



FORECASTING SEASONAL MILK PRODUCTION USING MARS ALGORITHM FOR MULTIPLE CONTINUOUS RESPONSES IN HOLSTEIN DAIRY CATTLES

Demet ÇANGA BOĞA^{1*}, Mustafa BOĞA², Mutlu BULUT³

¹Osmaniye Korkut Ata University, Department of Chemistry and Chemical Processing, Bahçe Vocational School, 80050, Osmaniye, Türkiye

²Niğde Ömer Halisdemir University, Bor Vocational School, 51700, Niğde, Türkiye


³Çukurova University, Department of Agricultural Engineering, 01120, Adana, Türkiye


Abstract: In this study, seasonal milk yield estimation will be made using multivariate adaptive regression spline (MARS) algorithm for multiple continuous responses in dairy cattle (Holstein hybrid). For the research, milking records for the years 2020-2021 were collected from 157 dairy animals using Holstein hybrid dairy cattle from a research farm in Konya, Türkiye. The amount of feed given in this experiment was not changed and the effect of the season on the estimation of milk yield was investigated in the study. The analyzed independent variables used in the study were pregnancy status (PS), number of days milked (MDN), Lactation Number (LN), age of cows (months), average seven-day milk yield (7-Day Average Milk-SDMY), last lactation milk yield (last_MY), number of inseminations (IN), peak yield (Pik_Yield) and target variables were calculated as $(\text{Yield}_{\text{Autumn/winter/spring/summer}} \text{ (kg)}) = \text{Mean milk mean of season}$. In this context, the ehaGoF package was used to measure the prediction performance of the simultaneous MARS model established with the earth package for MARS analysis. MARS estimation equations obtained simultaneously for four dependent variables (multiple responses) are given. By looking at the MARS equation, the MARS model estimation equation was determined for the optimum milk yield, the threshold values, the three threshold values determined in the model were determined as MDN, Age, Peak_Yield, and the corresponding values were respectively; 159 days, 39.6 (months) and 37.1 kg/day. Considering the estimation equation, it is seen that the independent variables MDN, SDMY and LN are the most important variables in determining the estimation equation. It is seen that the best fitting value for the estimation equation of the dependent variables is the $\text{Yield}_{\text{Winter}}$ variable.


Keywords: MARS, Multiple response, Dairy cattle, Milk yield

*Corresponding author: Osmaniye Korkut Ata University, Department of Chemistry and Chemical Processing, Bahçe Vocational School, 80050, Osmaniye, Türkiye

E mail: demetcng@gmail.com (D. ÇANGA BOĞA)

Demet ÇANGA BOĞA  <https://orcid.org/0000-0003-3319-7084>

Mustafa BOĞA  <https://orcid.org/0000-0001-8277-9262>

Mutlu BULUT  <https://orcid.org/0000-0002-4673-3133>

Received: October 31, 2023

Accepted: January 30, 2024

Published: May 15, 2024

Cite as: Çanga Boğa D, Boğa M, Bulut M. 2024. Forecasting seasonal milk production using MARS algorithm for multiple continuous responses in Holstein dairy cattle. *BSJ Agri*, 7(3): 203-214.

1. Introduction

As a hybrid race adopted for the conditions of Türkiye, Holstein has a milk yield of 6000-9000 kg (Torshizi, 2016). Many different factors such as the physiological status of this breed (age, number of lactations, number of days in lactation) can affect milk yield (Boğa et al., 2020; Omar, 2022). However, important environmental factors that do not have a genetic effect on milk yield are lactation period, calving age, calving season and calving stage (Torshizi, 2016). In addition, the effects of environmental factors such as calving year, calving interval, calving season, number of births, herd and milking frequency on milk yield should be investigated (Javed et al., 2007; Eydurán et al., 2013).

Therefore, to study these relationships, machine learning methodologies from traditional statistical methods have been increasingly adopted. In light of this, researchers commonly use machine learning approaches to improve predictive efficiency (Nayana et al., 2022). These

algorithms develop by distinguishing and defining the consistency of educational knowledge patterns that may apply to complex nonlinear datasets between yield and parameters. In recent years, many machine learning approaches have been used in the literature on predictive modeling in agriculture and animal data (Küçükönder et al., 2014; Küçükönder et al., 2015). Investigated the effect of the number and duration of lactation of Holstein breed of cows on milk yield with the artificial neural network (ANN) method. In their study, it was determined that this model created with the artificial neural network converged well with the real values and that the performance in milk yield estimation could give more successful results by increasing the number of parameters. Similarly, in the study of Boğa et al. (2020), the effect of the number of lactations, lactation days, first calving and reproductive age, and the number of inseminations (ratio) on cattle milk yield (mean last seven days) was determined. They evaluated the data on the use of a deep neural network in dairy cattle farms



and suggested that additional controlled management was needed in livestock and that the errors of the farm should be corrected. Akin et al. (2020a), applied MARS algorithm in agricultural applications. Altay et al. (2022) found that it would benefit herd management by closing the gaps in mastitis diagnosis with the development of data mining methods. Therefore, they stated that the application of CART and MARS algorithms could be a good choice for cattle breeders to find the threshold values of effective milk characteristics that accurately distinguish healthy and unhealthy cows. Nayana et al. (2022) compared wheat yield estimation for India and the most wheat-producing countries using MARS after extracting the main characteristics with Principal Component Analysis (PCA), considering parameters such as cultivated area and production for 1962-2018. Çanga (2022) developed MARS prediction models with first-order interaction effects using the MARS algorithm to predict carcass yield. The carcass weight of cattle of various breeds was determined using a MARS Data Mining Algorithm based on training and test sets. Akin et al. (2020b) used MARS statistical approach to predicting macronutrient-related growth responses of three strawberry species. Çelik and Yilmaz (2021) investigated the effects of silage type, silage consumption, and birth type and birth weight on body weight after fattening in curly lambs using MARS and Bagging MARS algorithms. Tyasi et al. (2021) conducted the body weight estimation of the Hy-Line Silver Brown Commercial Layer chicken breed using MARS. In this research, estimation equations were created to examine seasonal milk yield estimation, performance was evaluated with error analyzes such as RMSE, ME, Rsq, and the most suitable MARS model was selected using cross-validation and user-defined parameter optimization. Therefore, this study aims to develop prediction models that best predict milk production using the MARS algorithm.

2. Material and Methods

2.1. Data Identification

The experimental data set was taken from a private farm in Türkiye with Holstein hybrid dairy cattle used with an automatic milking system in a private farm in Konya province. For a period of twelve months, milking records obtained from 157 milking animals for 2020-2021 were collected. It was performed with daily milking number and collective (by placing corn silage, grass silage, wheat straw, soybean meal) or individual (pellet feed distributed through automatic feeder) consumption. Each cow is milked twice a day by the automatic milking system. In this experiment, the amount of feed given was not changed and the effect of the season on milk production estimation was investigated in the study. The milking dairy cattle used in the experiment were Holstein, and group fattening was done. Only all milking animals were given 3 kg/day extra milking milk feed during milking. During each animal group feeding, wheat straw: 0.5 kg/day, alfalfa hay: 6.5 kg/day, corn silage: 15 kg/day, dairy feed: 5.5 kg/day, corn flake: 2 kg/day, cottonseed meal per animal: 1.5 kg/day, soybean meal: 1kg/day, barley paste: 2.5 kg/day, premix: 0.05 kg/day, calcid: 0.05 kg/day. On average, 58% roughage and 42% concentrate feed are mixed and given to the animals daily as a total mixed ration.

The independent variables used in the study (Table 1), pregnancy status (PS), number of days milked (MDN), LN Lactation Number (LN), age of cows (months), daily mean milk yield (7 Day Mean Milk-SDMY), last lactation milk yield (last_MY), the number of insemination (IN), peak yield (Pik_Yield), insemination number (IN); dependent variables were formed as $Yield_{Autumn/winter/spring/summer}$ (Kg) = Mean milk mean of season. The study was conducted by MARS to build and train the most suitable model for the 4 dependent variables.

Table 1. Descriptive statistics of variables to be used in modeling

	N	Minimum	Maximum	Mean	Std. dev
MDN	157	9	998	170.40	151.74
LN	157	1	6	2.38	1.34
Age(month)	157	26	111	50.55	19.04
SDMY (kg)	157	10	46	26.51	8.29
Pik_Yield (kg)	157	16	48	32.17	6.15
last_MY(kg)	157	3155.00	9450.00	6444.76	1223.99
IN	157	0	9	1.80	2.06
YieldAutumn (kg)	157	45	983	422.07	284.56
Yieldwinter (kg)	157	58	1116	606.61	240.50
Yield spring (kg)	157	510	1184	504.73	354.26
Yieldsummer (kg)	157	42	879	321.32	267.38
				Frequency	
PS		0		120	
		1		37	

MDN= number of days milked, LN= lactation Number, Age= cow's age (month), PS= pregnancy status (1:Ppregnant; 0:Nonpregnant), SDMY= daily mean milk yield (7 day mean milk), Pik_Yield= peak yield of daily mean milk yield, last_MY= last lactation milk yield, IN= insemination number.

2.2. Multivariate Adaptive Regression Spline (MARS)

MARS algorithm used by Friedman (1999) to capture nonlinear relationships between predictors and response variable(s) is a powerful approach that does not require assumptions about functional relationships between dependent and input variables. The model that emerges as the weighted total basic function including the BFi (x) function is given by Equation 1 below (Akin et al., 2020a; Eyduran et al., 2020; Çanga and Boğa 2020; Çelik et al., 2021; Çanga 2022).

$$y = \sum_{i=1}^k a_i BFi(x) \tag{1}$$

Mars algorithm is formed by the linear breakdown of the basic function of BFi (x) with the following Equation 2a and 2b.

$$BF_1 = \max(0, x - t) \begin{cases} x - t, & x > t \\ 0, & x \leq t \end{cases} \tag{2a}$$

$$BF_2 = \max(0, t - x) \begin{cases} t - x, & x > t \\ 0, & x \leq t \end{cases} \tag{2b}$$

here, x is the variable range; t is the node. The linear combination of the basic functions obtained accordingly as in Equation 3:

$$Y_i = a_0 + a_1 BF_1 + a_2 BF_2 + \dots + a_k BF_k \tag{3}$$

and the estimation equation is obtained. Here Y_i is dependent variable, a_0 intercept, and a_1, \dots, a_k are coefficients of the related basic functions (Emamgolizadeh et al., 2015; Everingham and Sexton 2011, Çanga and Boga 2019; Akin et al., 2020a).

2.2.1. A MARS model application

“earth” and “*chaGoF*” packages were used for the MARS model in the study (Eyduran et al., 2019; Eyduran et al., 2020). With the earth package, the same basic functions are generated for MARS prediction models created simultaneously for more than one dependent variable. So MARS models produced with the “earth” package have different coefficients. To make this estimate, the Generalized cross-validation (GCV) method, a computational solution for linear models that provide an estimated exclusion cross-validation error metric, is used. According to the GCV criterion, MARS generalizes the model by eliminating the terms. GCV is given by Equation 4, a form of regulation that balances model complexity with the goodness of fit (Eyduran et al., 2019; Akin et al., 2020a; Akin et al., 2020b).

$$GCV = \sum_{i=1}^N \frac{(y_i - \hat{y})^2}{\left(1 + \frac{C}{N}\right)^2} \tag{4}$$

Here, $C = 1 + cd$, is the number of items in the N dataset; d is a degree of freedom; c is the basic function addition penalty. Y_i is an independent variable and \hat{Y} is an estimated value (Eyduran, 2020; Akin et al., 2020a).

2.2.2. Parameter tuning

The maximum degree of interaction and the number of

terms in the final state are two important setting parameters for the MARS model. These two values are set by the “prune” and “degree” caret application, respectively. The maximum number of terms of the pruned model is “prune”. The actual degree of interaction is calculated by “degree”. The prune can be calculated automatically by the user or by using an external resampling technique and the default pruning protocol uses GCV. In addition, the earth package helps to evaluate possible interactions between the functions of various functions by reducing the number of nodes. To find the best hyperparameter combination, we use cross-validation (for k = 10 times) using the “caret” function. In the last case, the closest CVRSq to RSq is obtained using the FOR loop created in the R package program. In other words, with the FOR loop, the best CVRSq value, that is, the optimum CVRSq value was created out of 100 cycles. MARS algorithm has the advantage of using input variables that only increase the accuracy of the model and obtain an automated type of feature selection. This will be installed with the necessary parameters and run in each dataset using all features and the Spline model as a classifier. Each dataset gives the most appropriate feature subset rated based on its relative importance. Based on the highest overall accuracy, the smallest number of attributes collected, and the lowest false alarm error, the best optimal feature subset was selected (Eyduran et al., 2019; Akin et al., 2020a).

2.2.3. Model validity

The most common model fit criteria to be used in measuring the predictive accuracy of the MARS algorithm (Goodness of Fit Criteria) are the goodness of fit criteria such as R-square, RMSE and MAE mentioned below (Eyduran and Zaborski, 2017; Eyduran et al., 2019; Akin et al., 2020a; Çelik et al., 2021; Nayana et al., 2022). The model was evaluated according to these values.

1) Determination coefficient (R^2):

It is the percentage of the total variation in the response variable explained by the regression line. The Equation 5 is expressed by X.

$$R^2 = 1 - \frac{SSE}{SST} \tag{5}$$

where $SSE = (y_i - \hat{y})^2$ is the sum of the squares of the differences between the predicted and the observed value, and $SST = (y_i - \bar{y})^2$ is the sum of the squares of the differences between the observed and the overall average value?

2) Average square error (RMSE), average estimation error (is the square root of the average square error). The formula is stated as given in Equation 6:

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - \hat{y})^2} \tag{6}$$

3) Average error (ME) is the average estimation error. It is less sensitive to outliers. It is given by the formula as in Equation 7:

$$ME = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}) \quad (7)$$

4) Mean absolute deviation (MAD) is the mean absolute estimate error. It is less sensitive to outliers. The formula is given as in Equation 8:

$$MAD = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}| \quad (8)$$

5) Pearson correlation coefficient between actual values and estimated values in terms of a dependent variable (r) (Equation 9):

$$PC = r_{y_i \hat{y}} = \frac{Cov(y_i, \hat{y})}{S_{y_i} S_{\hat{y}}} \quad (9)$$

6) Akaike information criterion (AIC) (Equation 10a and 10b):

$$AIC = \ln \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \right] + 2k; \text{ If } \frac{n}{k} > 40 \quad (10a)$$

$$AIC_C = \ln \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \right] + 2k + \frac{2k(k+1)}{n-k-1}; \text{ otherwise;} \quad (10b)$$

Standard deviation ratio (SD_{ratio}):

S_m: Standard deviation of model error terms,

S_d: The standard deviation of the dependent variable,

Cov(y_i, ŷ): The covariance between actual and predicted values in terms of a dependent variable,

S_{y_i}: The standard deviation of the actual values of the dependent variable and

S_ŷ: It refers to the standard deviation of the mined values of the dependent variable.

3. Results and Discussion

The basic functions (BF_i) and coefficients of MARS prediction equations obtained simultaneously for four dependent variables (multiple responses) are presented in Table 2.

Table 2. Basic functions and corresponding coefficients of MARS model for the estimation of seasonal dependent variables

Terms	Coefficients (a _i)				Basis functions
	Yield _{Autumn}	Yield _{winter}	Yield _{Spring}	Yield _{summer}	
1	-284.15	1682.55	635.20	579.83	Intercept
2	16.16	-9.36	1.78	-4.02	max(0, MDN - 89)
3	5.74	-10.58	3.64	1.78	max(0, 159 - MDN)
4	-15.21	9.63	-0.11	11.18	max(0, MDN - 167)
5	-2.88	-1.23	7.72	-6.69	max(0, MDN - 276)
6	3.09	1.11	-7.90	-0.10	max(0, MDN - 360)
7	50.20	15.60	245.78	99.27	max(0, LN - 2)
8	-7.67	-3.11	-25.60	-19.12	max(0, Age - 39.6)
9	5.56	1.87	12.17	13.70	max(0, Age - 51.7)
10	-1.56	18.89	4.39	1.42	max(0, 35.8 - SDMY)
11	-14.53	-7.24	-13.52	-5.16	max(0, 37.11 - Pik_Yield)
12	15.02	9.43	22.59	25.94	max(0, Pik_Yield - 37.11)
13	-0.02	0.00	-0.10	-0.04	max(0, 7797 - last_MY)
14	-0.02	0.01	-0.05	-0.02	max(0, last_MY - 7797)
15	0.43	-0.04	4.03	1.57	max(0, Age - 39.6) * PS
16	0.16	0.24	-0.12	-0.05	max(0, 120 - MDN) × max(0, 35.8 - SDMY)
17	-0.05	-0.01	-0.10	-0.04	max(0, MDN - 120) × max(0, 35.8 - SDMY)
18	-0.15	0.03	-0.22	-0.20	max(0, 159 - MDN) × max(0, Pik_Yield - 32.71)
19	0.00	0.00	-1.3896e-0.5	0.00	max(0, 89 - MDN) × max(0, 7797 - last_MY)
20	0.00	-8.54	0.00	0.00	max(0, 159 - MDN) × max(0, last_MY - 7063)
21	0.00	0.00	0.00	0.00	max(0, 159 - MDN) × max(0, 7063 - last_MY)
22	2.21	-0.63	6.17	0.17	max(0, 3 - LN) × max(0, 35.8 - SDMY)
23	0.00	0.00	0.00	0.00	max(0, Age - 39.6) × max(0, 4580 - last_MY)

The model consists of 23 terms left between each node and 31 basic functions with four-way interaction. The three threshold values determined in the model for MDN, Age, Peak_Yield are 159day, 39.6 (month), and 37.11kg, respectively, and these values are the most common in the multi-response MARS model equation that constitutes the basic functions. These values are selected to simplify the multi-response MARS model by deleting the corresponding basis functions. As a result, this simplified MARS model is made up of marked coefficients and basic functions (Akin et al., 2020b).

For the MARS model estimation equation for optimum milk yield, the corresponding values for the three threshold values MDN, Age, Peak_Yield are 159d, 39.6 (month), and 37.1kg, respectively. Threshold values are the most common ones in the multi-response MARS model equation, which constitutes the basic functions, and the equations based on these values are deleted. Thus, it is selected to simplify the multi-response MARS model by deleting the relevant basic functions (Akin et al., 2020a; Akin et al., 2020b). As a result, the simplest form of the model is composed of coefficients and basic functions related to the MARS model simplified in this way. To further optimize the target responses and to ensure optimum milk yield for the tested dependent variables, the $Yield_{Autumn/winter/spring/summer}$ estimation equations was first defined by looking at the MARS equation in Table 3.

First of all, when the four prediction models are examined by looking at both Table 3 and the prediction equations, the sign differences regarding some coefficients of the same basic functions are remarkable (Akin et al., 2020b; Çelik and Yılmaz 2018; Çanga 2022; Çanga and Boğa, 2019). Since the threshold value for the lactation day, which is one of the variables discussed in Table 2, is LD=159, in the case of $\max(0, MDN - 89)$, there is an increase of 16.16 units and 1.78 units in $Yield_{Autumn}$ and $Yield_{Spring}$ milk yield, respectively; $Yield_{winter}$ and $Yield_{summer}$ milk yields decreased by 9.36 and 4.02 units, respectively. When the third term in the prediction model is examined, when $MDN > 159$, the effect of $Yield_{winter}$, $Yield_{Autumn}$, $Yield_{Spring}$, and $Yield_{Summer}$ on milk yield is masked. $MDN < 167$ the effect of the fourth, fifth and sixth term is masked. When $LN > 2$, a positive effect is seen in all dependent variables, while the biggest effect is $Yield_{Spring}$ with 245.78 units. When $AGE > 39.6$, the effect of all dependent variables is masked. If $MDN > 120$ and $SDMY < 35.8$, the interaction effect will be negative in all dependent variables, while if $MDN < 120$ or $SDMY \geq 35.8$, the interaction effect will be masked. Similar comments can be made about other situations (Emamgolizadeh,

2011; Çelik et al., 2021; Fatih et al., 2021; Faraz et al., 2021). Within the scope of the Response Surface Method, it can be suggested that there is the same directional relationship between the dependent variables discussed with the optimization logic in terms of the ease of interpretations to be made. Therefore, similar to this study, when the study on tissue culture conducted by Akin et al. (2020b) was examined, the results obtained from the optimal design response surface method for the optimization of three dependent variables with the same directional relationship between them were analyzed with MARS algorithm. While it is observed that the basic functions of MARS equations produced for the three dependent variables have the same marked coefficients, when the 3 estimation equations are examined, the sign differences regarding some coefficients of the same basic functions draw attention. This is because there is an inverse relationship between the two variables (due to the negative correlation).

From here, the estimation equation for the milk yield for the autumn is obtained as follows:

$$Yield_{Autumn} = 284.15 + 16.16 \times \max(0, MDN - 89) - 15.21 \times \max(0, MDN - 167) - 2.88 \times \max(0, MDN - 276) + 3.09 \times \max(0, MDN - 360) + 50.20 \times \max(0, LN - 2) + 5.56 \times \max(0, Age - 51.7) - 1.56 \times \max(0, 35.8 - SDMY) - 0.02 \times \max(0, 7797 - last_MMY) - 0.02 \times \max(0, last_MMY - 7797) + 0.16 \times \max(0, 120 - MDN) - 0.05 \times \max(0, MDN - 120) \times \max(0, 35.8 - SDMY) + 0.00 \times \max(0, 89 - MDN) \times \max(0, 7797 - last_MMY) + 2.21 \times \max(0, 3 - LN) \times \max(0, 35.8 - SDMY)$$

Finally, when we substituted the threshold values for maximizing responses, i.e. $MDN = 159$, $Age = 39.6$, $Peak_Yield = 37.11$, the corresponding basic functions were deleted according to the rules in equation 2. For example, according to equation 2a and equation 2b ($\max(0, MDN - 167)$), since $MDN = 159$, the basic functions of those terms are masked as equal to 0, and in this case, the model of the relevant dependent variable is deleted and simplified (Akin et al., 2020; Faraz et al., 2021).

Table 3. MARS model created after the elimination process

Terms	coefficients (a_i)				Basis functions
	$Yield_{Autumn}$	$Yield_{winter}$	$Yield_{Spring}$	$Yield_{summer}$	
1	-284.15	1682.55	635.20	579.83	Intercept
2	16.16	-9.36	1.78	-4.02	$\max(0, MDN - 89)$
4	-15.21	9.63	-0.11	11.18	$\max(0, MDN - 167)$
5	-2.88	-1.23	7.72	-6.69	$\max(0, MDN - 276)$
6	3.09	1.11	-7.90	-0.10	$\max(0, MDN - 360)$
7	50.20	15.60	245.78	99.27	$\max(0, MDN - 2)$
9	5.56	1.87	12.17	13.70	$\max(0, Age - 51.7)$
10	-1.56	18.89	4.39	1.42	$\max(0, 35.8 - SDMY)$
13	-0.02	0.00	-0.10	-0.04	$\max(0, 7797 - last_MY)$
14	-0.02	0.01	-0.05	-0.02	$\max(0, last_MY - 7797)$
16	0.16	0.24	-0.12	-0.05	$\max(0, 120 - MDN) \times \max(0, 35.8 - SDMY)$
17	-0.05	-0.01	-0.10	-0.04	$\max(0, MDN - 120) \times \max(0, 35.8 - SDMY)$
22	2.21	-0.63	6.17	0.17	$\max(0, 3 - MDN) \times \max(0, 35.8 - SDMY)$

The final models for Yield_{Autumn/winter/spring/summer} are obtained as given in Equation 11:

$$\begin{aligned}
 \text{Yield}_{\text{Autumn}} &= \begin{cases} 284.15 + 16.16 \times \max(0, \text{MDN} - 89) + 50.20 \times \max(0, \text{LN} - 2) - \\ 0.02 \times \max(0, 7797 - \text{last}_{\text{MY}}) - 0.02 \times \max(0, \text{last}_{\text{MY}} - 7797) - \\ 0.05 \times \max(0, \text{LD} - 120) \times \max(0, 35.8 - \text{SDMY}) + 2.21 \\ \max(0, 3 - \text{LN}) \times \max(0, 35.8 - \text{SDMY}) \end{cases} \\
 \text{Yield}_{\text{winter}} &= \begin{cases} 1682.55 + 9.36 \times \max(0, \text{MDN} - 89) - 15.60 \times \max(0, \text{LN} - 2) \\ + 18.89 \times \max(0, 35.8 - \text{SDMY}) - 0.01 \times \max(0, \text{last}_{\text{MY}} - 7797) \\ - 0.01 \times \max(0, \text{MDN} - 120) \times \max(0, 35.8 - \text{SDMY}) - 0.63 \\ \max(0, 3 - \text{LN}) \times \max(0, 35.8 - \text{SDMY}) \end{cases} \\
 \text{Yield}_{\text{spring}} &= \begin{cases} 635.20 + 1.78 \times \max(0, \text{MDN} - 89) + 245.78 \times \max(0, \text{LN} - 2) \\ + 4.39 \times \max(0, 35.8 - \text{SDMY}) - 0.10 \times \max(0, 7797 - \text{last}_{\text{MY}}) - \\ - 0.01 \times \max(0, \text{MDN} - 120) \times \max(0, 35.8 - \text{SDMY}) - 6.17 \\ \max(0, 3 - \text{LN}) \times \max(0, 35.8 - \text{SDMY}) \end{cases} \\
 \text{Yield}_{\text{summer}} &= \begin{cases} 579.83 - 4.02 \times \max(0, \text{MDN} - 89) + 99.27 \times \max(0, \text{LN} - 2) \\ + 1.42 \times \max(0, 35.8 - \text{SDMY}) - 0.04 \times \max(0, 7797 - \text{last}_{\text{MY}}) \\ - 0.02 \times \max(0, \text{last}_{\text{MY}} - 7797) - 0.05 \times \max(0, 120 - \text{MDN}) \\ \times \max(0, 35.8 - \text{SDMY}) - 0.04 \times \max(0, \text{MDN} - 120) \times \max(0, 35.8 \\ - \text{SDMY}) + 0.17 \times \max(0, 3 - \text{LN}) \times \max(0, 35.8 - \text{SDMY}) \end{cases}
 \end{aligned} \tag{11}$$

MARS optimizes all stages of model design and implementation, including variable selection, transforming predictive variables with a nonlinear relationship, determining interactions of predictive variables, and creating new nested variable strategies to deal with missing values and avoid overfitting with comprehensive self-tests (Akin et al., 2020a). After MARS analysis, the overall GRsq, CVRSq, and RSq values for all dependent variables were found to be quite high and very close to each other for optimum simultaneous MARS modeling (Table 3). When all dependent variables are examined, it is seen that the highest value among GCV, GRsq, Rsq, sd and CVRSq values belongs to Yield_{Winter} dependent variable. Some metrics for MARS prediction models for four dependent variables are summarized in Table 4.

Table 4. Summary performance of MARS prediction models of dependent variables

	GCV	GRsq	CVRSq
Yield _{Autumn}	0.0213	0.906	0.838
Yield _{Winter}	0.0055	0.944	0.954
Yield _{Spring}	0.0225	0.853	0.625
Yield _{Summer}	0.0171	0.878	0.644
All	0.0664	0.887	0.765

When the earth package has more than one (k) continuous dependent variable (multiple responses), k simultaneous prediction models are created. This package tries to minimize the sum of GCV values of k-dependent variables ($GCV_1 + GCV_2 + GCV_3 + \dots + GCV_k$) (Milborrow, 2019). $GCV(\text{Yield}_{\text{Spring}}) + GCV(\text{Yield}_{\text{Summer}})$ is based on the principle of being minimum. When the following outputs are examined, GCV total value is equal to $GCV_{\text{ALL}} = 0.0213 + 0.0055 + 0.0225 + 0.0171 = 0.0664$. Therefore, GRsq, Rsq, CVRSq values were found to be quite high and close to each other. Therefore, it can be said that the generalization ability of simultaneous MARS

modeling is very good, that is, there is no excessive adaptation problem (Akin et al., 2020a).

Since the R package in question tries to optimize all models simultaneously, the results of MARS analysis to be obtained for more than one dependent variable (the GRsq value calculated simultaneously for all dependent variables) will not be as good as MARS analyses to be obtained separately for each dependent variable (i.e. the GRsq values calculated for each of the dependent variables) (Milborrow, 2019).

The ehaGoF package was used to evaluate the predictive performance of the MARS model established for all dependent variables, and the results obtained are shown in Table 5 below.

It can be argued that MARS models established due to the very low values of RMSE, RRMSE, CV, RAE, MAD, MAPE and MRAE goodness of fit criteria have a very good fit. When the literature is examined, the fact that the standard deviation rate of a model is lower than 0.10 means that the predictive accuracy of that model is quite good (Grzesiak and Zaborski, 2012; Eyduran and Zaborski, 2017; Eyduran et al., 2019; Faraz et al., 2021). It has been reported that the standard deviation rate should be lower than 0.40 to say that an established regression model can have a good fit (Grzesiak and Zaborski, 2012). In this study, Yield_{Winter} MARS result with the best (0.17) value of Standard deviation rate (SDR) was observed. In other words, it can be said that the MARS model established for this value has a much better fit than other values. As here, it is seen that the mean of ME, that is, error terms, is theoretically zero, and this is the desired value.

The fact that the determination coefficient (Coefficient of Determination, Rsq) and the Adjusted Coefficient of Determination (AdjRsqu = 0.998) of the established MARS regression model are close to 1 means that the said model explains almost all of the total difference (variation) of the dependent variable. In the study, when

Table 5. The goodness of fit criteria MARS algorithms

Criteria	MARS Results			
	Yield _{Autum}	Yield _{Winter}	Yield _{Spring}	Yield _{summer}
1 Rootmeansquareerror (RMSE)	0.35	0.04	0.43	0.07
2 Relative root mean square error (RRMSE)	14.81	6.73	19.27	20.70
3 Standard deviation ratio (SDR)	0.22	0.17	0.28	0.25
4 Coefficient of variation(CV)	14.85	6.75	19.33	20.76
5 Pearson's correlation coefficients (PC)	0.98	0.99	0.96	0.97
6 Performance index (PI)	7.50	3.39	9.82	10.51
7 Mean error (ME)	0.00	0.00	0.00	0.00
8 Relative approximation error (RAE)	0.02	0.00	0.03	0.03
9 Mean relative approximation error (MRAE)	0.01	0.01	0.01	0.01
10 Mean absolute percentage error (MAPE)	0.21	0.61	0.31	0.32
11 Mean absolute deviation (MAD)	43.89	29.02	72.03	49.60
12 Coefficient of determination (Rsqr)	0.0-95	0.97	0.92	0.94
13 The adjusted coefficient of determination (ARsqr)	0.94	0.97	0.91	0.93
14 Akaike's information criterion (AIC)	1344.42	1210.78	1483.22	1363.93
15 Corrected Akaike's information criterion (CAIC)	1352.72	1219.08	1491.52	1372.73

Model Selection (all responses)

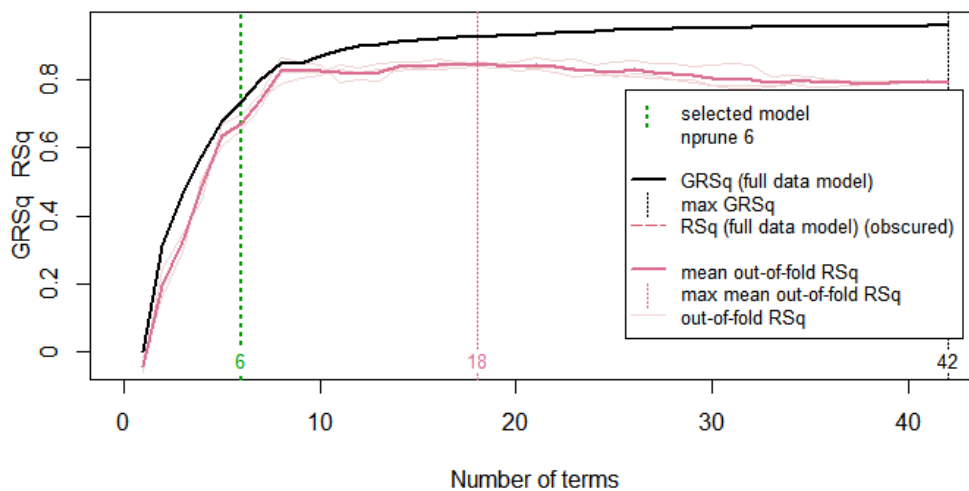


Figure 1. Selection graph of MARS model for all dependent variables.

Table 5 is examined, it is seen that the Rsq and AdjRs values of all dependent variables are very close to 1 and the highest value belongs to the dependent variable Yield_{Winter} (Rsq=0.971, AdjRsqr=0.966). Similar comments It is desirable that the adjusted coefficient of determination (Adjusted Coefficient of Determination, AdjRsqr =0.998) is close to the coefficient of determination (Rsq =0.999) (Akin et al., 2020a; Çelik et al., 2021).

The Pearson correlation coefficient (PC) between the observed and estimated values in terms of milk yield is desired to be very close to 1 for all dependent variables. When Table 5 was examined, it was determined that the Pearson correlation between the actual and predicted values was quite high and statistically significant in terms of the dependent variables examined. It is seen that the Yield_{Winter} value has the highest correlation (r=0.9869) (Çelik and Yılmaz, 2018; Akin et al., 2020a).

3.1. Graphical Representation of MARS Model

In the MARS model with four simultaneous continuous dependent variables, the black horizontal bold line represents the common GRSq value, the red horizontal dashed line represents the common Rsq value, the pale pink fine lines represent the corresponding Rsq value in each folk, and the horizontal thick pink line represents the common CVRsqr value, the dashed red vertical line represents the optimum number of terms corresponding to the common CVRsqr value in the peak, and the dashed fine dotted black vertical line (common for the four dependent variables) represents the point where the GRSq value is the maximum (Akin et al., 2020a). However, it can be stated that the dashed vertical dotted red line showing the point where the horizontal thick red line is maximum indicates the ideal number of terms determined by cross-validation. The earth pack accepts as the appropriate number of terms the number of terms

corresponding to the dashed vertical dotted black line, indicating the point where the GRSq value is maximum. In general, dashed vertical dotted red and black lines overlap. However, when Figure 1 is examined, it is seen that the MARS prediction model is established based on the GRSqvalue, that is, the GCV criterion has 6 terms for the MARS prediction model established based on cross validity for four dependent variables (Akin et al., 2020a). The Yield_{Spring} dependent variable is given as a graph showing the error value corresponding to each prediction value (Figure 2). In linear models, the constant variance assumption of errors is ideal. Therefore, it is desirable that the error values should spread evenly against the increase of the relevant estimation values and that the red straight horizontal line should be on the horizontal gray zero line. However, the fact that the

distribution of errors against the increasing Yield_{Spring} estimation values is in the form of a pipe around the zero point is proof of this assumption. However, it has been stated that the constant variance assumption, which is one of the most important assumptions of the linear model, is not important for MARS models (Milborrow, 2019; Akin et al., 2020a). Therefore, in the graph in Figure 2, it can be said that the constant variance is provided because the red horizontal straight line is almost above the zero horizontal line. This was confirmed to be in line with the results of the study by Akin et al (2020b). It is seen that observations 43, 67, and 135 for the Yield_{Spring} feature are outlier values that increase the error variance. Similar interpretations are made for other dependent variables.

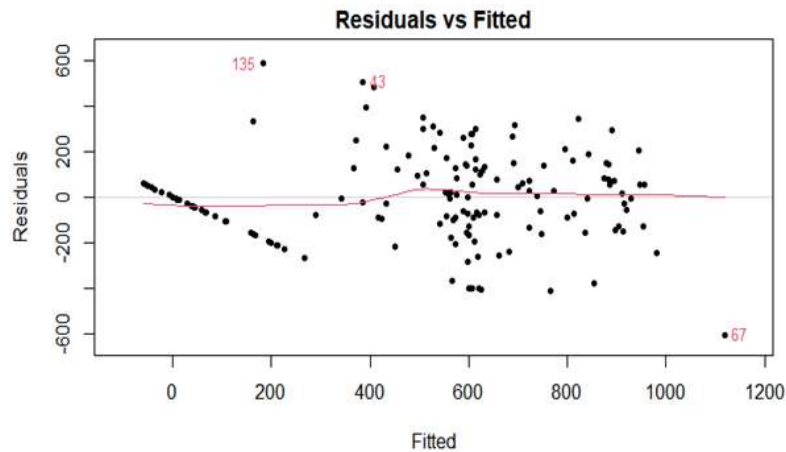


Figure 2. Distribution graph of errors according to the estimated Yield_{Spring} dependent variable.

Milborrow (2019) reported that the normal distribution of errors is generally not important for MARS models. The fact that the error values are almost above and around the cross line means that the errors show normal distribution. As can be seen from the graph in Figure 3, it can be said that observation values 43, 67 and 135 produce large error values (Figure 3). Similar to this research, Çanga (2022) compared the goodness-of-fit criteria of the training and test set model, reducing the bias by cross-validation. It represents the test data against the estimated graph obtained using the MARS model (Figure 4, 5, 6 and 7).

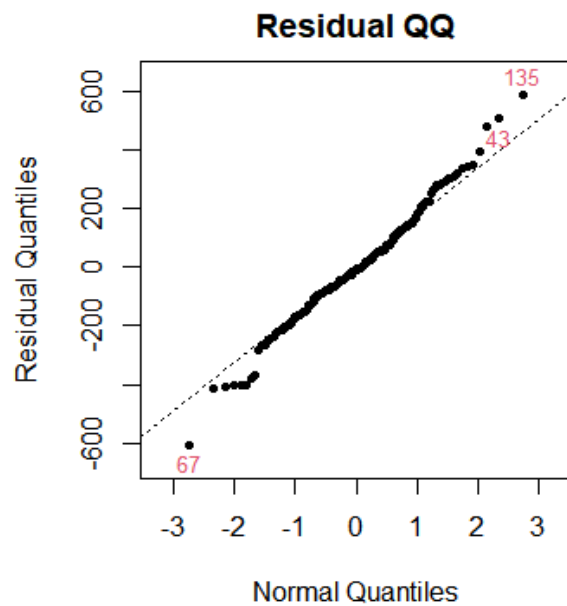


Figure 3. QQ graph of errors for Yield_{Spring} dependent variable.

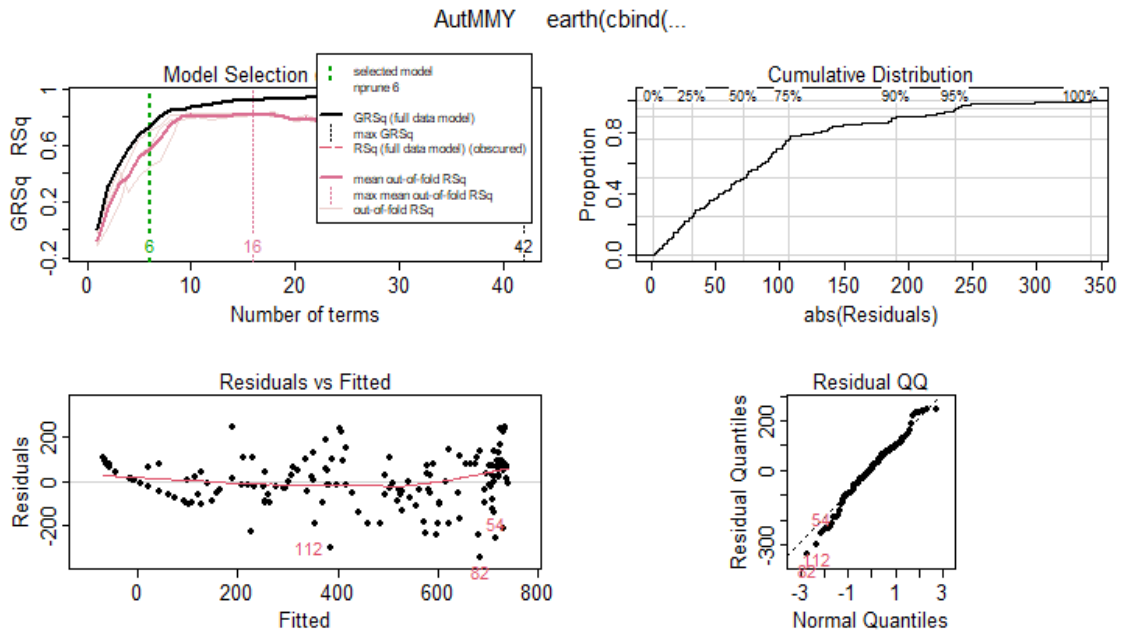


Figure 4. A graphical representation of the terms that make up the MARS model for Yield_{Autumn} estimation.

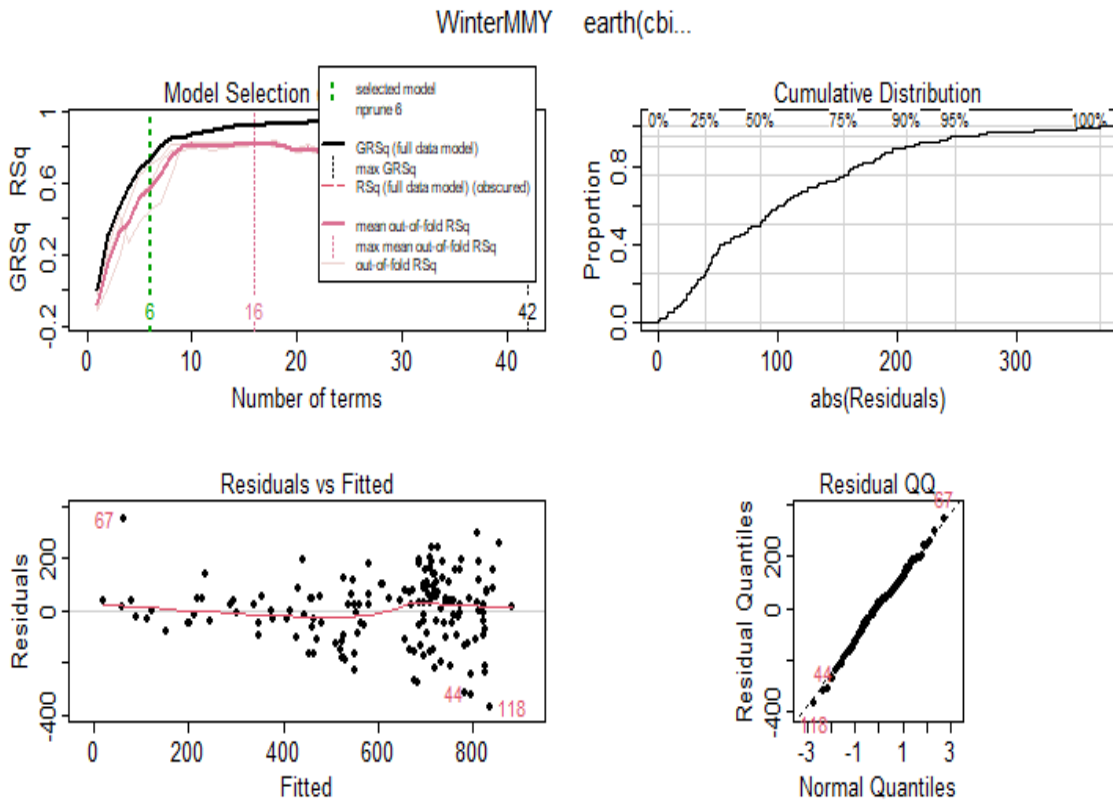


Figure 5. A graphical representation of the terms that make up the MARS model for Yield_{Winter} estimation.

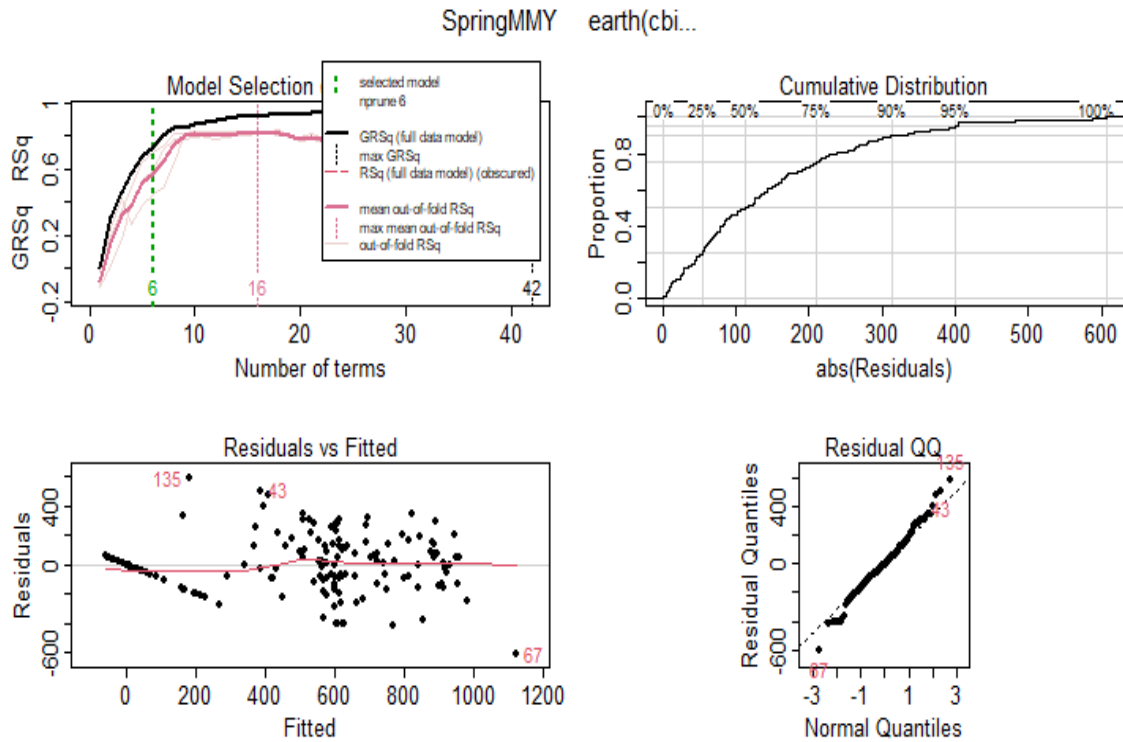


Figure 6. A graphical representation of the terms that make up the MARS model for Yield_{Spring} estimation.

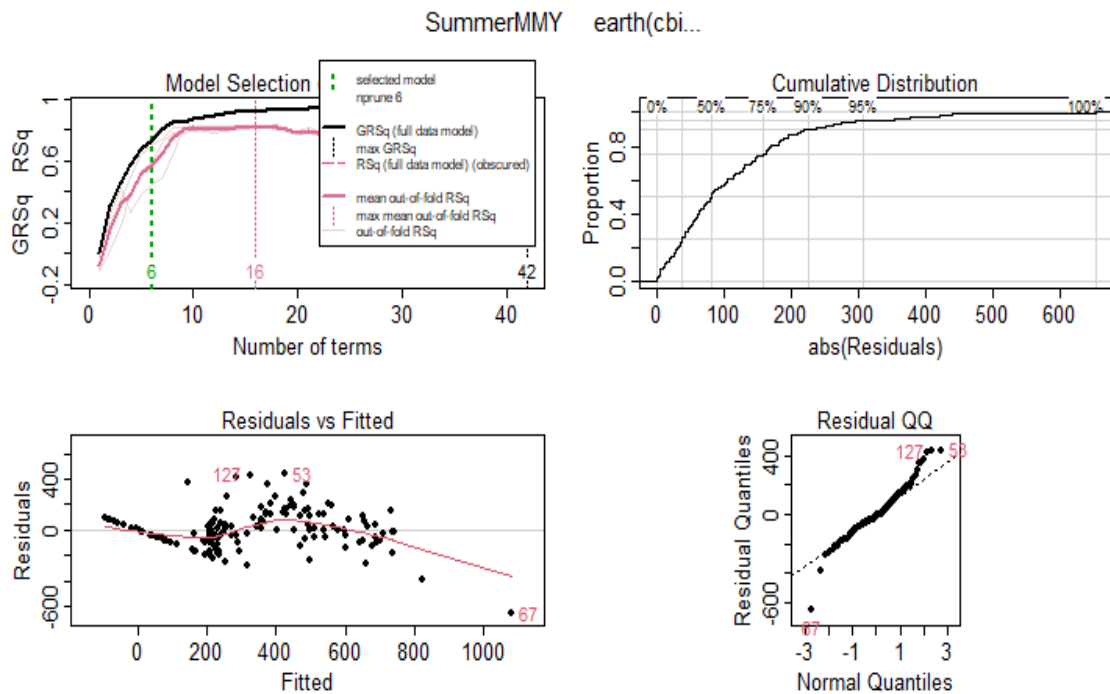


Figure 7. A graphical representation of the terms that make up the MARS model for Yield_{Autumn} estimation.

A more detailed interpretation can be made by giving a complete graphical holistic representation. In their study, Bayril and Yilmaz (2017) reported that the milk yields of animals differ in spring, summer, autumn, and winter and that milk yields are decreasing in summer and autumn. This situation is consistent with the results of the study. In parallel with this study, Novak et al. (2009) investigated the effects of hot months, lactation period,

and many lactations on milk yield. According to their findings, milk yields gradually increased from May to June and decreased in December. On the other hand, Bouallegue et al. (2013) reported that the milk yields of cattle were less in the summer months. Consistent with this study Vijayakumar et al. (2017) found that there was a relationship between the number of lactations, the number of days of lactation, the lactation phase-time, and

milk yield. Here, it is clear that the graphs that best fit among the four dependent variables are the Yield_{Winter} variable. In addition, when Table 5 was examined, it was observed that the Yield_{Winter} value had the highest correlation ($r=0.9869$) in terms of the dependent variables examined.

5. Conclusion

MARS algorithm is a method to advance a more understandable and easy prediction of agricultural and livestock. In the study, the direct effect of the parameters and complex interactions of the MARS model was clearly shown. It is seen that the best fit value for the estimation equation of four dependent variables is the Yield_{Winter} variable. The effect of the season on milk yield can be further investigated in this study by taking into account different parameters. Models with more and fewer features can be examined to determine the best-performing model. However, the results clearly show that machine learning models such as MARS perform more detailed and better than traditional statistical models when used in animal milk yield estimation.

It shows that the MARS model is suitable for Türkiye as a whole and for farmers producing milk to have information about the best yield periodically on the farm. Milk production estimation based on accepted statistics should be used to create short- and long-term plans to deal with future milk production, especially.

Author Contributions

The percentage of the author(s) contributions is presented below. All authors reviewed and approved the final version of the manuscript.

	D.Ç.B.	M.Bo.	M.Bu.
C	90	10	
D	100		
S	100		
DCP		100	
DAI	90	10	
L	70	20	10
W	80	10	10
CR	70	30	
SR	75	5	20
PM	80	10	20

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans. Livestock enterprises' own data was taken and there was no experiment conducted at the farm. Also, a signed consent form was also obtained from the owners

of the farms.

Acknowledgments

We thank the farm owners for their assistance in collecting this data.

References

- Akin M, Eydurun SP, Eydurun E. 2020a. R Yazılımı ile Mars (Multivariate Adaptive Regression Splines) Algoritması. Nobel Academic Publishing, Ankara, Türkiye, pp: 264.
- Akin M, Eydurun SP, Eydurun E, Reed BM. 2020b. Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. *J Plant Biotechnol*, 140(3): 661-670. <https://doi.org/10.1007/S11240-019-01763-8>.
- Altay Y, Aytakin İ, Eydurun E. 2022. Use of Multivariate Adaptive Regression Splines, Classification tree and Roc curve in diagnosis of subclinical mastitis in dairy cattle. *J Hellenic Vet Med Soc*, 73(1): 3817-3826. <https://doi.org/10.12681/jhvms.25864>.
- Bayril T, Yılmaz O. 2017. Holştayn sütçü ineklerde süt verim performanslarına buzağı cinsiyeti, servis periyodu, doğum sayısı ve buzağılama mevsiminin etkisi. *Dicle Üniv Vet Fak Derg*, 10(2): 89-94.
- Boğa M, Çevik KK, Burgut A. 2020. Classifying milk yield using deep neural network. *Pakistan J Zool*, 52(4): 1319-1325. <https://doi.org/10.17582/journal.pjz/20190527090506>.
- Bouallegue M, Haddad B, Aschi M, Ben H. 2013. Effect of environmental factors on lactation curves of milk production traits in Holstein – Friesian cows reared under North African condition. *Livestock Res Rural Devel*, 25(5): 37.
- Çanga D, Boğa M. 2019. Use of MARS in livestock and an application. III. International Scientific and Vocational Studies Congress, December 20, Nevşehir, Türkiye, pp: 31-37.
- Çanga D, Boğa M. 2020. Determination of the effect of some properties on egg yield with regression analysis met-hod bagging Mars and R application. *Turkish J Agri Food Sci Technol*, 8(8): 1705-1712. <https://doi.org/10.24925/turjaf.v8i8.1705-1712.3468>.
- Çanga D. 2022. Use of Mars data mining algorithm based on training and test sets in determining carcass weight of cattle in different breeds. *J Agr Sci*, 28(2): 259-268. <https://doi.org/10.15832/ankutbd.818397>.
- Çelik Ş, Eydurun E, Şengül AY, Şengül T. 2021. Relationship among egg quality traits in Japanese quails and prediction of egg weight and color using data mining algorithms. *Trop Anim Health Prod*, 53(3): 382. <https://doi.org/10.1007/s11250-021-02811-2>.
- Çelik Ş, Yılmaz O. 2018. Prediction of body weight of Turkish tazi dogs using data mining techniques: Classification. *Pakistan J Zool*, 50(2): 575-583. <https://doi.org/10.17582/journal.pjz/2018.50.2.575.583>.
- Çelik Ş, Yılmaz O. 2021. The relationship between the coat colors of kars shepherd dog and its morphological characteristics using some data mining methods. *IJLR*, 11(1): 53-61. <https://doi.org/10.5455/ijlr.20200604>.
- Emangolizadeh S, Bateni SM, Shahsavani D, Ashrafi T GH. 2015. Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). *J Hydrol*, 529(3): 1590-1600
- Everingham YL, Sexton J. 2011. An introduction to Multivariate Adaptive Regression Splines for the cane industry. Proceedings of the 2011 Conference of the Australian Society of Sugar Cane Technologists, May 4-6, Palm Cove, Australia.
- Eyduran E, Akin M, Eydurun SP. 2019. Application of

- Multivariate Adaptive Regression Splines through R Software. Nobel Bilimsel Eserler, Ankara, Türkiye, pp: 112.
- Eyduran E, Yakubu A, Duman H, Aliyev P, Tırink C. 2020. Predictive modeling of multivariate adaptive regression splines: An R Tutorial. In: Veri Madenciliği Yöntemleri: Tarım Alanında Uygulamaları. Rating Academy Ar-Ge Yazılım Yayıncılık Eğitim Danışmanlık ve Organizasyon Ticaret Limited Şirketi, Ankara, Türkiye, pp: 25-48.
- Eyduran E, Yılmaz I, Tariq MM, Kaygisiz A. 2013. Estimation of 305-d milk yield using regression tree method in brown Swiss cattle. *J Anim Plant Sci*, 23(3): 731-735.
- Eyduran E, Zaborski D. 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 J Zool*, 49(1): 257-265. <https://doi.org/10.17582/journal.pjz/2017.49.1.257.265>.
- Eyduran E. 2020. Package 'EhaGoF'. URL: <https://cran.r-project.org/package=EhaGoF> (accessed date: June 12, 2023).
- Faraz A, Tırink C, Eyduran E, Waheed A, Tauqir NA, Nabeel MS, Tariq MM. 2021. Prediction of live body weight based on body measurements in Thalli sheep under tropical conditions of Pakistan using cart and mars. *Trop Anim Health Prod*, 53(2): 301. <https://doi.org/10.1007/s11250-021-02748-6>.
- Fatih A, Çelik S, Eyduran E, Tırink C, Masood MT, Sheikh IS, Faraz A, Abdul Waheed A. 2021. Use of MARS algorithm for predicting mature weight of different camel (*Camelus dromedarius*) breeds reared in Pakistan and morphological characterization via cluster analysis. *Trop Anim Health Prod*, 53(1): 191. <https://doi.org/10.1007/s11250-021-02633-2>.
- Friedman JH. 1999. Multivariate adaptive regression splines. *Annals Stat*, 19: 67.
- Grzesiak W, Zaborski D. 2012. Examples of the use of data mining methods in animal breeding. In: *Data Mining Applications in Engineering and Medicine*. IntechOpen, pp: 304-321. <https://doi.org/10.5772/50893>.
- Javed K, Babar ME, Abdullah M. 2007. Within-herd phenotypic and genetic trend lines for milk yield in Holstein-Friesian dairy cows. *J Cell Anim Biol*, (4): 66-70.
- Küçükönder H, Boğa M, Burğut A ÜF. 2015. Yapay sinir ağları ile laktasyon süt veriminin modellenmesi. *Hayvansal Üretim*, 56(2): 22-27.
- Küçükönder H, Üçkardeş F, Nariç D. 2014. A data mining application in animal breeding: Determination of some factors in Japanese quail eggs affecting fertility. *Kafkas Univ Vet Fak Derg*, 20(6): 903-908. <https://doi.org/10.9775/kvfd.2014.11353>.
- Milborrow S. 2019. Earth: Multivariate Adaptive Regression Splines (MARS). *Annals Stat*, 19(1): 1-67. <https://doi.org/10.1214/aos/1176347963>.
- Nayana BM, Kumar KR, Chesneau C. 2022. Wheat Yield Prediction in India Using Principal Component Analysis-Multivariate Adaptive Regression Splines (PCA-MARS). *Agri Eng*, 4(2): 461-474. <https://doi.org/10.3390/AGRIENGINEERING4020030>.
- Novak P, Vokralova J, Broucek J. 2009. Effects of the stage and number of lactation on milk yield of dairy cows kept in open barn during high temperatures in summer months. *Archiv Tierzucht*, 2: 574-586.
- Omar MY. 2022. Comparison of reproductive performance between holstein and simmental cows in-terms of milk production, milk yield persistence, first lactation peak point and it is duration. MSc Thesis, Bursa Uludağ University, Institute of Health Science, Veterinary Sciences, Bursa, Türkiye, pp: 61.
- Torshizi ME. 2016. Effects of season and age at first calving on genetic and phenotypic characteristics of lactation curve parameters in Holstein cows. *J Anim Sci Technol*, 58: 8. <https://doi.org/10.1186/s40781-016-0089-1>.
- Tyasi TL, Eyduran E, Çelik S. 2021. Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa. *Trop Anim Health Prod*, 53(1): 7. <https://doi.org/10.1007/s11250-020-02443-y>.
- Vijayakumar M, Park JH, Ki KS, Lim DH, Kim SB, Park SM, Jeong HY, Park BY, Kim TI. 2017. The effect of lactation number, stage, length, and milking frequency on milk yield in Korean Holstein dairy cows using automatic milking system. *Asian-Australas J Anim Sci*, 30(8): 1093-1098. <https://doi.org/10.5713/AJAS.16.0882>.