Turkish Journal of Agricultural and Natural Sciences 12 (3): 640–652, 2025

https://doi.org/10.30910/turkjans.1671584

TÜRK TARIM ve DOĞA BİLİMLERİ DERGİSİ

TTDB

TURKISH JOURNAL of AGRICULTURAL and NATURAL SCIENCES

www.dergipark.gov.tr/turkjans

Original Article

Determination of Factors Affecting Net Profit in Buffalo Milk Production by Different Data Mining Algorithms: A Case Study of Iğdır Province

Köksal Karadaş [1], Osman Doğan Bulut [1]✉, Hakan Duman [2]

[1]Iğdır University, Faculty of Agriculture, Department Agricultural Economics, Iğdır-Türkiye
[2] Iğdır University, Iğdır Vocational School of Higher Education, Department of Transportation Sevices, Iğdır-Türkiye

[1]https://orcid.org/0000-0003-2682-6356, [2]https://orcid.org/0000-0003-1176-3313, [3]https://orcid.org/0000-0001-6166-5776

✉: dgnblt@gmail.com

ABSTRACT

Buffalo milk is an important animal product because it has more protein and fat content compared to other types of milk. This study aimed to identify the key factors influencing profitability in buffalo milk production using advanced data mining algorithms. Data were collected from 92 buffalo farms in Iğdır Province, Türkiye, in 2016 by using the Simple Random Sampling Method. Among the 4 models developed in the R program, the Multivariate Adaptive Regression Splines (MARS) model demonstrated superior predictive performance based on cross-validation and goodness-of-fit criteria. The results revealed that lactation year (LY) and lactation period (LP) were the most significant variables affecting net profit. Profitability was highest in the seventh lactation year, while extending the lactation period beyond 175 days contributed to linear profit increases. The findings suggest that buffalo producers should adopt management strategies focused on culling buffaloes after the seventh lactation and extending lactation periods to improve economic outcomes. This research highlights the effectiveness of data mining techniques in determining profitability factors and provides recommendations to optimize production efficiency in livestock systems. In future research, more comprehensive models can be developed using larger datasets and additional variables.

Key words: Buffalo milk, Net profit, Data Mining Algorithms, Iğdır province, Türkiye

Farklı Veri Madenciliği Algoritmaları ile Manda Sütü Üretiminde Net Kârı Etkileyen Faktörlerin Belirlenmesi: Iğdır İli Örneği

ÖZ

Manda sütü diğer süt türlerine kıyasla daha fazla protein ve yağ içeriğine sahip olması sebebiyle önemli bir hayvansal üründür. Bu çalışmada, gelişmiş veri madenciliği algoritmaları kullanarak manda sütü üretiminde kârlılığı etkileyen temel faktörlerin belirlenmesi amaçlanmıştır. Veriler, 2016 yılında Basit Tesadüfi Örnekleme Yöntemi kullanılarak Türkiye'nin Iğdır ilinde 92 manda çiftliğinden toplanmıştır. R programında geliştirilen 4 model arasında, Çok Değişkenli Uyarlamalı Regresyon Eğrileri (MARS) modeli, çapraz doğrulama ve uyum iyiliği kriterlerine dayalı olarak üstün tahmin performansı göstermiştir. Laktasyon yılı (LY) ve laktasyon süresinin (LP) net kârı etkileyen en önemli değişkenler olduğunu tespit edilmiştir. Kârlılık yedinci laktasyon yılında en yüksek iken, laktasyon süresinin 175 günden fazla uzatılması doğrusal kâr artışlarına katkıda bulunmuştur. Bulgular, manda üreticilerinin kârlılığını artırması için yedinci laktasyondan sonra mandaları ayıklamaya ve laktasyon dönemlerini uzatmaya odaklanan yönetim stratejileri benimsemeleri gerektiğini göstermektedir. Bu araştırma, kârlılık faktörlerini belirlemede farklı veri madenciliği tekniklerinin etkinliğini ortaya koymakta ve hayvancılık sistemlerinde üretim verimliliğini optimize etmeye yönelik öneriler sunmaktadır. Gelecek araştırmalarda, daha büyük veri kümeleri ve ek değişkenler kullanılarak daha kapsamlı modellemeler yapılabilir.

Anahtar kelimeler: Manda sütü, Net kâr, Veri madenciliği algoritmaları, Iğdır ili, Türkiye

INTRODUCTION

Buffaloes are more resistant to natural conditions and diseases than cattle, and they make much better use of pastures and sub-forest pastures, and they can easily convert low-quality feeds into meat and milk, and thus, they are important in terms of encouraging sustainable agricultural production by allowing low-cost animal products to be obtained (Becskei et al., 2020). Buffalo milk provides a safe source of high-quality nutrients and enables the production of valuable products for healthy nutrition, and it is more advantageous than cow's milk in terms of its physicochemical, compositional and sensory properties, and stands out in terms of nutrition and health (Mane and Chatli, 2015). Buffalo milk has complete proteins with high biological value and contains all the essential amino acids that the human body needs (Khedkar et al., 2016), and it is very suitable for the production of products such as butter and cream because it contains less water and more fat compared to cow's milk (Pudja et al, 2008). Buffalo cheese is very valuable with its color and texture, and the famous Italian cheeses Mozzorella and Borelli are made from buffalo milk (Park and Haenlein, 2008; Guo and Hendricks, 2010). Buffalo meat is preferred because it has low fat and cholesterol, and buffalo skin is preferred because it is thick (Sarıözkan, 2011). While the world buffalo population, which was 88321107 head in 1961, increased by 132% to 205141830 head by 2022, in the same years, the buffalo population in Turkey decreased by 85% from 1140000 head to 171835 head (FAO, 2022). Although there are studies on buffalo breeding, its importance, meat and milk quality in Turkey (Atasever and Erdem, 2008; Sarıözkan, 2011; Yılmaz et al., 2011; Toparslan and Mercan, 2018; Konca and Yılmaz Adkinson, 2021; Yılmaz Adkinson and Konca, 2021), there is no study on the reasons for the decrease in the number of buffalo in Turkey, and this situation shows that the importance of buffalo breeding and buffalo products on farms in Turkey is not sufficiently understood. In Turkey, the total milk production in 2022 is 21563490 tons, and only 0.2% (43588 tons) of this amount belongs to buffalo milk, while 0.65% (15386 tons) of the total red meat production amount of 2384047 tons in 2023 belongs to buffalo meat (TÜİK, 2023). The world buffalo milk yield per animal is 2022 kg, and the first three places are Iran 2844 kg, Pakistan 2298 kg and India 2205 kg, while Turkey ranks 15th with 597 kg. As of 2022, the world buffalo meat yield average is 247 kg per animal, while the first three places are Malaysia 450 kg, India 368 kg and Egypt 330 kg, while Turkey ranks 9th with 218 kg (FAO, 2022). Turkey is below the world average in terms of both buffalo milk yield and buffalo meat yield. The annual inflation rate of around 72% in Turkey in 2022 compared to the previous year (TÜİK, 2022) also increases the input prices of agricultural products. Low yield on the one hand and increasing input prices on the other hand necessitate taking some measures to work more profitably in buffalo milk production and determining the factors affecting profitability.

On the other hand, data mining, which is used to develop appropriate models for determining the factors affecting the milk yield of animals and selecting more productive animals, enables the emergence of hidden patterns in the data for a better understanding of the data relationship, as well as enabling high resource use efficiency and sustainable profitability in livestock production systems (Balhara et al., 2021). The aim of this study is to determine the factors affecting the profitability in buffalo milk production and to reveal the measures to be taken to obtain more profit from buffalo milk production. For this purpose, the performances of the models used in Data Mining were tested and the model that best explains the factors affecting the profitability in buffalo milk production was tried to be determined. In this context, the following hypotheses were established;

Hypothesis 1:

H_0: There is no relationship between the determined factors and profitability in buffalo milk production.

H_1: There is a relationship between the determined factors and profitability in buffalo milk production.

Hypothesis 2:

H_0: There is no difference between Data Mining Model Performances in determining the factors affecting profitability in buffalo milk production.

H_1: There is a difference between Data Mining Model Performances in determining the factors affecting profitability in buffalo milk production.

MATERIALS AND METHODS

Study Area

Igdir province, located 10 km away from Ararat Mountain, is on the easternmost border of Turkey with 3 neighboring countries of Armenia, Nakhchivan and Iran (Figure 1). Igdir is located between 390 55' latitude and 440 03' longitude and is known as 850 m above sea level.

Figure 1. The location of Igdır province in Türkiye

The distribution of buffalo numbers is shown in Table 1. The central district accounts for 40.64% of the total buffalo numbers in the province, while Aralık and Karakoyunlu districts have 29.50% and 29.86%, respectively. In contrast, Tuzluca district has no buffalo, suggesting that the conditions or resources in this area may not be suitable for buffalo farming.

Table 1. Data on Buffalo Number in Iğdır province (2023)

District	Number of Buffalos (Head)	%
Center	562	40.64
Aralık	408	29.50
Karakoyunlu	413	29.86
Tuzluca	0	0.00
Igdir Province Total	1383	100.00
Türkiye Total	161749	-

Sampling Method and Data Collection

The data obtained from the survey conducted with face-to-face interviews with 92 buffalo-breeding farms in Igdir province is the main material of this study. The survey study was conducted between September to October 2016 and the study covers the 2016 production period. The following sampling formula, which is included in the Simple Random Sampling Method, was used to determine the number of questionnaires used in the research (Arıkan, 2007; Yamane, 2010).

$$n = \frac{N.t^2.pq}{(N-1)D^2 + t^2pq} \quad (1)$$

n= Number of samples
N= Number of registered farms
D= Sampling error
t= Table value
p= The rate to be calculated
q= 1-p

$$n = \frac{270.1{,}96^2.0{,}1.0{,}9}{(270-1)0.05^2 + 1.96^2 0.5.0.5} = 91.68 \quad (2)$$

The distribution of surveys is shown in Table 2. The total number of surveys was distributed proportionally according to the number of members in the districts.

Table 2. Number of surveys by districts

Region	Number of members	Number of Surveys	Rate (%)
Iğdır Center	130	44	48
Aralık	84	29	31
Karakoyunlu	56	19	21
Total	270	92	100

Analytical Model

This study utilized data from 92 buffalo breeding enterprises in Iğdır Province of Turkey in 2016. The dependent variable was net profit (NP), while the independent variables were lactation year (LY), veterinary cost (VC), milking method (MM), mastitis control (MC), concentrate feed quantity (CFQ), roughage quantity (RFQ), lactation period (LP), milking time (MT), and milking interval (MI).

The data was randomly divided into a 70% training set and a 30% test set to determine the applied models' cross-validation (CV) performances. The training set was further divided into 10 folds for parameter tuning through n-fold cross-validation. This approach aimed to identify the most successful models on the training set without causing overfitting problems. In this framework, each cross-validation fold was sequentially set aside. Models, such as multivariate linear regression (LM), multivariate adaptive regression splines (MARS), artificial neural networks (ANN), and decision trees (CART) methods, were trained on the remaining data and compared based on 100 different tuning parameters. The most successful models were then evaluated on the 30% test set, and the best model was determined based on the preferred goodness of fitness criterion (RMSE).

Multivariate Linear Regression (LM)

One of the fundamental analytical methods, linear models, plays a central role in the field of applied statistics. It is widely utilized in machine learning studies, serving as a baseline to compare the performances of other machine learning models (Fox and Weisberg 2019). In addition to ordinary linear regression, such models encompass various variations such as partial least squares regression, ridge regression, and elastic nets (Kuhn and Johnson 2013). In this study, the LM method has been preferred as a fundamental reference point for evaluating the performances of other models. The general equation for LM model, which models the relationship where multiple independent variables influence a dependent variable, is expressed as follows (Eyduran et al. 2017):

$$\boldsymbol{Y = \beta_0 + \sum_{\{i=1\}}^{n} \beta_i X_i + \epsilon}$$

Where; Y is defined as the dependent variable, β_0 is expressed as the intercept, β_i is the ith parameter, X_i is the ith predictor (explanatory variable), and ε is defined as the random error.

Multivariate Adaptive Regression Splines (MARS)

The MARS algorithm, developed by Friedman (1991), is a non-parametric regression technique frequently preferred in data mining. The method is based on an effective algorithm that flexibly determines linear, non-linear, and interaction effects between response sets and predictors. The non-linearity of the model is achieved by providing various regression slopes for intervals of each predictor. The slopes of potential regression lines are determined by the connections between individual regression curves (Friedman 1991). This algorithm, which follows a two-stage process involving forward selection (first stage) and backward deletion (second stage), does not require assumptions about the distributions of variables or the functional relationship between response and predictor variables (Arthur, Temeng, and Ziggah 2020). It is important to note that the advanced transition stage may lead to overfitting problems. The advanced selection stage begins with a basic set of functions that produce the smallest training error. These functions are then iteratively added to advance the model, resulting in a more complex and sophisticated model (Weber et al. 2012). The model obtained in the forward pass stage may fit very well, but its generalization ability may be weak when exposed to a different dataset, indicating susceptibility to overfitting issues. To address these problems, the basic functions that contribute the least to the model are incrementally removed in a stepwise manner during the backward pass (Sahraei et al. 2021). Kuhn (2013) asserted that the MARS algorithm is more interpretable than complex models such as ANN. This algorithm has been extensively studied in previous research (Çelik 2019; Sahraei et al. 2021), providing detailed insights into its workings.

The foundation of the MARS (Multivariate Adaptive Regression Splines) system is based on the use of piecewise linear basis functions, as formulated by Friedman (1991):

$$\mathrm{BF1}(\mathrm{x}) = |\mathrm{x}-\mathrm{t}|_+ = \max(0, \mathrm{x}-\mathrm{t}) = \begin{cases} \mathrm{x}-\mathrm{t}, & \mathrm{x} > \mathrm{t} \\ 0, & \mathrm{x} \leq \mathrm{t} \end{cases} \ldots$$

$$BF2(x) = |t-x|_+ = max(0, t-x) = \begin{cases} \mathrm{t}-\mathrm{x}, & \mathrm{x} < \mathrm{t} \\ 0, & x \geq \mathrm{t} \end{cases} \ldots$$

Here, *t* represents the knots. The aforementioned formulations act as foundational elements for linear or nonlinear modeling aimed at estimating the function *f(x)*. The *notation* $/.|_+$ signifies the positive component. These functions are alternatively referred to as reflected pairs or mirror image functions. They can be defined for each input variable X_m based on its observed values x_{km} where *k=1,2,…,n*, as follows (Celik and Yilmaz 2018):

$$BF1 = max(0, X_m - x_{km})$$
$$BF2 = max(0, x_{km} - X_m)$$

If the dependent variable y relies on M terms, the MARS model can be formulated as shown in the following equation (Friedman 1991):

$$y = f(x) = \beta_0 + \sum_{i=1}^{M} \beta_i H_{ki}\left(X_{v(k,i)}\right) \ldots \ldots$$

Here, β_0 and β_i denote the parameters of the basis functions within the model, and the function H is defined as in the following equation (Friedman 1991):

$$H_{k,i}\left(X_{v(k,i)}\right) = \prod_{k} (-1)^k k_{ki}$$

Here $X_{v(k,i)}$ denotes the estimator corresponding to the *k*-th component of the *i*-th product. When the interaction order *K=1*, the model is additive; whereas for *K=2*, the model incorporates pairwise interactions (Celik and Yilmaz 2018; Friedman 1991).

In general, the generalized cross-validation (GCV) criterion is applied to select the best subset model. GCV is calculated as follows (Celik and Yilmaz 2018):

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\left[1 + \frac{M + d.(M-1)/2}{N}\right]^2}$$

Where, *N* represents the total number of observations, y_i is the dependent variable $\hat{y}_i$ denotes the predicted values from the MARS model, *d* is the penalty applied for each basis function included in the sub-model, and *M* indicates the total number of basis functions.

Artificial Neural Networks (ANN)

ANN were introduced to the literature for the first time in 1943 by neurophysiologist Warren McCulloch and mathematician Walter Pitts. McCulloch and Pitts (1943) introduced a simplified computational model based on the working principles of biological neurons in animal brains to perform complex calculations. This approach has since been followed by numerous different architectures, and the Multilayer Perceptron (MLP) algorithm which is preferred in this study is one of the algorithms developed within this context. The structure, composed of a combination of sensors in multiple layers, lies at the heart of the concept of deep learning. Their versatility, strength, and scalability make them ideal for overcoming large and highly complex machine-learning tasks, such as classifying billions of images (Géron 2017). This algorithm consists of many individual computation nodes connected according to various architectures. The calculations within each node are generally simple; in most cases, models lead to a binary response, but they can deliver powerful results in terms of performance (Titterington 2010). Sensors are based on artificial neurons called linear threshold units. Linear threshold units are structured in a way where each input connection corresponds to a weight, rather than a binary input-output structure. A critical component of these units is the activation function, which introduces non-linearity into the network, enabling it to capture and model complex patterns and dependencies in the data. Activation functions, such as sigmoid, ReLU (Rectified Linear Unit), and tanh, play a vital role by transforming the weighted sum of inputs into outputs that range between fixed limits. This transformation ensures that the network can solve non-linear problems and perform tasks like classification, regression, and feature extraction effectively. In multilayer

perceptrons, it consists of an input layer, one or more layers known as hidden layers, and a final layer called the output layer (Géron 2017). This method has been widely used in various studies, and explanatory information about the algorithm has been presented in previous works (Okut et al. 2013; Zhang et al. 2023).

Decision Trees (CART)

Among the non-parametric methods, tree-based models, categorized as 'divide and conquer' algorithms, operate by constructing a simple model for each region (Boehmke and Greenwell 2020). In this study, the CART algorithm developed by Breiman et. al. (1984) was preferred. In this method, capable of both regression and classification operations, the tree is constructed by progressing from the starting point to the leaf. The design of the tree places the root at the top, with the leaves representing the results at the bottom. At the root, all classifications in the initial dataset are mixed. Subsequently, the tree is developed towards the first node by employing a specific feature variable to partition the population into distinct categories (Nwanganga and Chapple 2020). For regression trees, the modeling process with the CART algorithm begins with the entire dataset. It involves searching through all values for each attribute to identify the optimal split value that minimizes the sum of squared errors by dividing the dataset into two parts. Subsequently, within each of these partitioned groups, the search continues to find the best combination of predictors and partition values that further minimize the sum of squared errors (Kuhn and Johnson 2013). Previous studies have shared detailed information about the CART algorithm (Akin et al. 2018; Faraz et al. 2021).

Explainable Artificial Intelligence Algorithms

The model with the best cross-validation value among the mentioned models was interpreted using explainable artificial intelligence algorithms, specifically variable importance plots and partial-dependence profiles.

Variable Importance (VI): Permutation-based variable importance metrics are a method for determining the significance and weight of an explanatory variable within a model. This method can be utilized to identify and exclude variables that do not impact the model predictions by comparing the importance of variables. Additionally, it can assist in determining the most critical variables and evaluating the validity of the model based on expert opinions (Biecek and Burzykowski 2021).

Partial-dependence Profiles (PD): The fundamental approach in creating partial-dependence profiles, developed by Friedman (2001), is to illustrate the impact of an explanatory variable on the expected value of a chosen prediction. For a single model, it is possible to create a general profile using all observations in a dataset or specific subsets with multiple profiles. These profiles assist in making complex black-box models more understandable and can also be used to compare different models. They succinctly summarize the effect of a particular explanatory variable on the dependent variable, making explanations more comprehensible (Biecek and Burzykowski 2021).

Application process of machine learning models

To evaluate the cross-validation (CV) performance of the models employed, the dataset was divided randomly into two parts: a 70% training set and a 30% test set. Before training, the data underwent pre-processing to eliminate variables with near-zero variance, linearize them using the Yeo-Johnson method (2000), normalize them by standard deviation, and create dummy variables for categorical variables.

To optimize the models, a range of 100 different combinations of fine-tuning parameters were tested on the training set using 10-fold cross-validation with 5 repetitions. The model with the most successful cross-validation value from the candidate models was selected for the next stage. Fine-tuning was performed using a standard grid search within the given parameter range. It is important to note that the LM algorithm has not undergone fine-tuning. The MARS algorithm's grid has a range of values for num_terms, from 2 to 5, and for degree, either 1 or 2. Additionally, it offers a selection of prune_method values, including 'backward', 'none', 'exhaustive', 'forward', or 'seqrep'. The algorithm was developed using 'cv' values.

The grid for fine-tuning ANN comprises a range of values for several parameters, including hidden units (hidden_units) from 1 to 10, a logarithmic penalty between log10(-10) and log10(0), epoch values (epochs) between 5 and 500, and a logarithmic learning rate (learn_rate) ranging from log10(-3) to log10(-0.5). The grid of the CART algorithm is built by utilizing the cost-complexity parameter (cost_complexity) with logarithmic values that range from log10(-10) to log10(-1), tree depth (tree_depth) that ranges from 1 to 15, and minimum node size (min_n) that ranges from 2 to 40.

Goodness of fit criteria were calculated on the test sets of the models with the most successful CV values on the training sets. Performances of the models were measured with goodness of fit criteria. Below are the formulas for the preferred goodness of fit criteria (Willmott and Matsuura 2005; Eyduran et al. 2017; Çelik 2019):

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (3)$$

Root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y - \hat{Y})^2} \quad (4)$$

Mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(\left|Y - \hat{Y}\right|\right) \quad (5)$$

Mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y - \hat{Y}}{Y}\right| . 100 \quad (6)$$

n is the number of cases in a set, Y is the real value of an output variable (NP), $\hat{Y}$ is the predicted value of an output variable (NP).

Models are ranked based on their measured RMSE values. The model with the lowest RMSE value for the test data set and with no significant difference between the RSME values of the training and test sets was accepted as the most successful model.

Variables Used in the Models

Dependent and independent variables in the data set are given in Table 3.

Table 3. Variables used in the models

Dependent Variable	**Type**	**Explanation**
Total net profit (TNP)	Continuous	$
Explanatory variables	**Type**	**Explanation**
Lactation year (LY)	Continuous	Year
Veterinary cost (VC)	Continuous	$
Milking method (MM)	Dummy	1: By Hand, 2: By Machine
Mastitis control (MC)	Dummy	1: Yes, 2: N0
Concentrated feed quantity (CFQ)	Continuous	Kg
Roughage feed quantity (RFQ)	Continuous	Kg
Lactation period (LP)	Continuous	Day
Milking time (MT)	Continuous	Minute
Milking interval(MI)	Continuous	Hour

Software Used

In this study, the R statistical environment (ver. 4.1.3) (Anonymous 2022) was preferred. The modeling was conducted through the tidymodels framework (ver. 1.1.0) (Kuhn and Wickham 2020). The LM model was trained using the R base library; the CART model was trained with the rpart package (ver. 4.1.19) (Therneau and Atkinson 2022). The MARS modeling was performed using the earth package (ver. 5.3.2) developed by Milborrow (2011). The ANN model was trained using the brulee package (ver. 0.2.0), developed for using the pytorch module in R (Kuhn and Falbel 2022). Variable importance metrics and partial dependence profiles for the models were created using the DALEX package (ver. 2.4.3) developed by Biecek (2018) and the DALEXtra package (ver. 2.3.0) created by Maksymiuk et al. (2020).

RESULTS AND DISCUSSION

As a result of the analyses, the most suitable models were determined based on their cross-validation RMSE values (CVRMSE). Among the 100 candidate models trained for each model, the fine-tuning parameters of the most successful models were identified. For instance, the MARS model had num terms = 4, prod degree = 1, prune method = 'cv' (CVRMSE = 162.11). For ANN, the parameters were hidden units = 8, penalty = 0.0408, epochs = 370, learn rate = 0.196 (CVRMSE = 218.18). The CART model had parameters cost complexity = 0.00209, tree depth = 15, min_n = 29 (CVRMSE = 218.11). The LM model had no fine-tuning parameters (CVRMSE = 222.17).

According to the goodness of fit criteria presented in Table 4, the MARS Model algorithm demonstrated the highest level of success with the lowest RMSE values in the test dataset. The performance values of the MARS model in the training set (RMSE: 160.661) and test sets (RMSE: 162.445) are not significantly different, which suggests that the model is not overfitting and is generalizable.

Table 4. Goodness-of-fit values

Model	Dataset	R^2	RMSE	MAE	MAPE
LM	train	0.639	206.143	148.374	15.258
	test	0.762	173.602	129.437	11.698
MARS	train	0.781	160.661	124.942	12.622
	test	0.776	162.445	131.771	11.823
MLP	train	0.716	184.065	135.921	13.680
	test	0.774	172.452	120.489	10.915
CART	train	0.676	195.391	146.404	15.296
	test	0.674	196.909	151.554	14.510

Upon analysis of the variable importance rankings of the models together (refer to Figure 2), it is clear that the two most significant variables affecting Net Profit in all models are LY and LP. When evaluating the models based on the variables affecting Net Profit, it is evident that only LY and LP have a significant impact. The ANN model, however, shows that CFQ and MC also have a partial effect, while in the CART and LM models, only CFQ has a partial effect. Nevertheless, it is important to note that removing these variables, as well as any variables other than LP and LY, does not affect the model performance.

It should also be noted that the Linear Regression (LM) model in this study was utilized solely for the purpose of comparing its performance with other machine learning models. As a result, the assumptions typically required for linear regression, such as normality, homoscedasticity, independence of residuals, and absence of multicollinearity among predictors, were not evaluated or validated. This approach was adopted to focus on the predictive accuracy and comparative performance of the models rather than ensuring the strict adherence of LM to its theoretical assumptions.

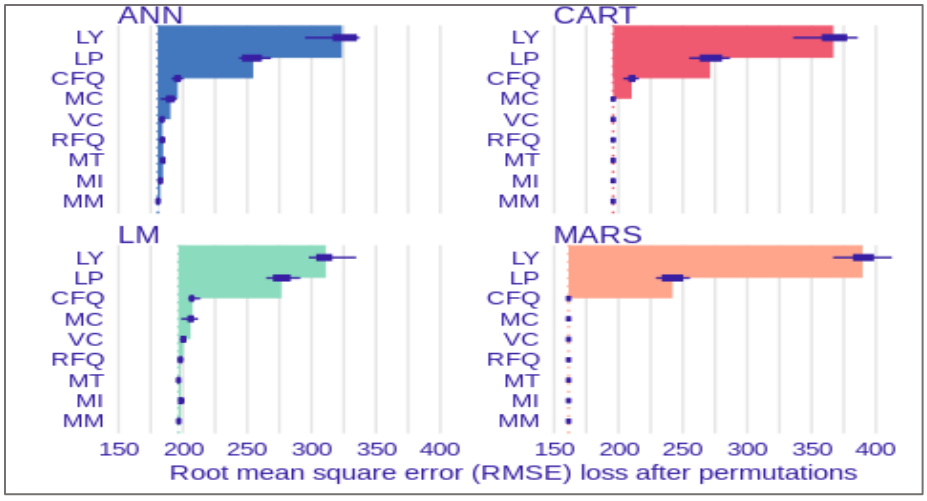

Figure 2. Variable importance plots generated for ANN, CART, LM, MARS models

In parallel with the results obtained in the variable importance plots, according to the partial-dependence profiles (Figure 3), it is observed that the changes in the variables CFQ, MI, MT, RFQ, VC have no effect on the predicted value of the MARS model (net profit). However, it was observed that the LY variable had a positive effect on total net profit up to a value of approximately 7.0, but a negative effect after this value. This suggests that in order to prevent the decrease in net profit, buffalo breeders should sell the buffaloes after the 7th lactation. There are very limited studies on the effect of lactation year on milk yield; Kaygısız (1999) stated that the highest yield from domestic buffalos was obtained in the 6th lactation and buffalos giving birth in the autumn season started lactation with a higher yield. Şahin et al. (2024) reported that the highest milk yield was obtained in Anatolian buffalos in the 3rd lactation between 90-120 days.

Call: earth(formula=y~., data=bake(prep(normal), train_data), pmethod="exhaustive", degree=1, nk=5)
coefficients
(Intercept) 1646.6552
h(1.49075-LY) -275.2942
h(LY-1.49075) -543.2853
h(0.372269-LP) -145.7062
h(LP-0.372269) 108.0481

Selected 5 of 5 terms, and 2 of 9 predictors (pmethod="exhaustive") Termination condition: Reached nk 5 Importance: LY, LP, VC-unused, CFQ-unused, RFQ-unused, MT-unused, Number of terms at each degree of interaction: 1 4 (additive model) GCV 32206.9; RSS 3819596; GRSq 0.7332048; RSq 0.7638881.

Based on the estimated coefficients obtained from the MARS model, the fitted equation can be expressed as follows:

$$\hat{y} = 1646.6552 - 275.2942 \cdot h(1.49075 - LY) - 543.2853 \cdot h(LY - 1.49075) - 145.7062 \cdot h(0.372269 - LP) + 108.0481 \cdot h(LP - 0.372269)$$

Where $h(z) = max(0, z)$ represents the positive part hinge function, and *LY* and *LP* are predictor variables included in the model.

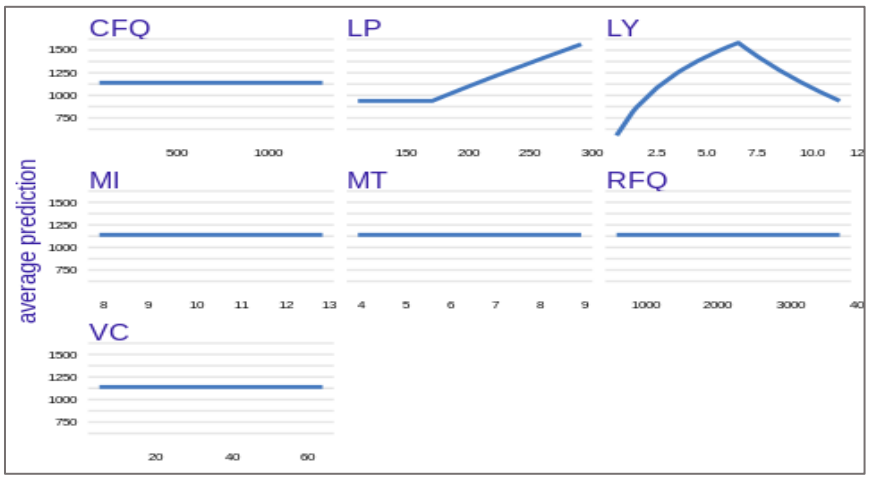

Figure 3. Partial-dependence profiles of the MARS model

The value of approximately 175 was determined by visually inspecting Figure 3, which presents the Partial Dependence Profiles of the MARS model. From this graph, it can be observed that increases in the LP variable have no significant effect on net profit until reaching around 175 days, after which net profit begins to increase linearly. Therefore, buffalo producers are advised to take measures to extend the lactation period beyond this threshold. While Kaygısız (1999) found the highest productive lactation period to be 124 days, he recommended planning optimum feeding and management programs to extend this period. Şekerden and Küçükkebapçı (1999) found that milk yield decreased after 210 days of lactation in Anatolian buffalos. Galsar et al. (2016) found the lactation period of Meshana Buffalo in India to be 298 days, Soysal et al. (2018) found the lactation period of Anatolian buffalo to be 231 days, Yılmaz Adkinson and Konca (2021) compared the lactation periods and milk yields of different buffalo breeds in the world and found that lactation periods varied between 210-350 days, Yenilmez et al. (2022) found the lactation period of Italian Mediterranean buffalos to be 247 days.

Luna-Polomera et al. (2021) compared Mixed nonlinear models in Murrah buffalos for 3 lactation periods (180 days, 210 days and 240 days) and concluded that the model with the highest efficiency and best fit at 180 days was WOOD. In addition, regarding water buffalo breeding and increasing profitability, Sweers et al. (2014) found that calf-calf interval had the highest impact on total economic performance in reproductive performance in water buffalo breeding in Germany, Sabia et al. (2015) found that the demand for water buffalo dairy products was increasing in many countries, more efforts were needed in the genetic evaluation of water buffalo populations and that one of the biggest challenges faced by the sector was low profitability. Işık and Gül (2016) reported that feed costs affect profitability the most in water buffalo breeding in Muş province and that farmers should be informed about modern techniques so that producers can earn more by increasing milk yield. Aydoğdu and Şahin (2022) stated that pure breeding and breed development activities should be expanded to increase meat and milk yield in buffalo breeding in Turkey.

CONCLUSION

This study has provided significant insights into the factors influencing profitability in buffalo milk production, leveraging advanced data mining algorithms to identify key determinants. Among the models tested, MARS model demonstrated superior performance in predictive accuracy and robustness, based on cross-validation and goodness-of-fit criteria. The analysis identified lactation year and lactation period as the most critical variables affecting net profit, with profitability reaching its peak during the seventh lactation year and increasing linearly for lactation periods exceeding 175 days. These findings emphasize the importance of implementing evidence-based management strategies in buffalo production systems.

The threshold of 175 days for the lactation period is particularly significant, as it marks the point beyond which net profit begins to increase more substantially. This suggests that efforts to extend the lactation period beyond this duration can directly enhance profitability. Practically, buffalo producers should focus on nutritional, health, and environmental management practices that support longer lactation cycles. By targeting this critical period, producers can optimize milk yield and improve overall economic returns, making the 175-day lactation period a key performance indicator in sustainable buffalo farming.

This research also underscores the utility of data mining methodologies in agricultural studies, offering a robust framework for analyzing complex relationships between variables and profitability. The results contribute to the growing body of knowledge on livestock management and provide actionable recommendations for improving production efficiency. Future studies should seek to validate these findings across diverse geographical and management contexts while exploring additional variables to further refine strategies for enhancing productivity and profitability in buffalo farming systems.

Declaration of interests

There is no conflict of interest among the authors of the article.

Author Contributions

***Köksal KARADAŞ*:** Conceptualization, data curation, formal analysis, investigation, methodology, writing and review.
Osman Doğan BULUT: Formal analysis, investigation, methodology, writing, review and editing.
Hakan DUMAN: Methodology, software, data analysis and writing.

ORCID

Köksal KARADAŞ http://orcid.org/0000-0003-2682-6356
Osman Doğan BULUT http://orcid.org/0000-0003-1176-3313
Hakan DUMAN http://orcid.org/0000-0001-6166-5776

Article History	
Submission received:	09.04.2025
Revised:	24.06.2025
Accepted:	27.06.2025

REFERENCES

Akin, M., Hand, C., Eyduran, E., & Reed, B. M. (2018). Predicting minor nutrient requirements of hazelnut shoot cultures using regression trees. *Plant Cell, Tissue and Organ Culture*, 132(3), 545–559. https://doi.org/10.1007/s11240-017-1353-x.

Anonymous. (2022). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Arthur, C. K., Temeng, V. A., & Ziggah, Y. Y. (2020). Multivariate adaptive regression splines (MARS) approach to blast-induced ground vibration prediction. *International Journal of Mining, Reclamation and Environment,* 34(3), 198–222.

Atasever, S., & Erdem, H. (2008). Manda yetiştiriciliği ve türkiye'deki geleceği. *Journal of the Faculty of Agriculture,* 23(1), 59–64.

Aydoğdu, M. A., & Şahin, Z. (2022). Analysis of the recent periods of changes in water buffalo presence and milk production quantities in Turkey. *International Journal of Social, Humanities and Administrative Sciences,* 8(51), 612–616.

Balhara, S., Singh, R. P., & Ruhil, A. P. (2021). Data mining and decision support systems for efficient dairy production. *Veterinary World,* 14(5), 1258–1262.

Becskei, Z., Savić, M., Ćirković, D., Rašeta, M., Puvača, N., Pajić, M., Đorđević, S., & Paskaš, S. (2020). Assessment of water buffalo milk and traditional milk products in a sustainable production system. *Sustainability,* 12(16), 1–13.

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research,* 19(84), 1–5.

Biecek, P., & Burzykowski, T. (2021). Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. With Examples in R and Python. New York: Chapman and Hall. https://pbiecek.github.io/ema/.

Boehmke, B., & Greenwell, B. (2020). Hands-on Machine Learning with R. Chapman and Hall/CRC.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. New York: Chapman & Hall.

Canbolat, Ö. (2012). Buffalo breeding and current situation in Turkey. Journal of Tarım Türk, 30, 176–180.

Çelik, Ş. (2019). Comparing predictive performances of tree-based data mining algorithms and mars algorithm in the prediction of live body weight from body traits in pakistan goats. *Pakistan Journal of Zoology,* 51(4), 1447–1456. https://doi.org/10.17582/journal.pjz/2019.51.4.1447.1456.

Çelik, Ş., & Yilmaz O. (2018) Prediction of Body Weight of Turkish Tazi Dogs using Data Mining Techniques: Classiication and Regression Tree (CART) and Multivariate Adaptive Regression Splines (MARS). *Pakistan Journal of Zoology,* 50(2),575-583.

Eyduran, E., Zaborski, D., Waheed, A., Çelik, Ş., Karadaş, K., & Grzesiak, W. (2017). Comparison of the predictive capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the indigenous beetal goat of pakistan. *Pakistan Journal of Zoology,* 49(1).

FAO. (2022). Food and Agriculture Organization of the United Nations. Crops and Livestock Products. https://www.fao.org/faostat/en/#data/QCL

Faraz, A., Tirink, C., Eyduran, E., Waheed, A., Tauqir, N. A., Nabeel, M. S., & Tariq, M. M. (2021). Prediction of live body weight based on body measurements in Thalli sheep under tropical conditions of Pakistan using CART and MARS. *Tropical Animal Health and Production,* 53(2), 301. https://doi.org/10.1007/s11250-021-02748-6.

Fox, J., & Weisberg, S. (2019). An R Companion to Applied Regression. London: SAGE.

Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 1–67.

Galsar, N. S., Shah, R. R., Gupta, J. P., Pandey, D. P., & Patel, K. B. (2016). Analysis of first production and reproduction traits of Mehsana buffaloes maintained at tropical and semi-arid region of Gujarat, India. *Life Sciences Leaflets*, 4297(77), 65–75.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

Guo, M., & Hendricks, G. (2010). Improving buffalo milk. In M. Griffiths (Ed.), Improving the safety and quality of milk (Vol. 2, pp. 402–416). Woodhead Publishing.

Işik, M., & Gül, M. (2016). Economic and social structures of water buffalo farming in Muş Province of Türkiye. i]Revista Brasileira de Zootecnia, 45(7), 400–408.

Kaygısız, A. (1999). Lactation curve traits of native buffaloes. *Tarım Bilimleri Dergisi,* 5(1), 1–8.

Khedkar, C. D., Kalyankar, S. D., & Deosarkar, S. S. (2016). Buffalo milk. In B. Caballero, P. Finglas, & F. Toldrá (Eds.), The encyclopedia of food and health 1, 522–528. Academic Press.

Konca, Y., & Yılmaz Adkinson, A. (2021). Manda eti üretimi ve kalite özellikleri Water buffalo meat production and quality characteristics]. European Journal of Science and Technology, 31(1), 420–428.

Kuhn, M., & Falbel, D. (2022). Brulee: High-level modeling functions with 'Torch' [Computer software]. https://CRAN.R-project.org/package=brulee

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles [Computer software]. https://www.tidymodels.org

Luna Palomera, C., Dominguez-Viveros, J., Aguilar-Palma, G. N., Castillo-Rangel, F., Sanchez-Dávila, F., & Macias-Cruz, U. (2021). Analysis of the lactation curve of Murrah buffaloes with mixed non-linear models. Chilean Journal of Agricultural & Animal Sciences (ex Agro-Ciencia), 37(1), 200–208.

Maksymiuk, S., Gosiewska, A., & Biecek, P. (2020). Landscape of R packages for explainable artificial intelligence. arXiv. https://arxiv.org/abs/2009.13248

Mane, B. G., & Chatli, M. K. (2015). Buffalo milk: Saviour of farmers and consumers for livelihood and providing nutrition. *Agricultural and Rural Development*, 2, 5–11.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5, 115–133. https://doi.org/10.1007/BF02478259

Milborrow, S. (2011). Earth: Multivariate adaptive regression splines [Computer software]. http://CRAN.R-project.org/package=earth

Nwanganga, F., & Chapple, M. (2020). Practical machine learning in R. Wiley. https://doi.org/[DOI if available]

Okut, H., Wu, X.-L., Rosa, G. J. M., Bauck, S., Woodward, B. W., Schnabel, R. D., Taylor, J. F., & Gianola, D. (2013). Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. *Genetics Selection Evolution,* 45(1), 34. https://doi.org/10.1186/1297-9686-45-34

Park, W. Y., & Haenlein, G. F. W. (2008). Buffalo milk: Utilization for dairy products. In Y. W. Park & G. F. W. Haenlein (Eds.), Handbook of milk of non-bovine mammals (pp. 195–274). Wiley-Blackwell.

Pudja, P., Djerovski, J., & Radovanović, M. (2008). An autochthonous Serbian product—Kajmak characteristics and production procedures. *Dairy Science and Technology,* 88, 163–172.

Sabia, E., Napolitano, F., Claps, S., Braghieri, A., Piazzolla, N., & Pacelli, C. (2015). Feeding, nutrition and sustainability in dairy enterprises: The case of Mediterranean buffaloes. In A. Vastola (Ed.), The sustainability of agro-food and natural resource systems in the Mediterranean basin (pp. 57–64).

Şahin, A., Aksoy, Y., Ulutaş, Z., Yıldırım, A., & Sarıkaya, Ö. (2024). Anadolu mandalarının ilk üç laktasyonlarına ait laktasyon eğrisi parametrelerinin ve eğri şeklinin belirlenmesi [Determination of lactation curve parameters and curve shape for the first three lactations of Anatolian buffaloes]. *Journal of Animal Sciences and Products,* 7(1), 12–18.

Sahraei, M. A., Duman, H. Muhammed Y., & Eyduran, E. (2021). Prediction Of Transportation Energy Demand: Multivariate Adaptive Regression Splines *Energy,* 224.

Sarıözkan, S. (2011). Türkiye'de manda yetiştiriciliği'nin önemi [The importance of buffalo breeding in Türkiye]. *Kafkas Üniversitesi Veteriner Fakültesi Dergisi*, 17(1), 163–166.

Soysal, M. İ., Genç, S., Aksel, M., Ünal, E. Ö., & Gürcan, E. K. (2018). Effect of environmental factors on lactation milk yield, lactation length and calving interval of anatolian buffalo in Istanbul. *Journal of Animal Science and Products*, 1(1), 93–97.

Sweers, W., Möhring, T., & Müller, J. (2014). The Economics Of Water Buffalo (Bubalus Bubalis) Breeding, Rearing And Direct Marketing. *Archiv Tierzucht*, 57(22), 1–11.

Therneau, T., & Atkinson, B. (2022). Rpart: Recursive Partitioning And Regression Trees [R package]. https://CRAN.R-project.org/package=rpart

Titterington, M. (2010). Neural networks. *WIREs Computational Statistics*, 2(1), 1–8. https://doi.org/10.1002/wics.50

Toparslan, E., & Mercan, L. (2018). Türkiye yerli manda popülasyonlarında yapılan moleküler genetik çalışmalar [Molecular genetic studies on native water buffalo populations in Türkiye]. *Academia Journal of Engineering and Applied Sciences,* Special Issue.

Turkish Statistical Institute (TÜİK). (2022). Data portal for statistics: Consumer price index. https://data.tuik.gov.tr/Bulten/Index?p=Tuketici-Fiyat-Endeksi-Aralik-2022-49651

Turkish Statistical Institute (TÜİK). (2023). Data portal for statistics: Livestock statistics. https://biruni.tuik.gov.tr/medas/?kn=79&locale=tr

Weber, G.-W., Batmaz, I., Köksal, G., Taylan, P., & Yerlikaya-Özkurt, F. (2012). CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering,* 20(3), 371–400.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1), 79–82.

Yenimez, K., Doğa, H., & Özbaşer, F. T. (2022). environmental factors influencing milk yield and lactation length in italian mediterranean buffaloes in Türkiye. *Journal of the Hellenic Veterinary Medical Society*, 73(3), 4296–4302.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *biometrika*, 87(4), 954–959.

Yılmaz Adkinson, A., & Konca, Y. (2021). Sütçü manda ırklarının performans ve verimliliğini etkileyen faktörler ve Türkiye'deki geleceği [Factors affecting the performance and productivity of dairy buffalo breeds and their future in Türkiye]. European Journal of Science and Technology, 25, 498–508.

Yılmaz, O., & Ertuğrul, M. (2011). Domestic livestock resources of Türkiye: water buffalo. *Tropical Animal Health and Production,* 44,707-714. https://doi.org/10.1007/s11250-011-9957-3

Zhang, J., Liu, Z., Shi, Z., Jiang, L., & Ding, T. (2023). Milk yield prediction and economic analysis of optimized rearing environment in a cold region using neural network model. *Agriculture,* 13. https://doi.org/10.3390/agriculture13122206