KSÜ Tarım ve Doğa Derg 24 (1): 242-251, 2021
KSU J. Agric Nat 24 (1): 242-251, 2021
https://doi.org/10.18016/ksutarimdoga.vi.716880

Prediction of Egg Weight Using MARS data mining Algorithm through R

Demet ÇANGA[1], Esra YAVUZ[2], Ercan EFE[3]

[1]Osmaniye Korkut Ata University, Department of Chemistry and Chemical processing, Bahçe, Osmaniye, Turkey, [2]Kahramanmaraş Sütçü Imam University, Institute of Science and Technology, Department of Animal Science, Kahramanmaraş, Turkey, [3]Kahramanmaraş Sütçü Imam University, Faculty of Agriculture, Department of Animal Science, Kahramanmaraş, Turkey

[1]https://orcid.org0000-0003-3319-7084, [2]https://orcid.org 0000-0002-5589-297X, [3]https://orcid.org/ 0000-0002-5131-323X

✉: demetcanga@osmaniye.edu.tr

ABSTRACT

Internal and external quality characters of poultry eggs are quitely important to determine egg weight. Also, the quality of eggs is important for both hatching and egg production. The purpose of this study was modelling of egg weight with the MARS (Multivariate Adaptive Regression Splines) method using inner and outsider quality characters of egg in Lohmann LSL Classic white hybrit flock. For this purpose, the eggs of the Lohmann LSL Classic white hybrid flock (n=60) were used. Weekly egg yields were evaluated from the 22nd week to the 62nd week. In the research, for the prediction of dependent and continuous variable egg weight; shape index (SI), shell breaking resistance (SBS), shell weight (SW), shell thickness (ST), yolk diameter (YD), yolk width (YW), yolk height (YH), color (YC), albumen length (AW), albumen height (AL) and albumen height (AH) were used. In order to obtain perfect goodness of fit, in the “earth” package of the R program, the definitions of penalty -1, degree = 2, nprune = 10 and nk = 60. The research, the mars prediction model was determined such as EW = 63.1-0.906 * max (0,75-SI)-0.32 * max (0, SI-75) -62.4 * max (0,0.57-ST) -354 * max (0, ST-0.57) + 1.13 * Groupa2 * max (0, 75-SI) + 1.49 * (0.0.57-ST) max * YD + 8.2*max(0, ST 0.57) * YD-0.02*(0 YD-38.5)max* YC-0.0366*YH * max(0,13-YC). As a result, some quality variables were found to be important in determining egg weight. Variables such as group a2, SI, YC, ST, YD, YH to estimate the weight of the egg determined as the dependent variable were used. Other variables are not included in this equation. In the poultry, the MARS prediction model may be a better alternative to classical nonlinear models in predicting egg weight since that it is easier and has higher accuracy.

Research Article

Article History
Geliş Tarihi : 09.04.2020
Kabul Tarihi : 30.06.2020

Keywords
MARS,
Egg weight,
Lohmann bread,
Goodness of fit,
Non-parametric regression model

R kullanarak Mars Veri Madenciliği Algoritması ile Yumurta Ağırlığı Tahmini

ÖZET

Kanatlı hayvanlarda, yumurta ağırlığını belirlemede yumurtanın iç ve dış kalite özellikleri oldukça önemlidir. Yumurtanın kalite özellikleri, gerek kuluçka üretimi ve gerekse yemeklik yumurta üretimi açısından büyük önem taşımaktadır. Bu çalışmanın amacı, Lohmann LSL Classic beyaz hibrit sürü yumurtaları kullanılarak yumurtanın ic dış kalite özellikleri ile yumurta ağırlığının tahminini MARS (Multivariate Adaptive Regression Splines) yöntemi ile modellemektir. Bu amacı gerçekleştirmek için Lohmann LSL Classic beyaz hibrit sürü (n = 60) yumurtaları kullanıldı. Haftalık yumurta verimleri 22. haftadan 62. haftaya kadar değerlendirilmiştir. Bağımlı ve sürekli değişken olarak belirlenen yumurta ağırlığını (EW) tahmin etmek için; şekil indeksi (SI), kabuk kırılma mukavemeti (SBS), kabuk ağırlığı (SW), kabuk kalınlığı (ST), yumurta sarısı çapı (YD), yumurta sarısı genişliği (YW), yumurta sarısı yüksekliği (YH), yumurta sarısı rengi (YC) albümin genişliği (AW), albümin uzunluğu (AL), albümin yüksekliği (AH) kullanılmıştır. Mükemmel uyum iyiliği elde etmek için, R programının “earth” paketinde, penalty = -1, derece

Araştırma Makalesi

Makale Tarihçesi
Received : 09.04.2020
Accepted : 30.06.2020

Anahtar Kelimeler
MARS,
Yumurta ağırlığı,
Lohman ırkı,
Uyum iyiliği
Parametrik olmayan regresyon modeli.

= 2, nprune = 10 ve nk = 60 tanımları yapıldı. Araştırma sonucunda mars tahmin modeli, EW = 63.1-0.906 * max (0,75-SI) -0.321 * max (0, SI-75)-62.4*max(0,0.57-ST)-354*max(0,ST 0.57)+1.13*Groupa2*max (0,75-SI)+1.49* max(0.0.57-ST) * YD + 8.2 * max(0, ST-0.57)*YD-0.02*max(0 YD-38.5)*YC-0.0366* YH*max(0,13-YC) olarak belirlendi. Sonuç olarak, bazı kalite değişkenlerinin yumurta ağırlığının belirlenmesinde önemli olduğu bulunmuştur.Bağımlı değişken olarak belirlenen yumurtanın ağırlığını tahmin ederken a2, SI, YC, ST, YD, YH görülürken, diğer değişkenler bu denkleme dahil edilmemiştir. Tavukçulukta, MARS tahmin modeli, daha kolay formül ve daha yüksek doğruluk ile yumurta ağırlığını tahmin etmede klasik lineer olmayan modellere daha iyi bir alternatif olabilir.

To Cite : Canga D, Yavuz E, Efe E 2021. Egg Weight Estimation with Mars Data Mining Algorithm Using R. KSU J. Agric Nat 24 (1): 242-251. https://doi.org/10.18016/ksutarimdoga.vi.716880.

INTRODUCTION

Today, nutrition is one of the most important problems of people. For a healthy diet, the energy and nutrients of the body should be taken completely. Chicken meat and egg production meets people's daily protein and vitamin needs. Eggs are a food source with full biological value (Doğan, 2008; Durmuş, 2015). For this reason, studies on egg quality have an important place in numerous researches on eggs. Quality of an egg depends on the inner (albumin weight, yellow weight) and outer (shell weight) quality characteristics. (Orhan et al., 2016; Altan, 1993). Orhan et al 2016 used a regression tree method based on the CHAID algorithm to estimate the egg weight and achieved a high accuracy of 98.988%. In his research, he obtained the highest EW (71.963 g) from eggs with AW 41 g and YW> 17 g. Aktan (2004) found significant correlations between egg weight and, albumen and yolk weight (0.489, 0.796). Alkan et al. (2010); reported that egg weight, shell weight, shell thickness, yolk weight and albumin weight are important features to determine egg quality. Akan (2011) stated that there is a positive correlation between egg weight and albumin weight. In recent years, one way to produce estimates in the decision making process is also to use statistical methods especially in the field of data mining. These methods involve artificial neural network (YSA), decision trees and multivariate adaptive regression splines (MARS) as well as the others., YSA are information processing systems based on the structure and functioning of the biological nervous system, especially the human brain. The MARS algorithm is a data mining technique that can be used to solve classification and regression-type problems (Friedman 1991, Hastie et al 2009). Regression analysis is the most commonly used statistical technique to investigate and model the relationship between variables. There are many regression models used for various purposes such as data analysis and parameter estimation. One of these regression models, MARS algorithm, is a non-parametric regression method that succesfully describes the complex relationships between sets of dependent and independent variables. It is a nonparametric process to adapt to adaptive regressions that use some piecewise functions to define the relationships between sets of response(s) and predictors. Therefore, a functional relationship between dependent and independent variables is not accepted before analysis. In the method, MARS, regression and tree techniques were combined (Kibet 2012, Yakubu 2012). The MARS algorithm aims to optimize the fit of a dependent variable to the data by using the least squares method such as regression. Unlike regression, MARS allows the definition of more complex terms than those in the model that are linear and additive. Like decision trees, the MARS algorithm segments data, but unlike decision trees, MARS allows the capture of linear and additional relationships to be split over all nodes at every step.

Categorical or continuous characteristics can be modelled in this technique (Kibet 2012). MARS divides the data segments at equivalent intervals (Friedman 1991, Sevimli 2009, Kibet 2012). In each segment, MARS divides the data into several subgroups. Many nodes have been created by MARS. These nodes can exist between different input variables or different ranges in the same input variable to separate subgroups. MARS performs successfully in finding optimal variable transformations and interactions, which are complex data structures that often hide high-dimensional data (Steinberg 2001, Deconinck et al 2005, Yerlikaya 2008, Oguntunji 2017, Aksoy et al 2018a, Celik 2019).

In this study, theoretical information about MARS algorithm was given. With the given algorithm, it was aimed to calculate the eligibility criteria for the predictive performance of the "earth" package and "ehaGOF" package that will be used effectively in MARS analysis. The effective use of more than one continuous or discontinuous independent variable was demonstrated in the context of estimating a continuous dependent variable. The outputs obtained with the prediction equation created was easily interpreted.

Therefore, the aim of this study was to estimate selected quality features and to show the most effective estimators of these features by determining egg weight by using MARS data mining algorithm. In poultry production, prediction studies with this method are not common and classical methods are stil dominant. With this study, a new approach was tried to be presented to researchers working in this field.

MATERIAL and METHOD

Material

The research data on egg external and internal traits in the prediction of egg weight were obtainedfrom the experiment conducted in KSU, Faculty of Agriculture, Department of Animal Science during the year 2015. The obtained data was measured from the eggs of Lohman LSL Classic white hybrid chickens randomly matched without selection between the ages of 17-20 weeks (n = 60). For the study, weekly egg yields was evaluated from the 22nd week to the 62nd week. The study was aimed to estimate the egg weight (EW) as a continuous dependent variable. Shape index (SI), shell breaking strength (SBS), shell weight (SW) , shell thickness (ST), yolk diameter (YD), yolk width (YW), yolk heigth (YH), yolk color (YC), albumen width (AW), albumen length (AL), albumen height (AH) were considered as independent variables for egg weight prediction. Descriptive statistics of the variables examined in the study are presented in Table 1.

Table 1. Descriptive statistics of the studied explanatory variables
Çizelge 1. İncelenen açıklayıcı değişkenlerin tanımlayıcı istatistikleri

Variables	N	Min.	Max.	Mean	Std. Dev.
EW	57	58	64	60.59	1.407
SI	57	71	83	76.24	2.689
SBS	57	0	3	0.82	0.669
SW	57	7	11	8.28	0.790
ST	57	0	1	0.40	0.152
YD	57	36	47	40.93	2.828
AW	57	61	88	70.37	5.960
AL	57	42	117	88.21	10.258
AH	57	5	12	8.45	1.506
YH	57	16	22	19.02	1.185
YC	57	10	15	12.95	1.156
YW	57	13	19	1575	1.034

Method

The Multivariate Adaptive Regression Splines (MARS) algorithm, which provides high dimensional relationships between dependent and independent variable sets, does not require any assumptions about the distribution of the variables. The MARS algorithm, also known as the nonparametric regression method, allows researchers to create a prediction equation (Sevimli 2009, Kibet 2012, Aksoy et al 2018b).

The MARS algorithm is based on an adaptive regression approach that uses forward and backward procedures to generate the basic functions and to select the positions of the nodes. In each forward procedure, the entire area is subdivided and the nodes are added to their corresponding basic functions. In the backward procedure, unnecessary basic functions are deleted. This sustainability process is known as “pruning” and the optimal number of nodes can be found using general cross validation (GCV) (Kibet 2012, Zhang and Goh, 2016, Aytekin et al 2018, Celik and Yılmaz 2018, Eyduran et al 2017c, Sevgenler 2019, Eyduran et al 2019a, Canga and Boga 2019). In general, more basic functions (selected from a set of possible basic functions) are added to the model to maximize the goodness of fit criteria for the least squares. As a result of these operations MARS automatically determines the most important independent variables and the most important interactions between them. MARS is very good at finding optimal variable transformations and interactions, as well as the complex data structure that often hides high-dimensional data. The MARS model discovered by Friedman (1991) is a flexible nonparametric regression model for high dimensional data. Friedman (1991) extended the MARS methodology to the model with nominal categorical explanatory variables for which normal definitions of regularity are not applied.

Data mining techniques can be a good option to describe complex relationships. MARS is a non-parametric data mining technique that does not require assumptions such as normality and fixed variance (Kibet 2012, Eyduran et al 2019a, Celik and Boydak 2020. The MARS algorithm reveals the nonlinear and interaction effects between predictors and responses. Model's prediction accuracy increases as GCV (prediction error) decreases (Eyduran et al 2019a, Sevgenler 2019, Erturk et al 2018; Celik 2019).

Formation of basis functions

Parametric and nonlinear MARS method, contrary to linear methods, takes into account subsets of variables (Xu et al, 2006). In other words, the space created by the predictive variables is divided into many overlapping regions. Thus, it is created spline functions and these regional regression spans are also called the basic function (Put et al, 2009). The structural model constrcuted with MARS uses the piecewise linear basis functions expansion, which is shown in the form $(x - t)_+$ and $(t - x)_+$.

If the desired condition cannot be met in order to indicate the positive part of the "+" subscript, the value of the basic function (BF) will be zero and this is expressed as follows (Friedman 1991, Steinberg 2001, Deconinck et al 2005, Banks 2001, Sevimli 2010, Orhan et al 2018, Celik and Boydak 2020).

$$BF_1(x) = (x-t)_+ = \max(0, x-t) = \begin{pmatrix} x-t, x>t \\ 0,\ x \le t \end{pmatrix} \quad (1)$$

$$BF_2(x) = (t-x)_+ = \max(0, t-x) = \begin{pmatrix} t-x, x<t \\ 0,\ x \ge\ t \end{pmatrix} \quad (2)$$

Another representation of the basic functions (x - t)+ and (t - x)+ is x - t = max (x-t, 0) and (t - x)+ = max (t-x, 0).

The equation of the generalized MARS prediction equation for the default system that generates the data is given as follows ((Friedman 1991, Hastie et al 2001, Banks 2003, Ko et al 2008, Kibet 2012, Eyduran et al 2017a, Eyduran et al 2017d, Celik and Yılmaz 2018, Sevgenler 2019):

$$\hat{y} = \beta_0 + \textstyle\sum_{n=1}^{M} \beta_m \prod_{k=1}^{K_m} h_{km}(X_{v(k,m)}) \quad (3)$$

where;
$\hat{y}$: Estimated value of dependent variable,
$\hat{\beta}_0$: constant,
$\hat{\beta}_m$: regression coefficient,
$h_{km}\left(X_{v(k,m)}\right)$: basic function, the index of the independent variable of component m of the multiplier k.
K_m , is the parameter that limits the degree of interaction. Backward procedure in the screening process, GCV is used to compare the performance of model subsets to select the best subset. Lower GCV values are better at this step. GCV is a form of regularization that reveals goodness of fit against model complexity. Pruning algorithm is made by GCV method. GCV takes into account both the error of residuals and the model complexity, and the GCV value is calculated by the formula in equation (4):

$$\mathrm{GCV}(\lambda) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\left[1 - \frac{M(\lambda)}{n}\right]^2} \quad (4)$$

Here:
n: Number of eggs in the experiment,
y_i: Dependent variable; observed weight for i-th egg,
$\hat{y}_i$: Predicted weight for i-th egg,
$M(\lambda)$: λ is the function of the complexity of the model that contains the terms.
Goodness of fit criteria used for measuring the predictive accuracy of the MARS model are formulated below (Kibet 2012, Eyduran et al 2017a, Eyduran et al 2017d, Celik and Yılmaz 2018, Sevgenler 2019; Celik 2019) :

1. Pearson correlation coefficient (r) between real and predicted values of the dependent variable ,

2. Akaike information criterion, AIC

$$AIC = n \cdot \ln\left[\frac{1}{n}\textstyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] + 2k,\ if \qquad n/k > 40 \quad (5)$$

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \text{ , otherwise}$$

3. Root-mean-square error, RMSE:

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

4. Mean error, ME :

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

5. Mean absolute deviation, MAD:

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

6. Standard deviation ratio, SD_{ratio}:

$$SD_{ratio} = \frac{s_m}{s_d} \quad (9)$$

7. Global relative approximation error, RAE:

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}{y_i}^2}} \quad (10)$$

8. Mean absolute percentage error, MAPE:

$$MAPE = \frac{1}{n}\textstyle\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| . 100 \quad (11)$$

9. Performance index:

$$\rho = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{(1+r)\frac{1}{n}\sum_{i=1}^{n} y_i} . 100 \quad (12)$$

where: n is the number of egg weight in data set, k is the number of model parameters, yi is the real value of the dependent variable (egg weight), $\hat{y_i}$ is the predicted value of yi, Sm is the standard deviation of the model errors, Sd is the standard deviation of the dependent variable. Sd ratio for ensuring a good fit should be less than 0.40 and, for a very good fit Sd ratio should be less than 0.40 and, for a very good fit Sd ratio should be less than 0.10 (Grzesiak and Zaborski 2012, Eyduran et al 2017, Orhan et al 2018, Celik 2019; Celik and Boydak 2020).

Data mining techniques can be a good option to describe complex relationships. MARS is a non-parametric data mining technique that does not require assumptions such as normality and fixed variance. The MARS algorithm reveals the nonlinear and interaction effects between predictors and responses. Model's prediction accuracy increases as GCV (prediction error) decreases (Eyduran et al 2019a, Sevgenler 2019, Erturk et al 2018).

Statistical analysis

In the study, "earth" and "ehaGoF" packages were loaded using R STUDIO program and analysis of MARS algorithm was provided (R Core Team, 2014). Definitions used for the MARS algorithm were given in appendix.

RESULTS and DISCUSSION

In this study, R commands for estimating egg weight, which is the dependent variable, are determined. The script prepared for MARS analysis related to egg weight is presented in Figure 1 (Eyduran et al 2019a, Sevgenler 2019).

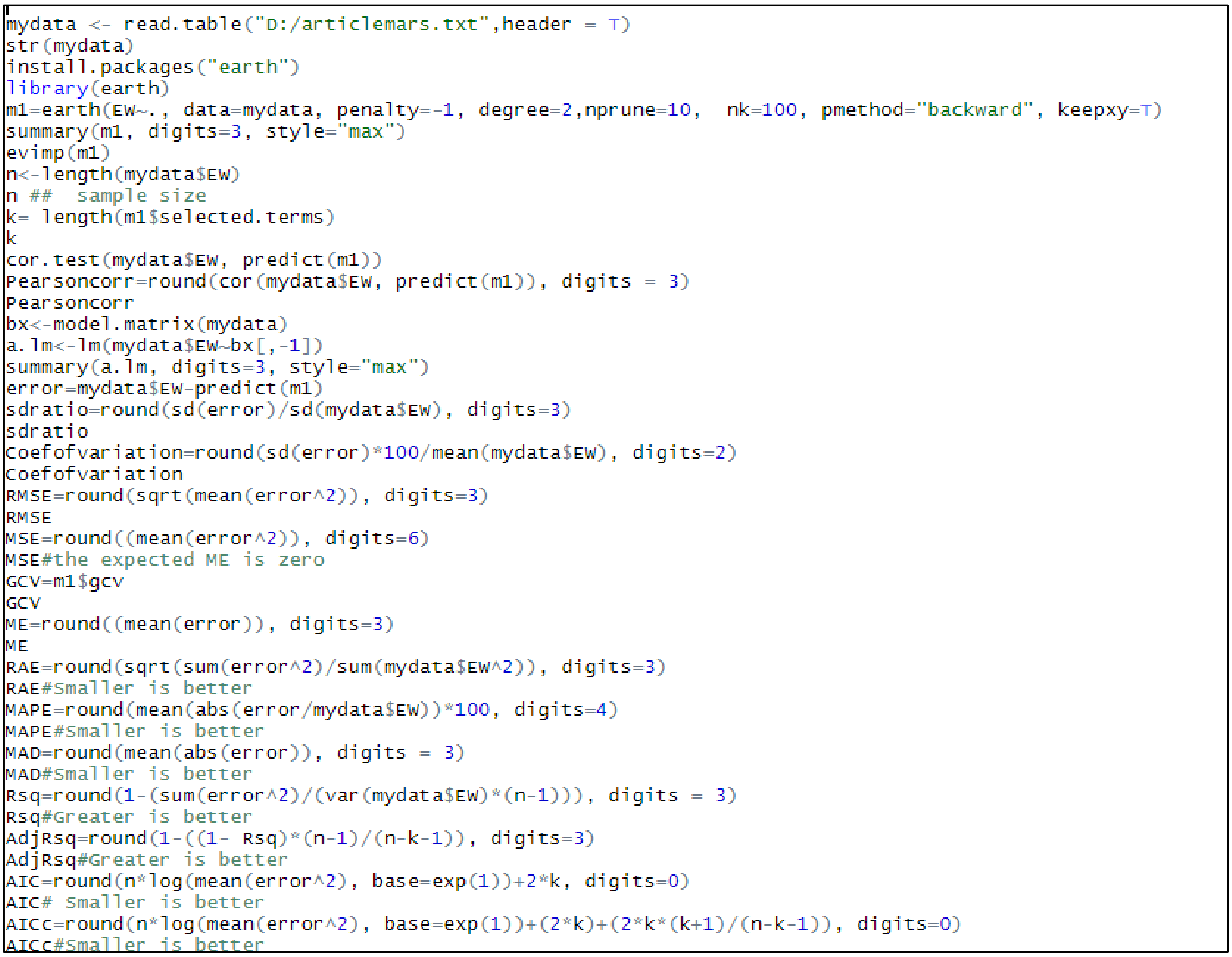
The codes of the prediction equation of the MARS model are given in the Appendix. The penalty =-1 and degree = 2 limitations are made in the earth package of the MARS algorithm used to estimate the egg weight. When summary results of MARS model were evaluated, it has been shown that the results have sufficient prediction performance (Friedman 1991, Steinberg 2001, Deconinck et al 2005, Erturk et al 2018, Millborrow 2018, Sevgenler 2019).

The prediction equation of the MARS results is given below:

EW= 63.1-0.906 * max(0,75-SI)-0.321 * max(0,SI -75)-62.4 * max(0,0.57 -ST)- 354 * max(0,ST -0.57)+ 1.13 * Groupa2*max(0,75-SI)+ 1.49 * max(0,0.57 -ST) * YD+ 8.2 * max(0,ST-0.57) * YD-0.0291 * max(0, YD-38.5) * YC -0.0366 * YH* max(0,13-YC).

When the basic functions are written in this equation, EW= 63.1-0.906 * BF1-0.321 * BF2- 62.4 * BF3- 354* BF4+ 1.13 * BF5+ 1.49 * BF6 + 8.2 * BF7- 0.0291 * BF8 -0.0366 * BF9" is obtained.

When the estimation equation is analyzed, it is expected that the variables in the 6th, 7th and 8th terms will have a positive effect on the dependent variable egg weight, while the other terms will have a negative effect (Eyduran et al 2019a, Orhan et al 2018, Celik and Boydak 2020, Sengul et al 2018). In Table 2, the coefficients of the estimation equation are given.



```
mydata <- read.table("D:/articlemars.txt",header = T)
str(mydata)
install.packages("earth")
library(earth)
m1=earth(EW~., data=mydata, penalty=-1, degree=2,nprune=10,  nk=100, pmethod="backward", keepxy=T)
summary(m1, digits=3, style="max")
evimp(m1)
n<-length(mydata$EW)
n ##  sample size
k= length(m1$selected.terms)
k
cor.test(mydata$EW, predict(m1))
Pearsoncorr=round(cor(mydata$EW, predict(m1)), digits = 3)
Pearsoncorr
bx<-model.matrix(mydata)
a.lm<-lm(mydata$EW~bx[,-1])
summary(a.lm, digits=3, style="max")
error=mydata$EW-predict(m1)
sdratio=round(sd(error)/sd(mydata$EW), digits=3)
sdratio
Coefofvariation=round(sd(error)*100/mean(mydata$EW), digits=2)
Coefofvariation
RMSE=round(sqrt(mean(error^2)), digits=3)
RMSE
MSE=round((mean(error^2)), digits=6)
MSE#the expected ME is zero
GCV=m1$gcv
GCV
ME=round((mean(error)), digits=3)
ME
RAE=round(sqrt(sum(error^2)/sum(mydata$EW^2)), digits=3)
RAE#Smaller is better
MAPE=round(mean(abs(error/mydata$EW))*100, digits=4)
MAPE#Smaller is better
MAD=round(mean(abs(error)), digits = 3)
MAD#Smaller is better
Rsq=round(1-(sum(error^2)/(var(mydata$EW)*(n-1))), digits = 3)
Rsq#Greater is better
AdjRsq=round(1-((1- Rsq)*(n-1)/(n-k-1)), digits=3)
AdjRsq#Greater is better
AIC=round(n*log(mean(error^2), base=exp(1))+2*k, digits=0)
AIC# Smaller is better
AICc=round(n*log(mean(error^2), base=exp(1))+(2*k)+(2*k*(k+1)/(n-k-1)), digits=0)
AICc#Smaller is better
```

Figure 1. R script file used to determine for egg weight
Şekil 1. Yumurta ağırlığını belirlemek için kullanılan R komut dosyası

If the difference between the node and observation value in above expressions regarding the basic functions given in Table2 is positive, this difference should be multiplied by the corresponding prediction coefficient in the model function. If the difference is negative, the basic function does not contribute to the model by taking the value of zero, in another word it is masked (Orhan et al 2018, Sahin et al 2018, Eyduran et al 2019a). Detailed examinations about this situation are mentioned in the literature (Sevimli 2009, Kibet 2012, Eyduran et al 2017a, Eyduran et al 2017d, Celik and Yılmaz 2018, Sevgenler 2019, Eyduran et al 2019a; Sahin et al 2018).

Table 2. Results of the MARS model regarding the basic functions and coefficients in the prediction of egg weight
Çizelge 2. Yumurta ağırlığının tahmininde temel fonksiyonlar ve katsayılara ilişkin MARS modelinin sonuçları

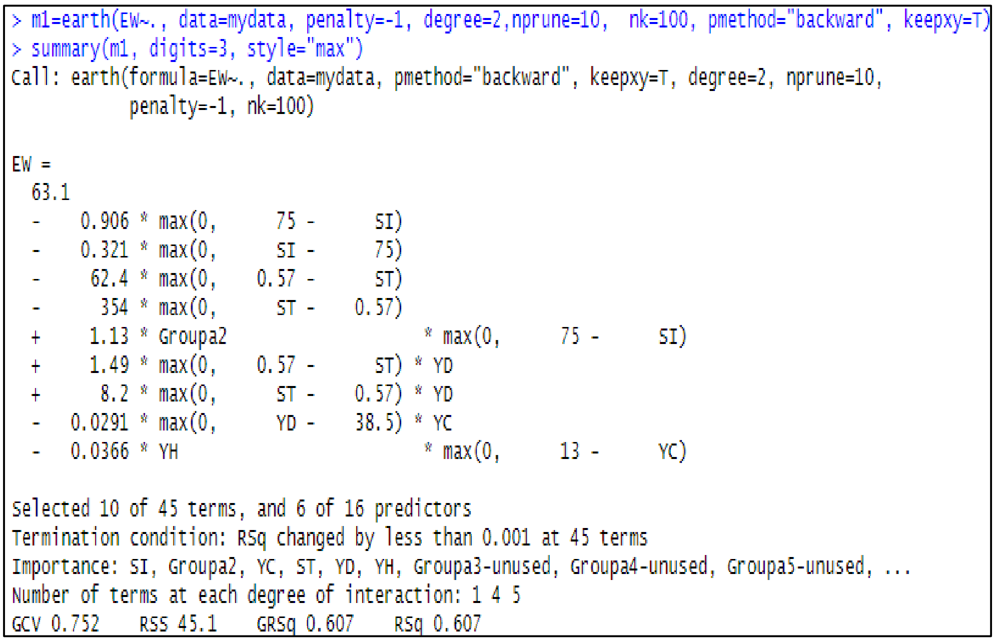
Terms	**Basis Function**		**Coefficients**
		Intercept	63.1
1	**BF1**	max(0,75-SI)	-0.906
2	**BF2**	max(0,SI -75)	- 0.321
3	**BF3**	max(0, 0.57-ST)	-62.4
4	**BF4**	max(0,ST -0.57)	-354
5	**BF5**	Groupa2*max(0,75-SI)	1.13
6	**BF6**	max(0,0.57-ST)*YD	1.49
7	**BF7**	max(0,ST-0.57) * YD	8.2
8	**BF8**	max(0, YD-38.5) * YC	-0.0291
9	**BF9**	YH*max(0,13-YC)	-0.0336

Selected 10 of 45 terms, and 6 of 16 predictors
Termination condition: RSq changed by less than 0.001 at 45 terms
Importance: SI, Groupa2, YC, ST, YD, YH, Groupa3-unused, Groupa4-unused, Groupa5-unused,
Number of terms at each degree of interaction: 1 4 5
GCV 0.752 RSS 45.1 GRSq 0.607 RSq 0.607

In Figure 2, the functions of the prediction equation obtained with the MARS algorithm and the coefficients of this function are given. In the MARS prediction model, the coefficient of determination was calculated as $R^2 = 0.61$. To obtain a smaller MARS model than the model obtained by default (with 10 terms selected), a small nprune with big nk and penalty = -1 and any desired term number is suggested. The predictive accuracy of the model increases as GCV decreases (Milborrow 2018, Eyduran et al 2019). In the study, a sufficient prediction model corresponding to the lowest GCV (0.752) value was produced. The results of the importance test of bound coefficients of the prediction equation produced by MARS algorithms are presented in Table 3 (Orhan et al 2018, Sengul et al 2018, Celik and Boydak 2020).

From Table 3, it was understood that all coefficients related to MARS estimation model were found statistically significant (P <0.001).

Looking at these results in Figure 3, the Pearson correlation coefficient between observed and predicted values was found to be 0.779 and the standard deviation ratio = 0.43, which means a sufficient fit (Grzesiak and Zaborski 2012, Erturk et al 2018, Eyduran et al 2019a). As a sophisticated approach, in the prediction model created by the MARS algorithm, the prediction of the dependent variable complies with the results of the previous studies such as Eyduran



```
> m1=earth(EW~., data=mydata, penalty=-1, degree=2,nprune=10,  nk=100, pmethod="backward", keepxy=T)
> summary(m1, digits=3, style="max")
Call: earth(formula=EW~., data=mydata, pmethod="backward", keepxy=T, degree=2, nprune=10,
            penalty=-1, nk=100)


EW =
  63.1
  -    0.906 * max(0,      75 -      SI)
  -    0.321 * max(0,      SI -      75)
  -     62.4 * max(0,    0.57 -      ST)
  -      354 * max(0,      ST -    0.57)
  +     1.13 * Groupa2                    * max(0,      75 -      SI)
  +     1.49 * max(0,    0.57 -      ST) * YD
  +      8.2 * max(0,      ST -    0.57) * YD
  -   0.0291 * max(0,      YD -    38.5) * YC
  -   0.0366 * YH                         * max(0,      13 -      YC)

Selected 10 of 45 terms, and 6 of 16 predictors
Termination condition: RSq changed by less than 0.001 at 45 terms
Importance: SI, Groupa2, YC, ST, YD, YH, Groupa3-unused, Groupa4-unused, Groupa5-unused, ...
Number of terms at each degree of interaction: 1 4 5
GCV 0.752    RSS 45.1    GRSq 0.607    RSq 0.607
```

Figure 2. Summary results of the MARS estimation model1
Figure 2. MARS tahmin modelinin özet sonuçları

Table 3 . Significance test results of coefficients in MARS algorithm
Çizelge 3. MARS algoritmasında katsayıların önem testi sonuçları

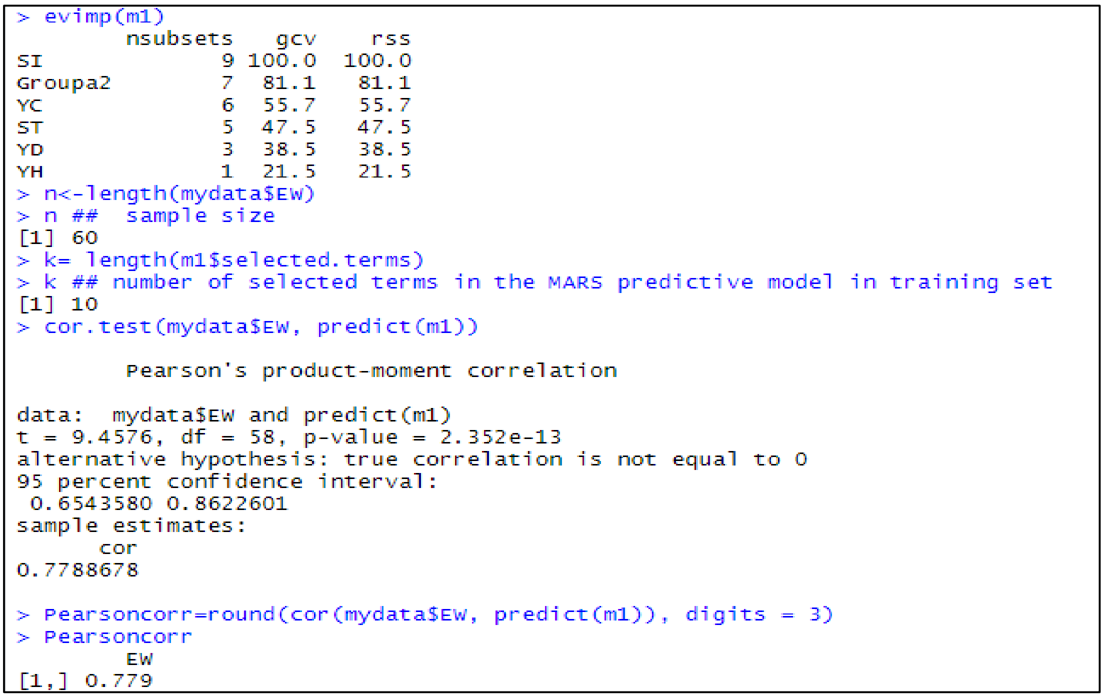
Basic function		**Estimate**	**Std.Error**	**P value**
	Intercept	63.1	0.44974	< 2e-16 ***
BF1	max(0,75-SI)	-0.906	0.07453	3.52e-05 ***
BF2	max(0,SI -75)	- 0.321	114.25216	0.002502 **
BF3	max(0, 0.57-ST)	-62.4	24.09248	0.001875 **
BF4	max(0,ST -0.57)	-354	2.68459	0.002479 **
BF5	Groupa2*max(0,75-SI)	1.13	1.4664	3.83e-06 ***
BF6	max(0,0.57-ST)*YD	1.49	0.60875	0.001663 **
BF7	max(0,ST-0.57) * YD	8.2	1.95565	0.001246 **
BF8	max(0, YD-38.5) * YC	-0.0291	0.04610	0.000108 ***

**: p<0.01; ***: p<0.001

and et al (2019a) and Celik and Boydak (2020). In the study, the SD rate of MARS was found to be 0.43. This shows that MARS model provides fit well. Here it has mainly been aimed at showing how to use the model. Therefore, it is suggested that it would be more appropriate for researchers to work on new researches with higher results.When these results are examined in Figure 3, the importance level of the variables has been determined with evimp (marsresult) command. Results of this command show the relative importance of statistics of goodness of fit. Shape index value has the highest relative importance (100%) in the prediction of egg weight. GCV criterion (81.1%) and RSS criterion (81.1%) of the 2nd group denoted by "groupa2" are also shown. Special statements can be made with the command "n <-length (mydata $ WEANINGW)" to test the significance of the terms.

In a study by Orhan et al (2016), they tried to find EW estimation using ridge regression (RR), multiple linear regression (MLR) and regression tree analysis (RTM) methods. In their study, they used only SW, AW, YW independent variables in EW estimation. However, in this study, almost all of the inner and outer quality features of the egg to determine egg weight were used. In this way, it was tried to determine which quality features should be used with the best MARS estimation equation in EW estimation.

Looking at Figure 3, it is seen that 6 of these features are important. In the EW estimation, it was seen that the SI variable was important in the first degree with its 100% significance level, while the study conducted by Orhan et al (2016) showed that the AW variable was important in the first place. When the results of this research are evaluated by looking at the literature, more detailed results about EW estimation appear as the number of variables related to quality characteristics increases (Orhan et al 2016, Sengul et al 2020).



```
> evimp(m1)
        nsubsets   gcv    rss
SI             9 100.0  100.0
Groupa2        7  81.1   81.1
YC             6  55.7   55.7
ST             5  47.5   47.5
YD             3  38.5   38.5
YH             1  21.5   21.5
> n<-length(mydata$EW)
> n ##  sample size
[1] 60
> k= length(m1$selected.terms)
> k ## number of selected terms in the MARS predictive model in training set
[1] 10
> cor.test(mydata$EW, predict(m1))

        Pearson's product-moment correlation

data:  mydata$EW and predict(m1)
t = 9.4576, df = 58, p-value = 2.352e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6543580 0.8622601
sample estimates:
      cor
0.7788678

> Pearsoncorr=round(cor(mydata$EW, predict(m1)), digits = 3)
> Pearsoncorr
        EW
[1,] 0.779
```

Figure 3. Summary results of the MARS estimation model2
Şekil 3. MARS tahmin modelinin özet sonuçları2

Finally, the model evaluation was performed using the ehaGoF package developed by Eyduran (2019b) for goodness of fit criteria (Table 4). In "ehaGoF" package output, the Akaike's information criterion (AIC) is preferred when the n/k is greater than 40. It is worth noting that otherwise there is a warning that the corrected Akaike's information criterion (AICc) should be used (Eyduran et al 2017b, c). Goodness of fit criteria are presented in Table 4 .

Table 4 . Goodness of fit criteria for MARS algorithm
Çizelge 4. MARS algoritması için uyum iyiliği kriterleri

	criterion	Value
1	Root mean square error (RMSE)	0.752
2	Relative root mean square error (RRMSE)	1.431
3	Standard deviation ratio (SDR)	0.627
4	Coefficient of variation (CV)	1.440
5	Pearson's correlation coefficients (PC)	0.779
6	Performance index (PI)	0.804
7	Mean error (ME)	0.000
8	Relative approximation error (RAE)	0.000
9	Mean relative approximation error (MRAE)	0.002
10	Mean absolute percentage error (MAPE)	1.140
11	Mean absolute deviation (MAD)	0.689
12	Coefficient of determination (Rsq)	0.607
13	Adjusted coefficient of determination (ARsq)	0.527
14	Akaike's information cCriterion (AIC)	2.889
15	Corrected Akaike's information criterion (AIC_c)	7.379

Also, goodness of fit statistics such as AIC, AICc, RMSE, Sdratio, R^2 and MAPE have been reported by some authors that models with the smallest value are the most suitable models (Orhan et al 2016, Celik et al., 2017; Eyduran et al 2017c, Celik 2019; Sengül et al 2020; Celik and Boydak 2020; Koyun and Çelik, 2020).

The findings were found to be compatible with the results of some recent studies on this subject (Eyduran et al 2017c; Eyduran et al 2019a, Zaborski et al 2019; Akin et al 2020, Celik and Boydak 2020). In this study, nprune = 10 and penalty = -1 were taken to prevent negative cross-validation value. Because in cases where the data set is small, the measurements should be made quite precisely (Milborrow 2018; Eyduran et al 2019a). In other words; with the backward elimination process, a model having the lowest GCV and nprune or less terms is selected. Therefore, the nprune command specifies the maximum number of terms allowed in the final model. This problem was solved by taking Penalty = -1.

Sum up , in the research, Theoretical information about the MARS algorithm is given and the prediction equation of the MARS algorithm is created with the codes that enable the "earth" package to be used effectively in terms of MARS analysis. Detailed results of the R resulting from these coding are included. Thus, it was thought that this would be important in future studies for the comprehensibility and progress of the subject.

CONCLUSION and SUGGESTIONS

In the study, independent variables of the shape index (SI), group a2, egg yolk color (YC) and shell thickness (ST), egg yolk diameter (YD), yolk height (YH) were effective in estimating the egg weight (EW) determined as the dependent variable. Other independent variables of SBS, SW, YD, AW, AL, AH, YW could not enter the model. The prdiction equation showed a medium level fit (SDR = 0.43) to the observed data. Looking at the coefficients in the estimation equation; While EW estimation, group a2, yolk diameter (YD), shell thickness (ST) variables positively affect; Shape index (SI), yolk color (YC), and yolk height (YH) variables affect negatively.

It can be concluded that this was due to the high variation in the data. In agriculture and animal science, this situation could be frequently encountered due to the sensitivity of the research and research material to environmental conditions. For this reason, researchers are especially recommended to work with low variation and high accuracy data in their studies with the MARS algorithm, which we recommend as a new method.

ACKNOWLEDGEMENT

This research was presented as an oral presentation at the International Congress on Domestic Animal Breeding Genetics and Husbandry (ICABGEH-2019) held on 11-13 September 2019 in Prague, Czech Republic.

Statement of Conflict of Interest

Author have declared no conflict of interest.

REFERENCES

Akin M, Eyduran, SP, Eyduran E 2020. Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. Plant Cell Tiss Organ Cult 140, 661–670.

Aksoy A, Erturk YE, Eyduran E and Tariq MM 2018a. Comparing predictive performances of MARS and CHAID algorithms for defining factors affecting final fattening live weight in cultural beef cattle enterprises. Pakistan Journal of Zoology, 50(6): 2279-2286.

Aksoy A, Erturk YE, Eyduran E, Tariq MM 2018b. Utility of MARS Algorithm for Describing Non-Genetic Factors Affecting Pasture Revenue of Morkaraman Breed And Romanov × Morkaraman F1 Crossbred Sheep Under Semi İntensive Conditions. Pakistan Journal of Zoology, 51(1):235-240.

Aytekin I, Eyduran E, Karadas K, Akşahan R, Keskin I 2018. Prediction of Fattening Final Live Weight from Some Body Measurements and Fattening Period in Young Bulls Of Crossbred And Exotic Breeds Using Mars Data Mining Algorithm. Pakistan Journal of Zoology, 50(1):189-195.

Banks DL, Olszewski RT, Maxion RA 2003. Comparing Methods for Multivariate Adaptive Regression. Communication in Statistics-Simulation and Computation, 32(2):541-571. [Electronic Journal], http://www.informaworld.com/smpp/title~content=t713597237.

Banks DL. 2001. Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression). In: International Encyclopedia of the Social & Behavioral Sciences. Eds: Smelser NJ, Baltes PB. Vol 8, 2nd ed, Elsevier, Amsterdam, p 5164-5169.

Banks DL. 2001. Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression). In: International Encyclopedia of the Social & Behavioral Sciences. Eds: Smelser NJ, Baltes PB. Vol 8, 2nd ed, Elsevier, Amsterdam, p 5164-5169.

Canga D, Boga M 2019. Hayvancılıkta Mars Kullanımı Ve Bır Uygulama. III. International Scientific and Vocational Studies Congress – Science and Health 27-30 June 2019, Ürgüp, Nevşehir / Turkiye.

Celik S, Eyduran E, Kaardaş K, Tariq MM. 2017. Comparison of predictive performance of data mining algorithms in predicting body weight in Mengali rams of Pakistan. Revista Brasileira de Zootecnia, 46(11): 863-872.

Celik S, Yilmaz O. 2018. Prediction of Body Weight of Turkish Tazi Dogs using Data Mining Techniques: Classification and Regression Tree (CART) and Multivariate Adaptive Reg-ression Splines (MARS). Pakistan Journal of Zoology, 50(2): 575-583 doi:10.17582/ journal.pjz/2018.50.2.575.583

Celik S. 2019. Comparing Predictive Performances of Tree-Based Data Mining Algorithms and MARS Algorithm in the Prediction of Live Body Weight from Body Traits in Pakistan Goats. Pakistan Journal of Zoology, 51(4):1447.

Celik, S. and Boydak E. 2020. Description of The Relationships Between Different Plant Characteristics in Soybean Using Multivariette Adaptıive Regression Splines (Mars) Algorithm. Japs, Journal of Animal and Plant Sciences, 30(2): 431-441.

Deconinck E, Xu QS, Put R, Coomans D, Massart DL, Heyden YV 2005. Prediction ofgastro-intestinal absorption using multivariate adaptive regression splines. Journal of Pharmaceutical and Biomedical Analysis, 39: 1021-1030.

Erturk YE, Aksoy A, Tariq MM 2018. Effect of Selected Variables Identified by MARS on Fattening Final Live Weight of Crossbred Beef Cattle in Eastern Turkey. Pakistan Journal of Zoology, 50(4):1403-1412.

Eyduran E, Zaborski D, Waheed A, Celik S, Karadas, K, Grzesiak W 2017a. Comparison of the predictive capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the indigenous Beetal goat of Pakistan. Pakistan Journal of Zoology, 49(1): 257-265.

Eyduran E, Akkus O, Kara MK, Tırınk C, Tarıq M M 2017b. Use of Multivariate Adaptive Regression Splines (Mars) in Predicting Body Weight from Body Measurements in Mengali Rams. International Conference on Agriculture, Forest, Food, Sciences and Technologies (ICAFOF), 11-17 May 2017, Nevşehir, Turkey.

Eyduran E, Tirink C, Karahan AE, Türkoğlu M 2017c. Prediction of an upper bound of gene-ralized cross validation in multivariate adaptive regression splines in agricultural studies. International Conference on Computational ans Statistical Methods in Applied Sciences, 9-11 Nov 2017, Samsun Turkey.

Eyduran E, Tirink C, Karahan AE, Türkoğlu M, Tariq MM 2017d. Comparison of Predictive Performances of MARS and CART Algorithms through R Software. International Conference on Computational ans Statistical Methods in Applied Sciences, 9-11 Nov 2017, Samsun, Turkey.

Eyduran E, Akin M, Eyduran SP 2019a. Application of Multivariate Adaptive Regression Splines in Agricultural Sciences through R Software. Nobel Bilimsel Eserler.

Eyduran E 2019b. ehaGoF: Calculates Goodness of Fit Statistics. R package version 0.1.0. https://CRAN.R-project.org/package=ehaGoF

Friedman JH 1991. Multivariate Adaptive Regression Splines. Annls. Stat. 19:1-141.

Hastie T, Tibshirani R, Friedman J 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, second ed. Springer.

Kibet CE 2012. A Multıvarıate Adaptıve Regressıon Splınes Approach to Predıct the Treatment Outcomes of Tuberculosıs Patıents in Kenya. Science in Biometry to The University of Nairobi, Yüksek Lisans Tezi,70s.

Ko M, Clark JG, Ko D 2008. Revisiting the Impact of Information Technology Investments on Productivity: An empirical investigation using multivariate adaptive regression splines (MARS). Information Resources Management Journal, 21(3):1-23.

Koyun M, Celik S. 2020. Investigation on Some Ectoparasites of Mesopotamian Spiny Eels (Mastacembelus mastacembelus) with Certain Data Mining Algorithms Based on the Effect of Weight and Sex. Pakistan Journal of Zoology, 52(2): 733.

Milborrow S 2018. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani. url: https://CRAN.R-project.org/package=earth (Erişim tarihi: 31.08. 2020).

Oguntunji AO 2017. Regression Tree Analysis for Predicting Body Weight of Nigerian Muscovy Duck (Cairina moschata). Genetika, 49(2): 743-753, 2017.

Orhan H, Teke Ç E, Karcı Z 2018. Laktasyon Eğrileri Modellemesinde Çok Değişkenli Uyar-lanabilir Regresyon Eğrileri (Mars) Yönteminin Uygulanması. KSU J Agric Nat 21(3): 363-373.

Put R, Xu QS, Massart DL, Vander Heyden 2004. Multivariate adaptive regressionsplines (MARS) in chromatographic quantitative structure–retention relationship studies. Journal of Chromatography A, 1055 : 11-19.

R Core Team 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, http://www.R-project.org.

Sahin G, Eyduran E, Turkoglu M, Sahin F 2018. Estimation of Global Irradiation Parameters at Location of Migratory Birds in Igdir, Turkey by Means of MARS Algorithm. Pakistan Journal of Zoology, 50(6): 2317-2324.

Sengul T, Celik S, And Sengul AY 2018. Bıldırcınlarda Göğüs Etinin Rengi ve Ph'sı Üzerine Yaş, Cinsiyet ve Canlı Ağırlığın Etkisi. Türk Tarım ve Doğa Bilimleri Dergisi, 5(4): 523-529.

Sengul AY, Sengul T, Celik S 2020. Relationships Between Body Weight and Some Egg Quality Traits in Japanese Quails. Turkish Journal of Agriculture-Food Science and Technology, 8(2): 308-312.

Sevgenler H 2019. Keçilere Ait Kimi Özelliklerin Canlı Ağırlık Üzerindeki Etkilerini Belirlemek Amacıyla Kullanılan Veri Madenciliği Algoritmalarının (Cart, Chaıd Ve Mars) Karşılaştırılması. Iğdır Üniversitesi Fen Bilimleri Enstitüsü Bahçe Bitkileri Anabilim Dalı, Yüksek Lisans Tezi, 57s.

Sevimli Y 2009. Çok Değişkenli Uyarlanabilir Regresyon Uzanımlarının Bir Split Mouth Çalışmasında Uygulaması. Marmara Üniversitesi, Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Yüksek Lisans Tezi, 87 s.

Steinberg D 2001. An alternative to neural networks: Multivariate adaptive regression splines (MARS), PC AI, January/February, pp. 38 -41.

Yakubu A 2012. Application of regression tree methodology in predicting the body weight of Uda sheep. Anim. Sci. Biotechnol., 45: 484-490.

Yerlikaya FA 2008. New Contribution to Nonlinear Robust Regression and Classification with Mars and Its Applications to Data Mining for Quality Control in Manufacturing, Master Thesis, METU, Ankara.

Xu QS, Daeyaert F, Lewi PJ, Massart DL 2006. Studies of relationship between biological activities and HIV Reverse Transcriptase Inhibitors by Multivariate Adaptive Regression Splines with Curds and Whey. Chemometrics and Intelligent Laboratory Systems, 82: 24-30.

Zhang W, Goh AT 2016. Multivariate Adaptive Regression Splines And Neural Network Models For Prediction Of Pile Drivability. Geoscience Frontiers, 7(1): 45-52.

APPENDİX

For the MARS algorithm using the R studio program, the following definitions have been made:

```
m1=earth(EW~., data=mydata, penalty=-1,
degree=2,nprune=10, nk=60, pmethod="backward",
keepxy=T)
summary(m1, digits=3, style="max")
```

The following R definitions are used to test the significance of coefficient for terms created from important variables as follows : .

```
bx<-model.matrix(mydata)
a.lm<-lm(mydata$EW~bx[,-1])
summary(a.lm)
evimp(m1)
```