TÜBİTAK

Turkish Journal of Veterinary and Animal Sciences

http://journals.tubitak.gov.tr/veterinary/

Research Article

Turk J Vet Anim Sci
(2022) 46: 809-819

doi:10.55730/1300-0128.4257

# Detection of correct pregnancy status in lactating dairy cattle using MARS data mining algorithm

**Demet ÇANGA**[1], **Mustafa BOĞA**[2,*]

[1]Department of Chemistry and Chemical Processing, Bahçe Vocational School, Osmaniye Korkut Ata University, Osmaniye, Turkey
[2]Department of Food Processing, Bor Vocational School, Niğde Ömer Halisdemir University, Niğde, Turkey

**Received:** 30.04.2022 • **Accepted/Published Online:** 26.11.2022 • **Final Version:** 08.12.2022

**Abstract:** In this study, it is aimed to determine pregnancy outcomes by using multivariate adaptive regression splines (MARS) algorithm for classification type problems. For this purpose, data obtained from a private dairy farm in the Konya region of Türkiye in 2020 were used to determine pregnancy outcomes in Holstein dairy cattle. It has been determined how to perform statistical analyses on solving classification-type problems with the MARS algorithm and how to use R packages (caret and earth) by creating an R script file. After the analysis, the MARS estimation equation was created and in finding the probability of being pregnant: While lactation period, cow age, number of lactations, insemination number, and total lactation milk yield variables are important, it was seen that 7-day mean milk yield and last lactation milk yield were not significant. Using the train function of the caret package, the number of terms that produce the highest accuracy and the degree of interaction are determined. Goodness-of-fit tests of the optimum model were calculated. Within the scope of the evaluation of the generalization ability of the model, training and test sets were created, the classification success graph of the MARS algorithm, the model building phase were summarized, and the generalization ability of the established model was measured. When the pregnancy status is taken as a positive reference, the correct classification rate (sensitivity) of the animal with positive pregnancy status was found to be 0.9574. The correct classification rate (specificity) of pregnant animals was found to be 0.8370. The overall classification ratio of the training set (accuracy) was found to be 0.8777. The area under the ROC curve (AUC) was found to be 0.947, which indicates that the optimum specificity value is close to 1.

**Key words:** Logistic regression, classification, binary data, train and test set, Holstein breed

## 1. Introduction

Data mining (DM) is an information technology that is gaining momentum today and is also technologically advancing and is defined as a method applied to reveal the desired important information hidden in large data sets [1]. It can be recommended to use data mining algorithms, especially in case of violation of the distributional assumptions regarding the examined variables. CART, CHAID, Exhaustive CHAID, MLP, and naive Bayes, artificial neural networks (ANN), and MARS are examples of data mining algorithms. In the field of livestock, these algorithms, which were used for classification purposes in previous studies, are especially important [2]. Using the most appropriate data mining algorithm on subjects such as characterization of breeds, characterization of sex, characterization of progeny ratio, characterization of mastitis will be important in terms of developing the right strategies in animal husbandry. Therefore, the use of the most appropriate algorithm for classification contributes to the breeders in terms of defining the breed standards of the breeds studied [2]. In other words, these algorithms can be used for two-level dependent variables (sex, birth type, pregnancy status) in the field of livestock. Different methods such as classification functions (CF), ANN, multivariate adaptive regression curves (multivariate adaptive splines, MARS), logistic regression (LR) and classification trees and discriminant analysis are used to classify the research model [3]. Logistic regression is considered one of the most popular approaches for classification of binary data [4]. Researchers from various disciplines such as statistics, machine learning, and data mining have engaged in classification using logistic regression from available data. LR to study successful pregnancy in cows and buffaloes involving animal husbandry has been used successfully in many problems such as detecting lameness in cows and clinical mastitis [5,6]. Süt and Şimşek [7] compared six different decision tree algorithms (classification and regression tree (CART),

* Correspondence: mboga@nigde.edu.tr

CHAID, exhaustive-CHAID, QUEST and boosted tree classifiers and regression (BTCR)) with each other in terms of classification performance to estimate the death rate resulting from head injury accidents. The performances of the evaluated algorithms were compared using criteria such as sensitivity, specificity, positive/negative predictive, and accuracy rate. In addition, the areas under the ROC curve of all algorithms were estimated. Grzesiak et al. [5] used naive Bayes classification (NBC) and CART methods to determine the effect of factors affecting fertility in dairy cattle. In the study, lactation number, artificial insemination season, cow's insemination age, rate of HF genes in cows, pregnancy rate, gestation period, milk protein and fat yield, and sex at previous calving were taken as independent variables. The dependent variable was evaluated as binary (one cow conception after one or two artificial inseminations) and poor (one cow conception after more than two artificial inseminations) progeny. Piwczynski [8] evaluated 6586 heads of Polish Merino aged 2 to 8 years in terms of reproductive performance index from ten herds in Pomarze and Kujavay region of Poland. In the study, CART classification algorithm was used to define the variables responsible for the variation in the number of lambs obtained from the mated sheep. It was determined that the most important independent variables were maternal age, herd, birth type, and 16th month live weight of sheep. In addition, it was reported that the factors included in the model (herd, maternal age, birth type, body weight) were effective on the number of lambs obtained per mated sheep. Piwczynski et al. [9] used some classification algorithms to determine the effects of factors affecting stillbirths and calving ease in a population of 1257 Holstein cows. In this study, classification trees obtained using CART and QUEST algorithms were evaluated according to three separation criteria (Pearson's chi-squared, entropy function, and Gini index) and five goodness-of-fit criteria. In the study, when the order of importance of the variables that affect the ease of calving is examined, it has been reported that the live weight, lactation order, rearing system, length of gestation, and calf sex are followed, respectively. It was determined that only calf birth weight variable was effective on stillbirth. Yılmaz [2] compared the prediction performance of some data mining algorithms in terms of birth type in sheep. For this purpose, CART, CHAID, exhaustive CHAID, naive Bayes, and MLP algorithms have been applied. Accuracy (%), sensitivity, specificity, and area under the ROC curve were calculated to find the best algorithm within the prediction performance. Within the scope of the research, sex (female and male), farms (Mastung, Quetta, Noshki), maternal age, birth weight, and lambing season and year were used as independent variables. It was observed that all algorithms used in the research showed superior performance.

MARS method, commonly known as a modification of the CART algorithm, is a powerful data mining tool for solving regression and classification problems. In the study on MARS by Grzesiak and Zaborski [10], the best model was used to predict pregnancy in cows in the test set and validate the quality of this prediction. It is practically important to determine the factors affecting the status of the two-level dependent variable and to set the standards for the breed of the animal studied. When the literature is examined, it has not been determined how to perform statistical analyses to be solved with MARS algorithm instead of LR in classification-type problems involving two-level (sex, birth type, pregnancy status) dependent variables in the field of livestock. In addition, no studies were found on which R packages (earth and caret) to be used and how to create training and test sets within the scope of evaluating the generalization ability of the model [2]. However, it has been determined that there is no practical and application-oriented detailed information about the interpretation of the area under the ROC curve using MARS and R, the modeling phase, and the measurement of the generalization ability of the established model [11].

In this study, a classification approach is presented for binary data in which logistic regression is updated with MARS to address this deficiency. The model in the study was used to predict conception in cows in the test set and to validate the quality of this prediction. In this respect, the classification performance of the MARS algorithm for the pregnancy status of Holstein cattle used in the study was performed. To determine algorithm, criteria such as correct classification rate, sensitivity, specificity, and area under the ROC curve were calculated.

## 2. Materials and methods

### 2.1. Materials

The animal data of the experiment were obtained from a private farm named Gökcan Agriculture and Livestock in Konya-Karapınar. These data consist of milk yield records of 172 head dairy cows of Holstein breed for 2020. While animal concentrated feed raw materials are soybean meal, sunflower seed meal, corn, barley, razmol carob, soybean husk, molasses; roughage raw materials consist of alfalfa grass, corn silage, and sorghum Sudan grass. The rations of these animals are made by the relevant engineer on the farm using these feed materials. The automatic drinkers of the animals are adjusted. Thus, it was ensured that the daily water needed by the animal was always in front of them and was available individually and freshly.

The classification variable in the study was coded as a binary variable. Pregnancy status was accepted as positive and coded as "1", other statuses were negative (fresh, inseminated, and cut) and coded as "0". The dependent

variable used to perform the classification in the study was determined as pregnancy status (PS). PS was determined for dependent variable used to perform the classification in the study. The number of lactations (LN), lactation period (LP), insemination number (IN), cow age (AGE), 7-day mean milk yield (SDMY), total lactation milk yield (TMY), and last lactation milk yield (LLMY) were determined for independent variables.

### 2.2. Method

#### 2.2.1. MARS algorithm

The MARS method proposed by Friedman [12] can be defined as a combination of machine learning with a purely classical approach. It is the selected weighted sum of spline functions (or basis functions) used to generate the variation of individual explanatory variables [13]. The resulting model may be additive or involve interactions between variables. MARS does not make any assumptions about the basic functional relationships between dependent and independent variables. This model is a generalized additive model and is based on the divide-and-conquer§ strategy. It divides the training and test datasets into separate linear splines with different gradients. In the model, the variable is included in the model and has different weight coefficients and different signs depending on whether it is above or below a certain threshold. In general, the splines are interconnected in the model, and these piecewise curves, also called basis functions (BF), produce a flexible model that can handle linear and nonlinear behavior. The connection points between the parts produced by this model are called nodes. By marking the end of one data field and the beginning of another data field, candidate nodes are placed at random locations within each input variable range [14]. MARS generates BF by progressively searching for all possible univariate candidate nodes and the interaction between all variables. It allows the adaptive regression algorithm to automatically select the node location. The MARS algorithm, which includes a forward stage and a reverse stage, places candidate nodes at random locations within the range of each predictive variable to identify a pair of $BF_i$ in the forward stage. At each step, the model adjusts the nodes and appropriate BF pairs to minimize residuals in the sum of squares. This BF addition continues until the maximum number is reached and often results in a model that is too complex and too cohesive. The backward step involves deleting the redundant BF that contributes the least to the model's goodness of fit. Each main function is then refitted, and each reduced suboptimal model is tested with the generalized cross-validation (GCV) method to avoid overfitting. The model with the lowest GCV score is considered the best [15]. Multivariate adaptive regression spline data mining algorithm was conducted as described by Çelik and Yılmaz [14]. Briefly, the MARS algorithm was used as follows:

$$\hat{y}= \beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_m}h_{km}\left(X_{v(k,m)}\right) \quad (1)$$

where $\hat{y}$ is the predicted value of the dependent variable, $\beta_0$ is the constant in the MARS prediction equation (intercept), $\beta_m$ is the coefficient of the fundamental functions in the MARS estimation equation, $h_{km}$ $(X_{v(k.m)})$ is the basic function in the MARS estimation equation, v(k,m *)* is the index of the independent variable used in the $m^{th}$ component of the kth factor, and $K_m$ is parameter value that limits the degree of interaction [14, 16]. The pruning algorithm is done with the GCV technique [16–22]. GCV considers both errors and model complexity.

$$GCV(\lambda)=\frac{\sum_{i=1}^{n}(y_i-y_{ip})^2}{\left[1-\frac{M(\lambda)}{n}\right]^2}, \quad (2)$$

where n is the number of observations in the data set, $y_i$ is the dependent variable value of the observation value, $y_{ip}$ is dependent variable predictive value of the observation value, and M(λ) is the penalty function for the complexity of the model containing λ terms. The MARS algorithm is constructed with piecewise linear basis functions ($BF_i$) of the following form:

$$BF_1=\max(0,x-t)\begin{cases}x-t, & x>t\\ 0, & x\le t\end{cases} \quad (3a)$$

$$BF_2=\max(0,t-x)\begin{cases}t-x, & t>x\\ 0, & t\le x\end{cases} \quad (3b)$$

where x is the variable range and t is the node. The MARS model is a linear combination of fundamental functions:

$$Y_i=a_0+a_1BF_1+a_2BF_2+\ldots+a_iBF_i \quad (4)$$

where $Y_i$ is the dependent variable, $a_0$ is the intersection, and $a_1$ and $a_2$ are the coefficients of the respective principal functions [15–17, 21, 23]. The MARS model is a linear combination of basis functions: An example of three-part linear functions or curves ($BF_1$, $BF_2$, and $BF_3$) interconnected at two points or nodes created by a simple RStudio software [24] for better understanding of the model is shown in Figure 1.

Splines (MARS) model is represented by a three-part linear basis function ($BF_1$, $BF_2$, and $BF_3$) connected by nodes (shown in blue) and where $BF_1$ = max (0, x-3) and BF2 = max (0, 3 - x) and $BF_3$ = max (0, x -6). In this case, the nodes are t = 3 and 6. These two nodes limit the input range x in three regions where different linear relationships are detected between the response and x explanatory variable (Figure 1) [18,25–27]. In addition, in the study,

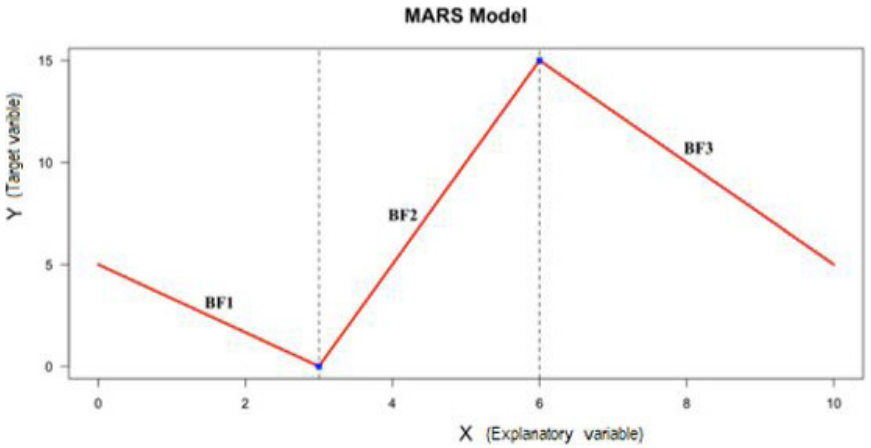


**Figure 1**. A graphical representation of a simple multivariate adaptive regression splines model. As shown in red line represented by a three-part linear basis function ($BF_1$, $BF_2$, and $BF_3$) connected by nodes. Nodes show regions of relationship change between the explanatory and target variable. These two nodes limit the input range for the variable X in three regions where different linear relationships between the response and the explanatory variable were detected.

the optimum number of terms (*nprune*) that provides the highest accurate classification rate (accuracy) was found with the *train* function of the *caret* package. Determining the degree of interaction, how R command lines should be created has been shown in detail [25]. The probability of being pregnant (positive) within the scope of binary logistic regression analysis:

$$P(\text{Pregnancy(positive)}=1)=\frac{\exp^{GLM1}}{1+\exp^{GLM1}} \quad (5)$$

**2.2.2. ROC curve**

Receiver operating characteristics (ROC) analysis used in research is a method that is frequently used to measure the performance of research data or to compare the performance of more than one research data. The area under the ROC curve is used as an important criterion for the accuracy of research tests [2, 18]. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a given decision threshold. A test in perfect classification (no overlap in the two distributions) has an ROC curve passing through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test is [28, 29]. The ROC curve graph used in classification-type surveys is a graph used to summarize the performance of the classifier over all possible values (Figure 2).

It is generated by plotting the ratio of true-positive values (sensitivity) (x-axis) versus the ratio of false-positive values (specificity) (y-axis) when you change the threshold for assigning observations to a particular class. The ROC curve is used to generate a sensitivity/specificity report. The area under the curve (AUC) is a measure of how well a parameter can be distinguished between two classes [25, 29, 30].

**2.2.3. Calculation of confusion matrix and statistics value**

A confusion matrix is a table often used to describe the performance of a classification model on a set of test data for which the actual values are known. An example confusion matrix is given in Table 1.

Equations based on a confusion matrix related to the ROC analysis used in the research are given as follows [2, 6, 20, 29, 30]:

$$\text{Accuracy}=\frac{A+D}{A+B+C+D} \quad (6a)$$

$$\text{Sensitivity}=\frac{A}{A+C} \quad (6b)$$

$$\text{PPV}=\frac{(\text{Specificity}^*(1\text{-prevalence}))}{((1\text{-Sensitivity})^*\text{prevalence})+^*(1\text{-prevalence})))} \quad (6c)$$

$$\text{NPV}=\frac{(\text{Sensitivity}^*\text{prevalence})}{((1\text{-Sensitivity})^*\text{prevalence})+((\text{Specificity})^*(1\text{-prevelance}))} \quad (6d)$$

where positive predictive value is expressed as PPV and negative predictive value is expressed as NPV.

$$\text{Prevalence}=\frac{A+C}{A+B+C+D} \quad (6e)$$

$$\text{Detection Prevalence}=\frac{A+B}{A+B+C+D} \quad (6f)$$

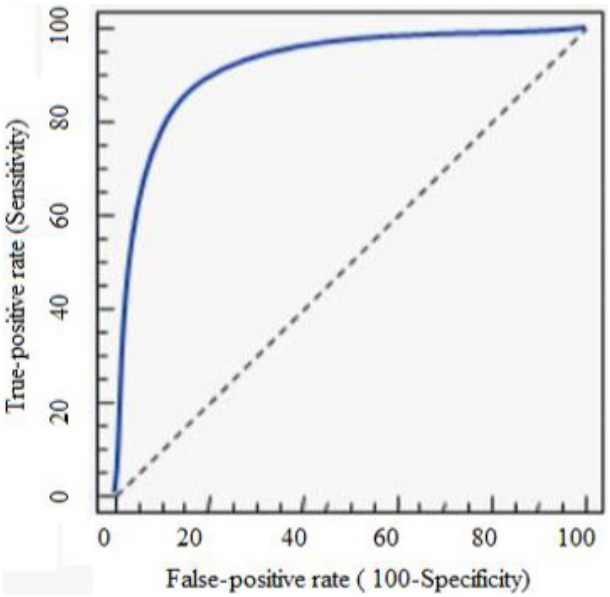


**Figure 2**. ROC curve plotted with (100-Specificity), selectivity values corresponding to the sensitivity value.

$$\text{Balanced Accuracy} = \frac{(\text{Sensitivity}+\text{Specificity})}{2} \quad (6g)$$

$$\text{Detection Rate} = \frac{A}{A+B+C+D} \quad (6h)$$

### 2.2.4. Statistical analysis

Descriptive statistics of quantitative and qualitative features were estimated using the "psych" package from the R package [31]. In the study, R-software [ 24] was used to generate insemination results using other independent variables related to milk yield in Holstein dairy cattle.

**Table 1.** An example model confusion matrix.

| | **Estimated value** | |
|---|---|---|
| **Real value** | **1** | **0** |
| **1** | (A) | (B) |
| **0** | (C) | (D) |

## 3. Results

Lactation ranks of the cows within the scope of the evaluation in the dairy cattle farm, lactation milk yield and descriptive values for each lactation group age are given in Table 2.

In the study, MARS approach was applied instead of binary logistic regression analysis to perform the classification situation for 8 dependent and independent variables in 172 data of dairy cattle. For this purpose, an R script file was created. To see the structure of the focused data set, *str (data)* should be defined in the R Console or R Script window. *Data $PS <-as.factor (data$PS)* has been defined to save the related variable as a factor variable. With the definition of table (data$PS) in R Console or Script window, it was determined that the number of nonpregnant (negative: 0) cows was 114 heads and the number of pregnant (positive: 1) animals was 58 heads [25]. It is seen that the percentage of being pregnant is (58 / 172) × 100 = 34%. Necessary definitions are made to see the ratio of pregnant (positive/negative) animals assigned to the training and test set, respectively. In addition, the ratio of pregnant (positive/negative) animals for the training and test set should be the same balanced. This is an important condition for the data set on study. The positive/negative (92:47 = 1.96) ratio of the pregnancy

**Table 2.** Descriptive statistics for the variables.

| **Variables** | **N** | **Min** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max** |
|---|---|---|---|---|---|---|---|
| LP (days) | 172 | 3 | 93 | 162 | 191 | 223.2 | 1036.0 |
| LN | 172 | 1 | 1 | 2 | 2.314 | 3 | 6 |
| AGE (months) | 172 | 24 | 33 | 45.50 | 50.61 | 65.25 | 113.00 |
| PS | 172 | 0 | 0 | 0 | 0.3372 | 1 | 1 |
| SDMY (kg) | 172 | 0.00 | 19.02 | 26.90 | 25.76 | 32.00 | 51.30 |
| TMY (kg) | 172 | 39 | 2840 | 4034 | 4655 | 6062 | 18070 |
| LLMY (kg) | 172 | 0 | 12 | 5882 | 4710 | 8114 | 12605 |
| IN | 172 | 0 | 0.750 | 1 | 1.884 | 3 | 10 |

LP: lactation period (days); LN: number of lactations; AGE: cow age (months); SDMY:7-day mean milk yield (kg); PS: pregnancy Status; TMY: total lactation milk yield (kg); LLMY: last lactation milk yield (kg); IN: number of inseminations

status in the training set and the positive/negative (22:11 = 2) ratio in the test set are calculated, and as can be seen in the calculation result, it is balanced/proportional [25]. Here, with the help of the initial split function in the R Script window, $p = 0.70$ is defined to divide the training and test set by 70% and 30%). In the research, the number of terms producing the highest accuracy and the degree of interaction were determined by using the train function of the caret package. As a resampling method in the research, a definition of cross-validation of 10 was selected. For the MARS model, the model with the best performance was determined between the 2- and 40-term MARS candidate models created for the 1st, 2nd, and 3rd interaction degrees. According to the results of the analysis, it is seen that the most suitable MARS model with the highest accuracy level created by the train function of the caret package includes 9 terms (nprune = 9) and quadratic interaction (degree = 2) terms (Table 3). How to calculate the probabilities of being pregnant (1) or not pregnant (0) belonging to the dependent variable levels of each individual forming the researched data set with the MARS estimation equation can be shown by using Equation (5) [20, 25].

According to the MARS estimation equation, while the constant suffix was found to be 3.961, 3 basic variables had a positive effect, and 5 basic variables had a negative effect. The BF2 variable of the largest positive effect was 0.091 and the BF3 basic variable was the largest negative effect with –25.414. When Table 3 was evaluated, it was seen that while LP, IN, AGE, LN, TMY variables were important in finding the possibility of being pregnant, other variables were not. As a result of the analysis, the MARS estimation equation was found as:

GLM = 3.9614 - 0.051 × max (0, 223 - LP) + 0.0911 × max (0, LP - 223) - 25.4142 × max (0, 1 - IN) - 1.4577 × max (0, IN - 1) + 0.016 × max (0, LP - 223) × LN - 0.0019 × max (0, LP - 223) × AGE + 0.0505 × max (0, 223 - LP) × max (0, 1 - IN)- 0.1334 × max (0, 4305 - TMY) × max (0, IN - 1) .

The graph showing the variation of the accuracy rate according to the number of terms and degree combinations is given in Figure 3.

When Figure 3 is examined, it is seen that the highest accuracy can be obtained with the 9-term MARS model containing first-degree interaction (degree = 2) [2, 11, 20, 25]. The overall accuracy rate of the cross-validation for the training data set was found to be 0.8489. It was understood that 66.2% of the pregnancy status of the 92 heads of the training set was negative and the remaining 33.8% was positive. With cross-validation, 58.3% of the nonpregnant animals, which constitute 66.2% of the training set, were classified correctly and 7.9% were incorrectly classified. Likewise, 26.6% of the animals that were pregnant, 33.8% of the training set were classified correctly, and the remaining 7.2% were incorrectly classified [2, 11, 25].

To draw the ROC curve and find the area under the ROC curve, R definitions and the "*library (pROC)*" package were used. The representation of ROC curves for individual classifiers, which includes both the specificity and the complement of sensitivity, is shown in Figure 4.

The fact that the area under the ROC curve is close to the value of "1" is a sign that the match between specificity and sensitivity is perfect. However, to mention that the classification performance of the MARS algorithm is particularly good, it is necessary to look at the performance

**Table 3.** Coefficients of the MARS model and results of MARS analysis.

| **Basis Functions (BFi)** | | **Coefficients** |
|---|---|---|
| | Intercept | 3.961 |
| **$BF_1$** | max (0, 223 - LP) | –0.051 |
| **$BF_2$** | max (0, LP - 223) | 0.091 |
| **$BF_3$** | max (0, 1 - IN) | –25.414 |
| **$BF_4$** | max (0, IN - 1) | – 1.458 |
| **$BF_5$** | max (0, LP - 223) × LN | 0.016 |
| **$BF_6$** | max (0, LP - 223) × AGE | –0.002 |
| **$BF_7$** | max (0, 223 - LP) × max (0, 1 - IN) | 0.051 |
| **$BF_8$** | max (0, 4305 - TMY) × max (0, IN - 1) | –0.133 |

LP: lactation period (days); LN: number of lactations; AGE: cow age (months); TMY: total lactation milk yield (kg); IN: insemination number; BFi: piecewise linear basis functions. The values corresponding to the piecewise linear functions of the fundamental functions in the estimation equation created here are given as $BF_1$, $BF_2$,..,.$BF_8$, respectively.

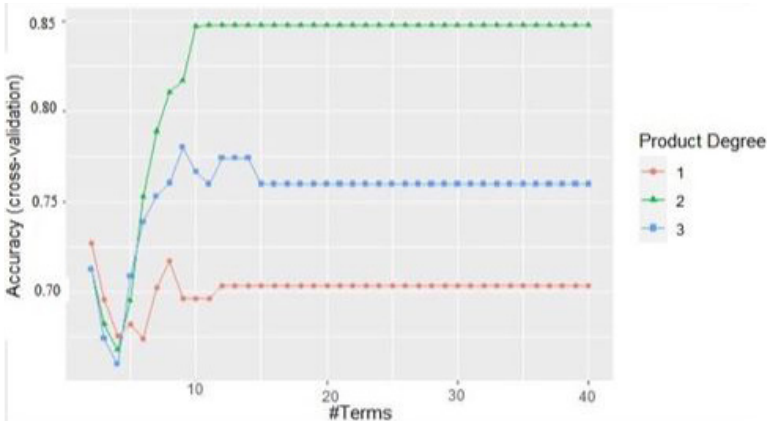


**Figure 3.** Classification performance graph of candidate MARS models. Red color: model with 1st degree interaction; green color: model with 2nd degree interaction; blue color: model with 3rd degree interaction. The product degree determines the highest classification accuracy.

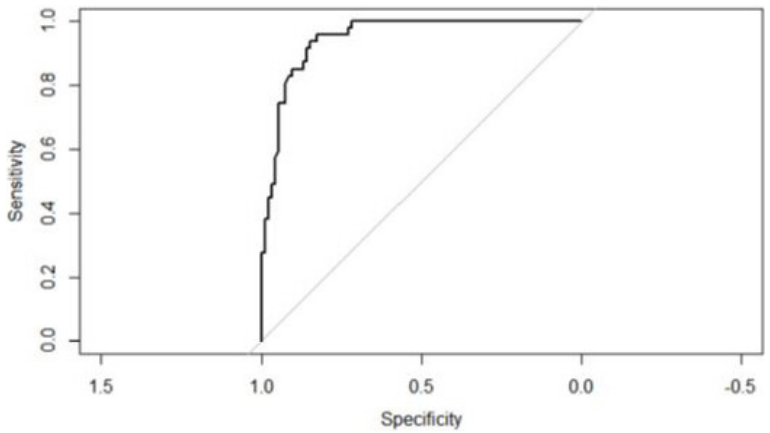


**Figure 4.** ROC curve for the MARS algorithm (pregnant).

of the test data set. Accordingly, definitions are made in the R Script window and the value of the area under the ROC curve (area under the curve (AUC) : 0.947) is checked [2, 25, 28].

As can be seen from Figure 4, the optimum sensitivity (at the point where the ROC value for specificity is 0.5) for all classifiers is 0.947 and the optimum specificity is close to 1. These values agree with the values in the studies by Grzesiak et al. [6]. To determine whether an animal is pregnant or not, a threshold value must be found. The probability value corresponding to the point where the sum of sensitivity and specificity is highest will form the threshold point. After the analysis with R, as can be seen from Table 4, the threshold point (A + B) was found to be 0.35 cows with a probability of being pregnant (positive) value of 0.35 or greater were classified as pregnant. Otherwise, animals with a probability of being pregnant less than 0.35 are classified as not pregnant (negative) (Table 4).

The classification success of the MARS algorithm is demonstrated with the confusion matrix (Table 5).

The pregnancy statuses of 45 animals out of 47 animals with positive pregnancy status in the training set were classified as positive. Accordingly, when the pregnancy status is taken as a positive reference, the correct classification rate (sensitivity) of the animal with positive pregnancy status was found to be 45 / 47 = 0.9574. Of the 96 pregnancy-negative animals in the training set, which were not actually pregnant, 77 were classified as pregnant-negative. The correct classification rate (specificity) of pregnant animals was found to be 77 / 92 = 0.8370. The overall classification ratio of the training set was Accuracy = (45 + 77) / 139 = 0.8777 (Table 5) [25]. According to the relative importance graph created for the MARS algorithm within the scope of classification problems, it is seen that the variables LP, IN, AGE, and LN are important in determining the estimation equation, respectively (Figure 5).

**Table 4.** ROC curve results for the MARS algorithm (pregnant).

| Threshold point (Positive if greater than or equal to) | Specificity (A) | Sensitivity (B) | **A+B** |
|---|---|---|---|
| 0.05 | 0.620 | 1.000 | 1.62 |
| 0.15 | 0.728 | 0.979 | 1.71 |
| 0.25 | 0.804 | 0.957 | 1.76 |
| **0.35** | **0.848** | **0.936** | **1.78** |
| 0.45 | 0.902 | 0.830 | 1.73 |
| 0.55 | 0.924 | 0.745 | 1.67 |
| 0.65 | 0.946 | 0.702 | 1.65 |
| 0.75 | 0.957 | 0.574 | 1.53 |
| 0.85 | 0.967 | 0.468 | 1.44 |
| 0.95 | 1.000 | 0.213 | 1.21 |

ROC: receiver operating characteristic; Threshold point: the probability value corresponding to the point where the sum of sensitivity and specificity is highest will form. Sensitivity: the correct classification rate of the animal with positive pregnancy status; Specificity: the correct classification rate of pregnant animals' status.

**Table 5.** Cross-validated (10-fold) confusion matrix.

| **Confusion matrix** | | |
|---|---|---|
| | Real value | |
| Estimated value | **1** | **0** |
| **1 (pregnant)** | 45 | 15 |
| **0 (Not pregnant)** | 2 | 77 |

When Figure 5 is examined, it is seen that LP, IN, AGE, and LN variables are important in estimating the pregnancy status determined according to the relative importance graph in determining the estimation equation with the MARS algorithm within the scope of classification problems, and this situation agrees with the literature [3, 5, 7]. In Table 6, it is seen that the MARS model created gives reliable results because the general classification ratio in both the training set and the test set is above the sensitivity and specificity ratios of 0.80.

## 4. Discussion

Similarly, a detailed calculation of the values in the test set was made by Grzesiak et al. [6]. Yağanoğlu [32] showed that in the ROC curve method, which was applied to find the important cutoff point, the variable with the highest sensitivity and selectivity value emerged and the ROC curve method had an important diagnostic power in revealing whether the sheep were pregnant or not.

In the present study, the AUC value (AUC = 0.947) was taken as the determining variable to determine the appropriate cut-off point. The AUC value also consists of sensitivity and selectivity values. Accordingly, the AUC value, which has high sensitivity and high selectivity, is also high. Thus, it shows parallelism in separating pregnant and nonpregnant animals in this study.

In this case, the value of 0.947 obtained in the present study is compatible with those obtained by Akın et al. (0.986) [25] and Akben (0.905) [29]. In addition, due to the high ratios in both training and test sets, no overfitting problem was encountered. If these ratios were too high in the training set and too low in the test set, an overfitting problem would be observed. In addition, as can be seen, the values in the training set and the test set were found to be quite close to each other. Grzesiak et al. [6] evaluated the dependent variable as two-level (binary) as good and bad progeny and reported the accuracy of estimations (accuracy classification) as 83%. This value was found very close to the data of the study (accuracy classification: 81%). In this respect, it is like our current study. Süt and Şimşek [7] compared different decision tree algorithms (CART, CHAID, exhaustive-CHAID) with each other in terms of classification performance. The performances of these algorithms were found to be 0.801 and 0.954, respectively, using the criteria of sensitivity and specificity. Again, in line with this study, the areas under the ROC curve of all algorithms were calculated ($p < 0.001$). It was determined that the algorithm with the smallest area under the ROC curve was CART (0.801), and the hit rate for this algorithm was 91.1%. Yılmaz [2] calculated the accuracy rate (%),

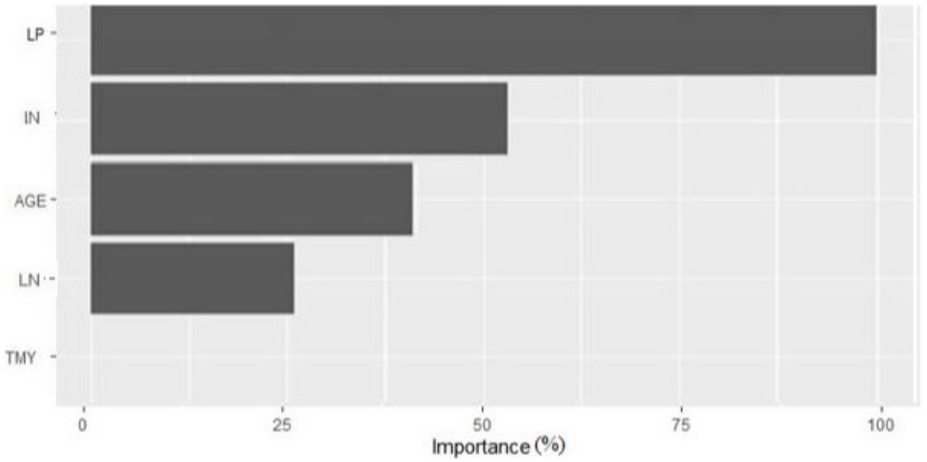


**Figure 5.** Relative importance graph of significant independent variables. LP: lactation period (days), AGE: cow age (months), LN: number of lactations, IN: insemination number, TMY: total lactation milk yield (kg).

**Table 6.** Classification success graph of MARS (training and test set) algorithm results.

| **Statistics** | **Training set mars result** | **Test set mars result** |
|---|---|---|
| Accuracy | 0.8777 | 0.8485 |
| 95% CI | (0.8114, 0.9271) | (0.681,0.9489) |
| No information rate | 0.6619 | 0.6667 |
| p-value [Acc > NIR] | 5.005e-09 | 0.0167 |
| Kappa | 0.7441 | 0.6809 |
| McNemar's test p-value | 0.0036 | 0.3711 |
| Sensitivity | **0.9574** | **0.9091** |
| Specificity | **0.8370** | **0.8182** |
| Pos. pred. value (PPV) | 0.7586 | 0.7143 |
| Neg. pred. value (NPV) | 0.9630 | 0.9474 |
| Prevalence | 0.3381 | 0.3333 |
| 'Positive' class | 1 | 1 |
| Detection rate | 0.3165 | 0.3030 |
| Balanced accuracy | 0.8920 | 0.8636 |
| Detection prevalence | 0.4173 | 0.4242 |

sensitivity, specificity, and the area under the ROC curve to find the best algorithm within the prediction performance. Like the present study, the areas under the ROC curve of all algorithms were calculated ($p < 0.001$). The studied data is divided into two main parts as training set (80%) and test set (20%) for MLP algorithms. Within the scope of the research, sex (female and male), farms (Mastung, Quetta, Noshki), maternal age, birth weight, lambing season, and year were used as independent variables. It was determined that sensitivity, specificity, and overall correct classification rate was over 90% in all of the CART, CHAID, Naive bayes, C5, and multilayer perceptron classification algorithms. In other words, it was seen that all algorithms used in the research showed superior performance. It has been observed that the algorithm with the smallest area under the ROC curve is CART (0.801). The singularity rate was found to be 98.9%. In this study, the rate of being pregnant was found to be 81.28%, while 0.947 was below the ROC curve. The correct classification rate we found is compatible with those obtained by Süt and Şimşek [7] and Yılmaz [2]. As a result of the literature review, no studies were encountered except for a few studies on the use of the classification status of the method in the field of livestock. However, by using the results obtained, it is thought that a preliminary idea can be obtained in the determination of the pregnancy status by considering the lactation day, insemination number, age, total milk yield, and lactation number.

In conclusion, in estimating the pregnancy status, analysis was made with the MARS algorithm, and it was seen that the variables LP, IN, AGE, and LN, respectively, determined according to the relative importance graph, were important. Sensitivity was 95.74% for the training set and 90.91% for the testing set; 83.70% for the specificity training set and 81.82% for the testing set; the overall correct classification rate was determined to be 87.77% for the training set and 84.85 for the test set. In addition, the area under the ROC curve was 0.947 and the pregnancy rate was 81.28%. It has been determined that there is no practical and detailed information on how to interpret the results of the analysis. In the future, within the scope of classification problems, these approaches can also be applied to predict multiple traits in animal species in determining the overall

classification rate, prediction equation in both the training and test sets. With DM algorithms, which are easy to use and interpret, the independent variables that highlight the desired or undesirable feature in animal husbandry and the definition of combinations of these variables will form the basis for future studies.

**Funding**
This research did not receive any specific grant from funding or financial support.

**Author contributions**
M. Boğa contributed to data acquisition and critical revision. D. Çanga contributed to methodology, study design, literature searches, and test set with the MARS algorithm.

**Conflict of interest**
We declare that there are no conflicts of interest between the authors in this article.

**References**

1. Küçükönder H, Üçkardeş F, Narinç D. A Data mining application in animal breeding: Determination of some factors in japanese quail eggs affecting fertility. Kafkas University, Faculty of Veterinary Medicine 2014; 20 (6): 903-908. https://doi: 10.9775/kvfd.2014.11353
2. Yılmaz M. Comparison of different data mining algorithms used in animal science. Msc, Iğdır University, Iğdır, Türkiye, 2017.
3. Küçükönder H, Boğa M, Burğut A, Üçkardeş F. Modelling of the lactation milk yield through artificial neural networks. Hayvansal Üretim 2015; 56 (2): 22-27.
4. Samarasinghe S. Neural Networks for Applied Sciences and Engineering. 1st Edition Boca Raton, New York: Auerbach Publications, 2006. https://doi: 10.1201/9780849333750
5. Grzesiak W, Zaborski D, Sablik P, Pilarczyk R. Detection of difficult conceptions in dairy cows using selected data mining methods. Animal Science Papers and Reports 2011; 29 (4): 293-302.
6. Grzesiak W, Zaborski D, Sablik P, Żukiewicz A, Dybus A, Szatkowska I. Detection of cows with insemination problems using selected classification models. Computers and Electronics in Agriculture 2010; 74 (2): 265-273. https://doi: 10.1016/j.compag.2010.09.001
7. Süt N, Şimsek O. Comparison of regression tree data mining methods for prediction of mortality in head injury. Expert Systems with Applications 2011; 38(12): 15534-15539. https://doi:10.1016/j.eswa.2011.06.006
8. Piwczynski D. Using classification using trees in statistical analysis of discrete sheep reproduction traits. Journal of Central European Agriculture 2009; 10 (3): 303-309.
9. Piwczynski D, Nogalski Z, Sitkowska B. Statistical modeling of calving ease and stillbirths in dairy cattle using the classification tree technique. Livestock Science 2013; 154 (1-3): 19-27.
10. Grzesiak W, Zaborski D. Examples of the use of data mining methods in animal breeding. In: Data Mining Applications in Engineering and Medicine.1st ed. Adem Karahoca, London: In Tech, Rijeka, Croatia; 2012. pp.303-324.
11. Eyduran E, Yakubu A, Duman H, Aliyev P, Tırınk C. Predictive modeling of multivariate adaptive regression splines: An R Tutorial. In: Çelik Ş (editor). Veri Madenciliği Yöntemleri: Tarim Alanında Uygulamaları 1th ed. Çanakkale, Türkiye; 2020; pp.25-48.
12. Friedman JH. Multivariate adaptive regression splines. In Annals of Statistics, 1999; 19 (1): 1-141.
13. Doğan İ. Investigation of the factors which are affecting the milk yield in Holstein by CHAID analysis. Veterinary Journal of Ankara University 2003; 50 (1): 065-070 (in Turkish with an abstract in English). https://doi: 10.1501/Vetfak_0000002231
14. Çelik Ş, Yılmaz O. Prediction of body weight of turkish tazi dogs using data mining techniques : Classification. Pakistan Journal of Zoology 2018; 50 (2): 575-583.
15. Everingham YL, Sexton J, and White J. *An introduction to multivariate adaptive regression splines for the cane industry.* In: Proceedings of the 2011 Conference of the Australian Society of Sugar Cane Technologists; Mackay, QLD, Australia; pp. 1-22.
16. Çanga D, Boğa M. Determination of the effect of some properties on egg yield with regression analysis method bagging Mars and R application.Turkish Journal of Agriculture - Food Science and Technology 2020; 8 (8): 1705-1712 (in Turkish with an abstract in English). https://doi:10.24925/turjaf.v8i8.1705-1712.3468
17. Çelik S, Yilmaz O. The relationship between the coat colors of Kars shepherd dog and its morphological characteristics using some data mining methods. International Journal of Livestock Research 2021;11(1):53–61.https://doi: 10.5455/ijlr.20200604.
18. Çelik Ş, Eyduran E, Şengül AY, Şengül T. Relationship among egg quality traits in japanese quails and prediction of egg weight and color using data mining algorithms. Tropical Animal Health and Production 2021; 53 (3): 382. https://doi: 10.1007/s11250-021-02811-2
19. Faraz A, Tirink C, Eyduran E, Waheed A, Tauqir NA et al. 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 2021; 53 (2). https://doi: 10.1007/s11250-021-02748-6

20. Akin M, Eyduran SP, Eyduran E, Reed BM. Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. Plant Cell, Tissue and Organ Culture 2020; 140 (3): 661-670. https://doi:10.1007/s11240-019-01763-8

21. Çanga D, Yavuz E, Efe E. Prediction of egg weight using MARS data mining algorithm through R. Journal of Agriculture and Nature 2021; 24 (1): 242-251. https:// doi: 10.18016/ksutarimdoga.vi.716880

22. Çanga D. Use of Mars data mining algorithm based on training and test sets in determining carcass weight of cattle in different Breeds. Journal of Agricultural Sciences 2021; 28 (2): 259-268. https://doi: 10.15832/ankutbd.818397

23. Emamgolizadeh S, Bateni SM, Shahsavani D, Ashrafi T, Ghorbani H. Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). Elsevier 2015; 529 (3): 1590-1600. https://doi: 10.1016/j.jhydrol.2015.08.025

24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2020.

25. Akın M, Eyduran SP, Eyduran E. R Yazılımı ile Tarım Bilimlerinde Regresyon ve Sınıflandırma Tipi Problemlerin Çözümünde Mars Algoritması. First Edition Publisher: Nobel Akademik Yayıncılık, 2020 (in Turkish).

26. Eyduran E, Akın M, Eyduran SP. Application of Multivariate Adaptive Regression Splines through R Software. Ankara: Nobel Academic Publishing, 2019.

27. Kibet K, Erick C. A Multivariate adaptive regression splines approach to predict the treatment outcomes of tuberculosis patients in Kenya. MSc, University of Nairobi, Kenya, 2012.

28. Zweig MH, Campbell G. Receiver-Operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry 1993; 39 (4): 561-577.

29. Akben SB. Determination of the blood, hormone and obesity value ranges that indicate the breast cancer, using data mining based expert system. Innovation and Research in BioMedical Engineering 2019; 40 (6): 355-360. https://doi: 10.1016/j.irbm.2019.05.007

30. Ayyıldız M, Tekin EM. Examination of some hemogram values by ROC curve in the diagnosis of parvo viral enteritis in dogs. Eurasian Journal of Veterinary Sciences 2021; 37 (2): 101-108. (in Turkish with an abstract in English). https://doi: 10.15312/EurasianJVetSci.2021.332

31. Revelle W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, 2020.

32. Yağanoğlu A. Comparison of pregnancy tests in sheep with Roc analysis. MSc, Atatürk University, Erzurum, 2010.