

FIGURE A1 Distribution of data based on random forest algorithm

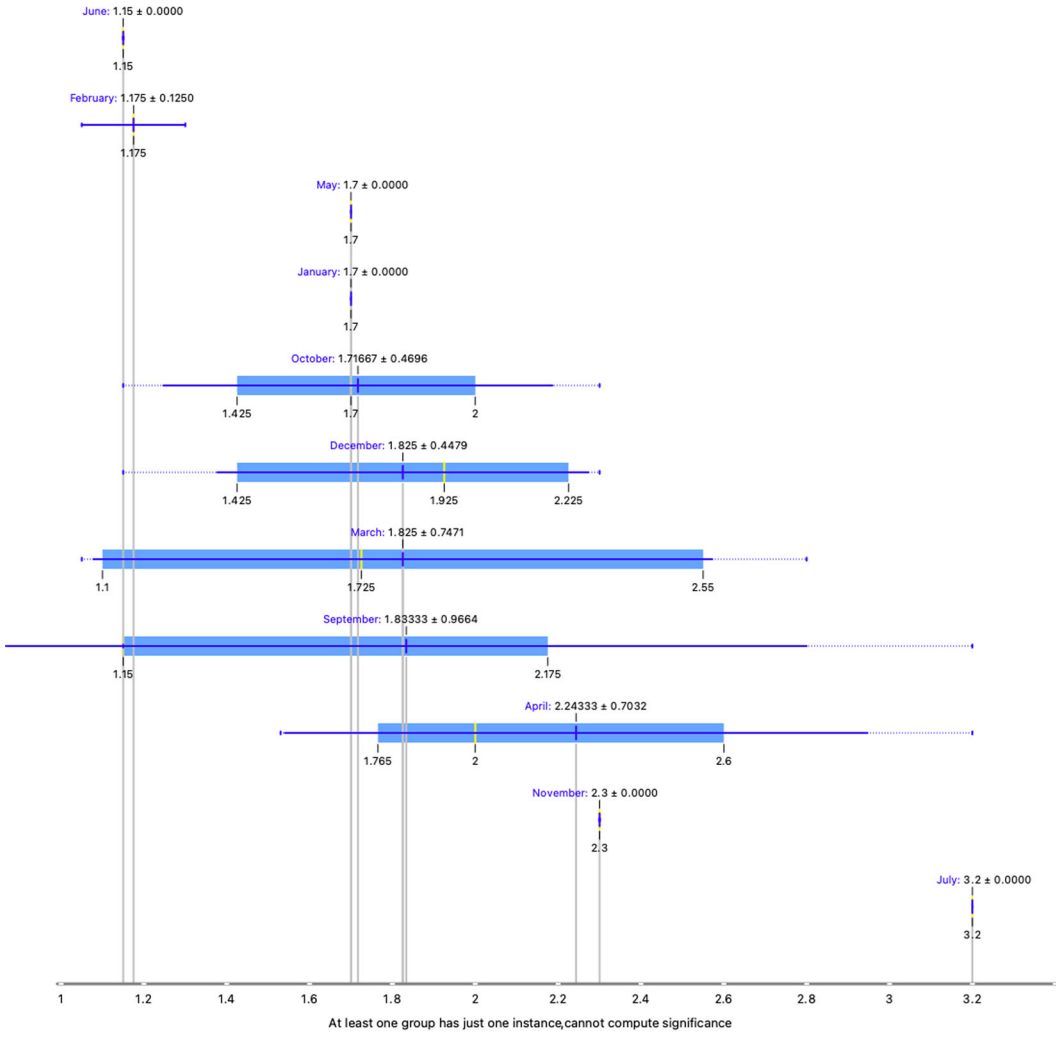

FIGURE A2 Distribution of data based on AdaBoost algorithm

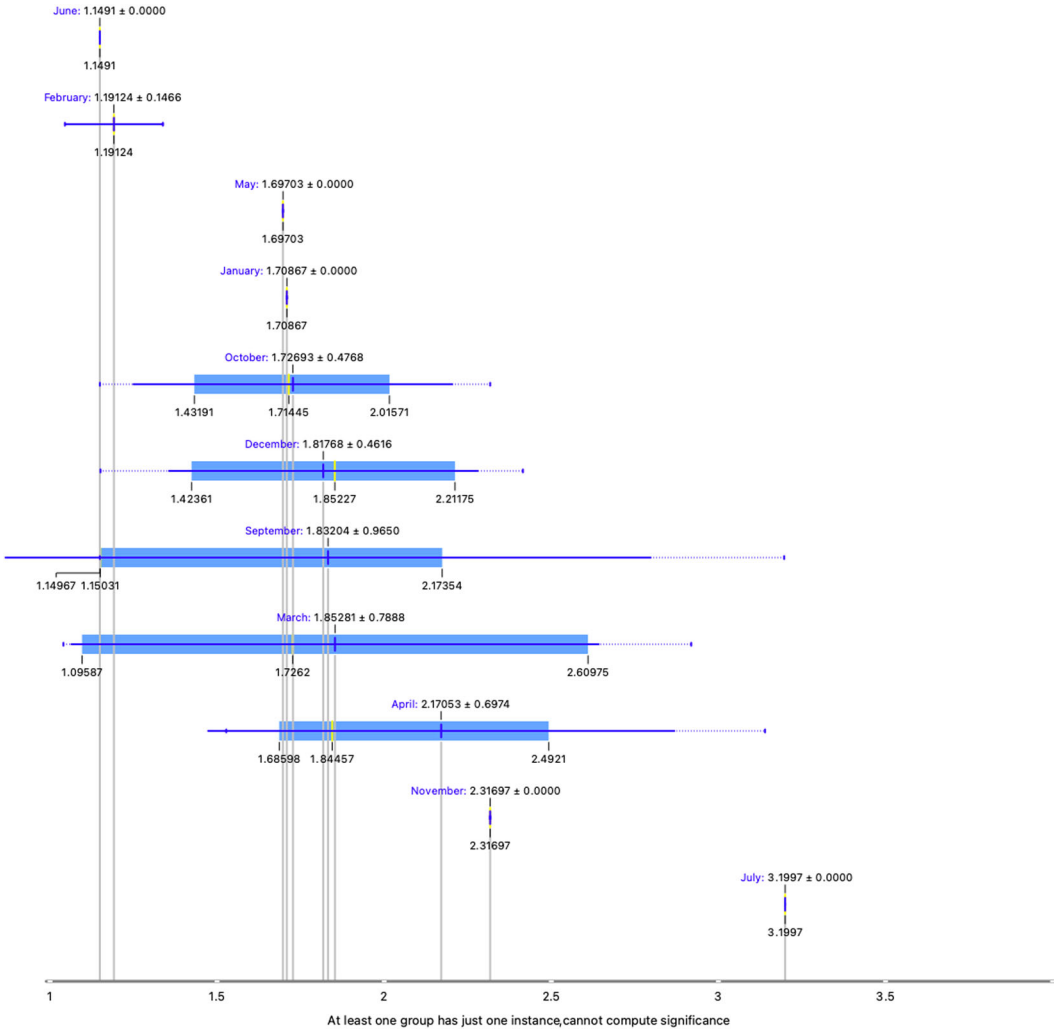

FIGURE A3 Distribution of data based on gradient boosting algorithm

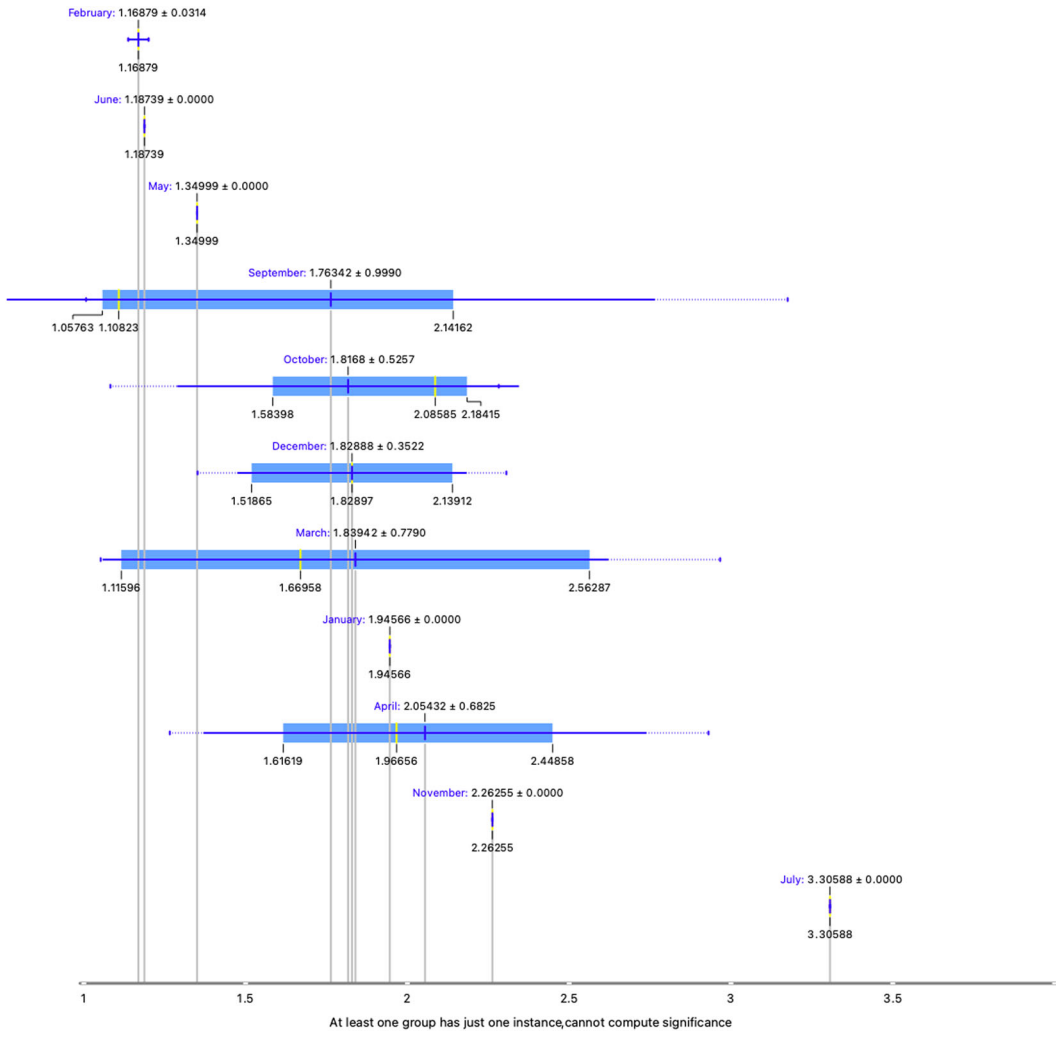

FIGURE A4 Distribution of data based on a neural network algorithm

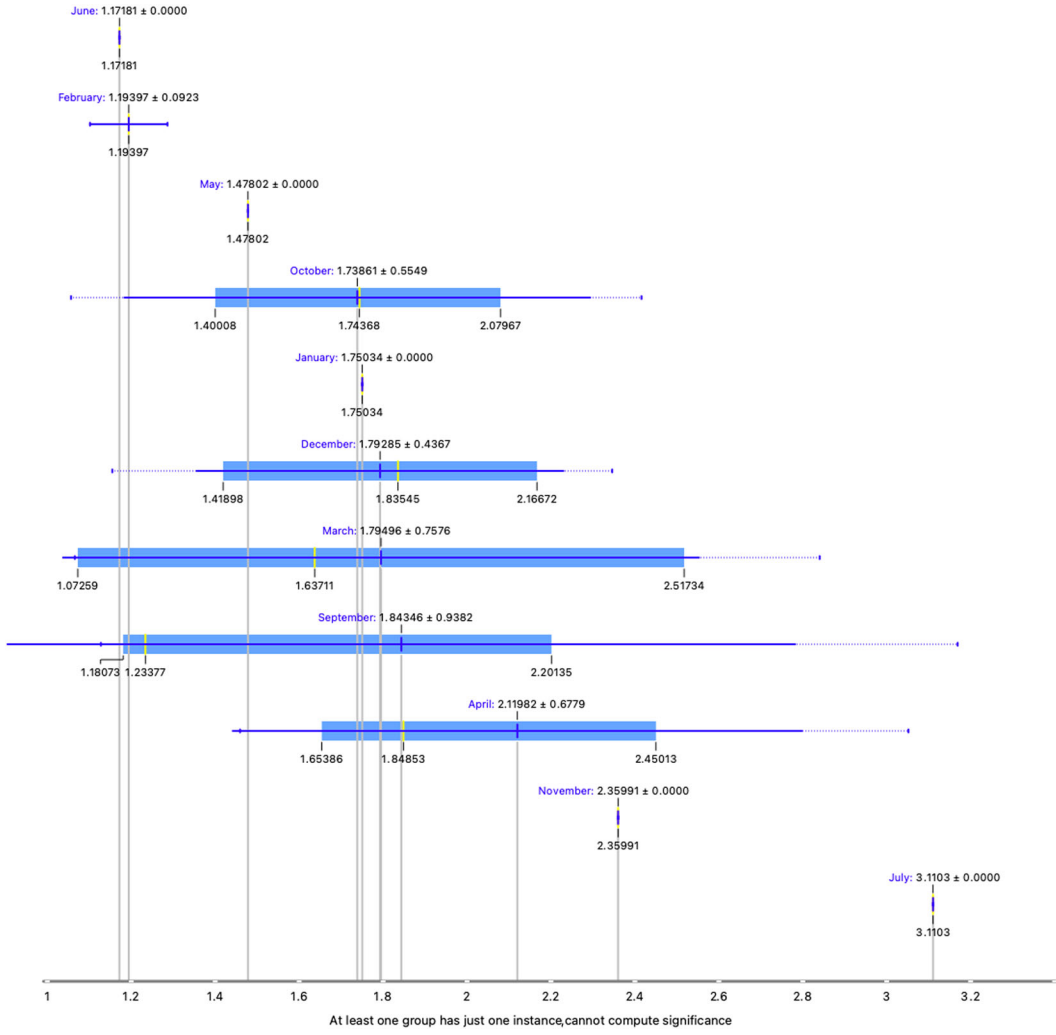

FIGURE A5 Distribution of data based on support vector machine