Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

JAS

Use of MARS Data Mining Algorithm Based on Training and Test Sets in Determining Carcass Weight of Cattle in Different Breeds

Demet ÇANGA[a]

[a]*Osmaniye Korkut Ata University, Department of Chemistry and Chemical processing, Bahçe, Osmaniye, TURKEY*

ARTICLE INFO
Research Article
Corresponding Author: Demet ÇANGA, E-mail: demetcanga@osmaniye.edu.tr
Received: 30 October 2020 / Revised: 02 April 2021 / Accepted: 15 May 2021 / Online: 25 March 2022

ABSTRACT

This research was carried out with the purpose of estimating hot carcass weight by using parameters such as race, carcass weight and age with Multivariate Adaptive Regression Spline (MARS) algorithm. To achieve this goal, 700 cattle data belonging to the years 2017-2018, which were taken in equal numbers from 7 different breeds, were used. A total of 700 data were used, taking equal numbers of data from each breed. In order to test the accuracy of the model created in the research, the data set was divided into two data subsets as training and test subsets. In order to test the compatibility of these separated subsets with the MARS model, a new package program named "ehaGoF" which estimates 15 goodness of fit criteria was used. According to the analysis results, the MARS model with the smallest SD_{RATIO} (0.157, 0.130) and the highest determination coefficient (R^2) (0.975, 0.983) of the training and test sets, respectively, was determined. Looking at the other fit values, it is seen that the training and test set are quite compatible. In terms of hot carcass weight among the breeds, it was determined that the Limousine race performed higher than the other breeds. As a result, the implementation of the MARS algorithm can allow livestock breeders to obtain effective clues by using independent variables such as breed, age, and body weight in estimating hot carcass weight.

Keywords: Carcass weight, MARS algorithm, Multiple regression analysis, Beef cattle, ehaGoF, K-fold Cross Validation

1. Introduction

Beef, Turkey as well as to people all over the world in an adequate and balanced nutrition emerge as one of the most important resources. Although there are many cattle breeds in the world, most of the meat and milk production is met from certain breeds. native cattle breed in Turkey, is reported to be low meat yield in conventional farming conditions and the fattening end weight varied between 186-387 kg, and the daily live weight gain of 673-973 g (Kumlu 2000; Sak & Duru 2017). The carcass obtained from cattle is affected by many parameters. Among these, there are many factors such as the breed, sex, fattening period, care and nutrition of the cattle (Sak & Duru 2017; Seker et al. 2017). In the study conducted to determine the fattening performance and carcass characteristics of Simental, Aberdeen-Angus, Hereford, Limousine and Charolais breeds, it was reported that while Charolais performed better than others for carcass weight, the Simental breed was higher for daily weight gain (Sak & Duru 2017). Galiç & Takma (2019) have conducted on the determination of live weight and the genetics factors affecting this situation. In addition, it was stated in the studies that subjective and objective methods can be used in estimating the carcass composition in live animals. At the same time, the fact that the devices used to estimate the carcass composition in live animals are very expensive is seen as the most important obstacle in front of the studies. Therefore, since the most important thing in animal production is the production of animal products economically, it may be important to do different researches on this subject (Kor & Ertuğrul 2000). There are many scientific studies written for this purpose. However, in the breeding practice, it is very important to estimate the properties of the independent variables that the researcher should use economically. Such forecasts can assist the farmer in the decision process regarding herd management. Rather, these situations can be decided by looking at the physical characteristics of the animals subjectively in farm conditions. However, one way of producing such estimates, especially when the number of data is large, can be obtained by using statistical methods such as data mining. These methods include, among others, artificial neural networks (ANNs), decision trees, and Multivariate Adaptive Regression Spline (MARS) (Kibet 2012; Eyduran et al. 2018; Orhan et al. 2018; Eyduran et al. 2019). With the MARS algorithm used in this study, linear models are explained by dividing nonlinear multivariate models with more than one independent variable. As is known, regression analysis investigates the relationship between two or more variables with a cause-effect relationship. The main purpose of chains MARS analysis, which is an application of the techniques popularized by Friedman (1991) to solve regression type problems; estimating the result variable or the value of a continuously dependent variable with the set of independent variables. MARS offers the opportunity to be explained with linear models by breaking down multivariable nonlinear models (Sevimli 2009; Eyduran et al. 2017a; Eyduran et al. 2018; Orhan et al. 2018; Eyduran et al. 2019). As in other scientific fields, the selection of breeds gives effective results in determining the weight of the carcass, which is one of the main issues within the scope of animal breeding. To obtain these results, powerful statistical methods, ie data mining algorithms, are required. MARS, one of these algorithms, is a statistically significant tool that can capture the relationship between dependent and independent variables. There are other studies that use the MARS algorithm in agriculture and animal husbandry (Aksoy et al. 2018; Aytekin et al. 2018; Celik &Yılmaz 2018; Celik et al. 2020; Canga & Boga 2019; Canga et al. 2019; Eyduran et al. 2019). The research also included goodness of fit values and the comparative use of training and test sets. Here, the training data set is a sample data set used for learning and created in accordance with the parameters of a classifier. The test data set is a data set that is independent of the training data set, but follows the same probability distribution as the training data set. For this, the estimation equation was created based on all the values in the train set, and then the accuracy of the values created with the test set was compared with the values in the train set.

When the literature is reviewed, estimation of economically determined variables such as estimation of live carcass weight has not been investigated by MARS algorithm yet. Therefore, the use of live weight and carcass weight of animals used for meat production in Turkey and there is a gap in terms of increasing productivity. This research was carried out to estimate the hot carcass weight efficiency using the MARS algorithm. Therefore, the aim is to draw attention to this issue and to contribute to the literature by leading more comprehensive research in this field.

2. Material and Methods

2.1. Materials

In the research, data belonging to the cattle belonging to the year 2017-2018 brought to the slaughterhouse in the open prison in Niğde province for slaughter from the provinces other than Niğde and Niğde were used. In the data used in the study, a total of 700 male calf data, 100 from each of the Aberdeen-Angus, Simmental, Limousine, Holstein–Friesian, Charolais, Zebu and Hereford breeds, were used. Live weight, age and breed were used as independent variables in the estimation of hot carcass weight determined as the dependent variable.

The animals were slaughtered between 2017-2018 in line with the observational decisions of the technical staff in the slaughterhouse belonging to the enterprise. For this, first of all, descriptive statistics values of the data are shown in Table 1.

Table 1- Descriptive statistics of the explanatory variables studied

	Descriptive Statistics					
	N	*Minimum*	*Maximum*	*Mean*	*Std. Error*	*Std. Deviation*
AGE (year)	700	1	3	1.63	0.025	0.650
LIVEWEIGHT (kg)	700	440	1020	696.520	4.299	113.749
CARCASSWEIGHT(kg)	700	237	515	353.390	2.178	57.611

2.2. Methods

MARS method, which is one of the non-parametric regression methods, was developed by the statistician Jerome H. Friedman in the early 1990s (Tunay 2001; Mukhopadhyay & Iqbal 2009; Sevimli 2009). This regression method is designed for both continuous and binary response variables. MARS is a nonparametric regression method that makes no assumptions about the underlying functional relationships between dependent and independent variables (Kibet 2012; Oguz 2014; Eyduran et al. 2019). The main purpose of this method is to predict the values of the result variable or continuous variable regardless of the set of independent variables. The biggest advantage of the MARS model is that it defines both the individual effects of independent variables and their interactions in the model and presents them graphically (Chou et al. 2004; Sevimli 2009; Orhan 2018).

2.2.1. Mars model

Regions and spline functions are formed, and these regression extensions, which are regional, are called the basic functions (Put et al. 2004). The basic functions that occur are in a piecewise linear relationship with the dependent variable. Fundamental functions and model parameters (estimated by least squares method); consists of the results of the determinants that gave the data entries. Basic functions (BF) are mechanisms used in generalized searches for nodes. BF is a set of functions used to represent information contained within one or more variables. The structural model created with MARS uses the piecewise linear basis functions expansion, shown in the form of $(x - t)_+$ and $(t - x)_+$. The "+" subscript used here indicates the positive part and indicates that the basic function will take the result of zero when the desired condition is not met, and the formation process of the BF is defined as follows (Friedman 1991; Deconinck et al. 2005; Kayri 2010).

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

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

Here "t" is the node value and each function is linear piecewise at the value of "t". The fundamental functions (x - t) $_+$ and (t - x) $_+$ (linear extensions) are also called a reflected pair, denoting the right and left regions of the node "t", respectively (Hastie et al. 2001; Sevimli 2009; Oğuz 2014; Eyduran et al. 2017b; Eyduran et al. 2018; Orhan et al. 2018).

2.2.2. MARS model selection criteria

How to measure the accuracy of the model is the most important issue in regression problems. For this purpose, the generalized cross validation (GCV) value developed by Craven & Wahba (1979) in the selection criteria of the most suitable MARS model measures the accuracy of the mean squares errors (Sevimli 2009). As a result, GCV is a form of regulation that transforms model complexity into goodness of fit. With the GCV approach, BF that have the least contribution to the model are thrown into the model, preventing the addition of excessive number of extension functions in the final model. The GCV criterion, which is the goodness of fit criterion, can be defined as follows (Xu et al. 2006; Grzesiak et al. 2010; Ali et al. 2015; Zhang & Goh 2016; Celik & Yılmaz 2018; Çanga & Boga 2019; Eyduran et al. 2019; Zaborski et al. 2019; Celik & Boydak, 2020; Eyduran & Gulbe 2020).

$$GCV(M) = \frac{1}{N}\sum_{i=1}^{N} \left(\left[y_i - \hat{f}_M(x_i)\right]^2 \Big/ \left[1 - \frac{C(M)}{N}\right]^2 \right) \tag{3}$$

Where; *N*, number of observations; *C(M)*, constant basic function; *C(M) = M + dM*, d is the smoothing parameter. Studies have shown that the best value for d is between $2 \le d \le 4$ (Friedman, 1991; Salford 2001). The most appropriate MARS model is the value with the smallest GCV measurement (Xu et al. 2004). The quality of the MARS model in the study was evaluated using the following criteria (Grzesiak et al. 2010; Ali et al. 2015; Zhang & Goh, 2016; Çelik &Yılmaz, 2018; Çanga & Boga, 2019; Eyduran et al. 2019; Eyduran et al. 2019; Zaborski et al. 2019; Celik & Boydak 2020). Pearson's correlation coefficient between actual values and predicted values (*r*).

1. Akaike Information Criteria (*AIC*):

$$AIC = n.\ln\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2\right] + 2k\ , if; \frac{n}{k} > 40 \tag{4}$$

2. Root-mean-square error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2} \tag{5}$$

3. Mean error (*ME*)):

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right) \tag{6}$$

4. Absolute mean deviation (*MAD*):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_{ip}\right| \tag{7}$$

5. Standard deviation rate (*SD_{ratio}*):

$$SD_{ratio} = \frac{S_m}{S_m} \tag{8}$$

6. Relative approximate error rate (*RAE*):

$$RAE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - y_{ip}\right)^2}{\sum_{i=1}^{n} {y_i}^2}} \tag{9}$$

7. Average absolute error percentage (*MAPE*):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{ip}}{y_i}\right|.100 \tag{10}$$

8. Performance index (ρ):

$$\rho = \frac{\sqrt{\sum_{i=1}^{n} \left(y_i - y_{ip}\right)^2}}{(1+r)\frac{1}{n}\sum_{i=1}^{n} y_i}.100 \tag{11}$$

Where; *n*, the number of observations in the data set; *k*, Number of model parameters (selected terms); y_i, i. dependent variable value of the observation; y_{ip}, i. dependent variable estimation value of the observation; S_m, Standard deviation of model error terms and S_d, Standard deviation of the dependent variable expresses (Eyduran et al. 2020).

2.3. Cross Validation process for training and test set

Cross validation is any of the model validation techniques used to evaluate how statistical analysis results will be generalized to an independent data set.

The purpose of cross validation is to test the model's ability to predict new data that is not used in its prediction, to mark problems such as selection bias, and to give an idea of how the model will generalize a generalization method. In a cross validation process, the analysis is first divided into complementary subsets of a data sample. Then, the analysis is performed in a subset called the training set and the analysis is verified in the other subset called the test set. To reduce variability, in most methods, multiple rounds of cross validation are performed using different parts. That is, cross validation combines fit measures in predictions to get a more accurate estimate of model forecast performance (Geisser 1993; Ripley 1996; Salford 2001).

Cross validation is done in a predetermined number of k. In the literature, 10-fold expression of cross-validation is also common. The data is divided into k pieces of equal size and evaluated k times. When doing cross validation, researchers first divide the data sets into two sets, training and test sets. Next, they choose X% of the training data set as the actual test set and the remaining (100-X) % as the verification set, where X is a constant number. If we define x as 80, the 80% model is repeatedly trained and verified on these different sets. There are multiple ways to do this and it is commonly known as cross validation. K-fold cross validation is the most used cross validation method (Devijver & Kittler 1982; Kohavi 1995). K-fold cross validation is a special case of cross validation where we iterate over a dataset k times. In each round, the data set is split into k parts and one split part is used for validation, then the remaining k - 1 parts are combined into a training subset for model evaluation. Figure 1 below shows the step-by-step cross-validation process (Anonymous 2021):

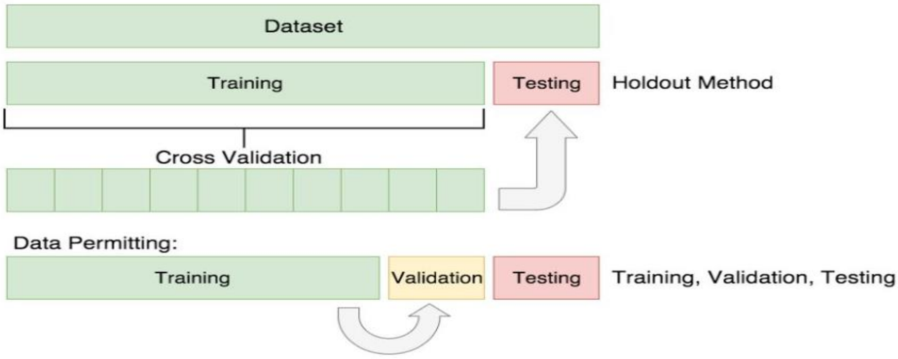

Figure 1- Visual representation of train/test split and cross validation

Some descriptive statistical values in the analysis were performed using IBM SPSS 23.0 software, and the MARS model was developed using the R software "*earth"* package and "*ehaGoF*" (Milborrow 2011; R Core Team 2014; Milborrow 2018; Eyduran et al. 2019; Eyduran & Duman 2020; Eyduran & Gulbe 2020). Among the observed and predicted values of carcass weight in the study, the smallest *GCV*, *SDRATIO*, *RMSE*, *MAPE*, *MAD*, *AIC*, *AICc* and the MARS model with the highest determination coefficient (R^2) and Pearson coefficient (*r*) were accepted as the best.

3. Results and Discussion

In the research, the carcass yield was analyzed with the MARS method. Graphics of hot carcass weight, breeds, body weight and age, which are the dependent variables for better understanding of the data, are given in Figure 2.

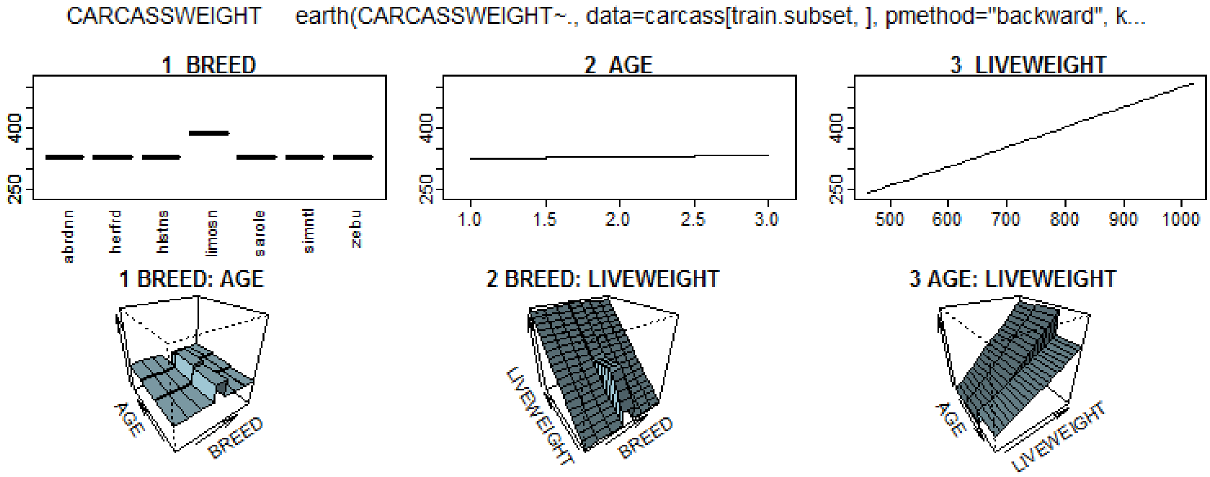

Figure 2- Distribution of breeds (1), age (2), live weight (3) by carcass weight

When looking at the data obtained from 700 cattle beef cattle from 7 different breeds of different ages; it is seen that the carcass weight is in the animals belonging to the highest limusin race, while the other breeds are very close to each other (Figure 2). Duru & Sak (2017) aimed to determine the fattening performance and carcass characteristics of Simmental, Aberdeen-Angus, Hereford, Limousine and Charolais breeds. In this study, 606 male cattle aged 10-12 months imported from Uruguay and France in 2015 were used. All animals were fed unlimitedly with the same ration during the fattening period of about 7-9 years. Daily weight gain (DWA) Simmental 1362.9; Aberdeen-Angus 1275.9; Hereford 1214.2; Limousine 1266.9; Charolais was obtained as 1101.1 g. This situation was found to be similar to that the limusin breed in our study was more than the charolais and hereford breeds. Differences with other breeds indicate that it can also be affected by different factors such as environment, care and feeding. The age of animals with a carcass weight between 300 and 350 varies between 1 and 3 years (Figure 2). According to the research data, it is seen that as the body weight increases, the carcass weight increases positively (Figure 2). Age is one of the important factors affecting the fattening performance in beef cattle. The development of culture breeds and their hybrids continues until maturity. This period is 18 months. In domestic breeds, it is 2.5-3 years old (Yıldız 2020). Looking at Figure 3, it is seen that there is a positive relationship between the real and predicted values on the training and test set, respectively, according to the results of MARS analysis (r = 0.994).

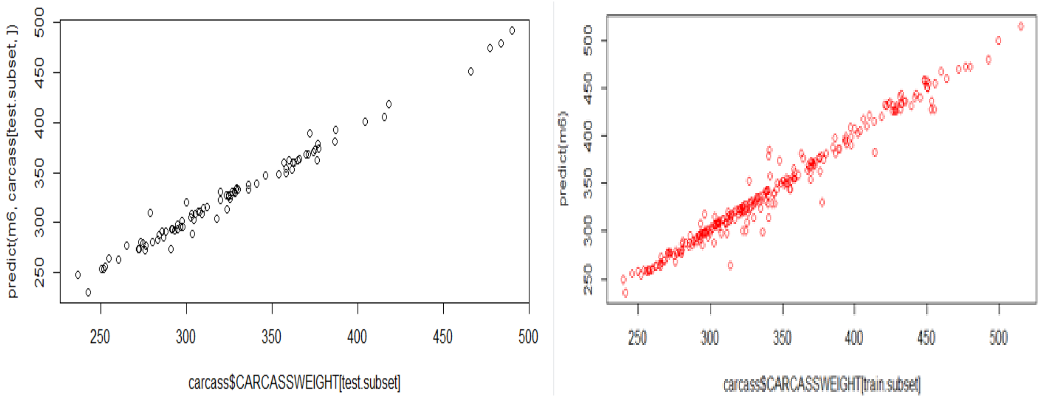

Figure 3- Relationship between real and predicted values for test set and training set

The prediction model formed as a result of MARS analysis is mathematically written as follows:

CARCASSWEIGHT = 296.807 + 0.492 * max (0, LIVEWEIGHT-585) - 0.444 * max (0,585-LIVEWEIGHT) -9.481 * BREEDLimousine + 0.554 * BREEDLimousine * max (0,737.22-LIVEWEIGHT) - 99.854 * BREEDLimousine * max (0, AGE-2) +147.262 max (0, AGE-2) -6.151 max (0,2-AGE) -0.921 * BREEDLimousine * max (0,585-LIVEWEIGHT) -0.218 * max (0, AGE-2) * LIVEWEIGHT.

This prediction function is represented in terms of basic functions as follows:

CARCASSWEIGHT = 296.807 +0.492 * BF1 - 0.444 * BF2- 9.481 * BF3 + 0.554 * BF4 - 99.854 * BF5 + 147.262 * BF6-6.151 * BF7 - 0.218 * BF8.

Here y is defined as the carcass weight. BF is defined as the basic function. When looking at the MARS estimation model by considering equation 2 and equation 3; For the fundamental function BF1 max (a, b) = a, a> b otherwise the result will be b. So, when looking at the equation, there are two nodes, 585 and 737. Thus, the two nodes in 585 and 737 divide the interval into three intervals in which different linear relationships are determined (Canga & Boga 2019; Eyduran et al. 2019; Celik & Boydak 2020). The results of the MARS algorithm for carcass yield are presented in Table 2. As can be seen, all coefficients for the carcass yield were statistically significant (P<0.001). In interpretation, for example, when the body weight is less than 585; max (0, LIVEWEIGHT-585) = 0, so the effect of the MARS term number 1 is masked in the carcass yield. Likewise, when the age of the animal is less than 2; max (0, AGE-2) = 0, i.e. the effect of MARS terms 5th, 6th, and 9th is masked on carcass yield (Celik & Yılmaz 2018; Celik et al. 2020; Canga & Boga 2019; Canga et al. 2019; Eyduran et al. 2019; Sevgenler 2019; Zaborski et al. 2019; Canga & Boga 2020; Sengul et al. 2020).

When LIVEWEIGHT is> 585 cm (BF1), the carcass yield is expected to be higher and when LIVEWEIGHT is greater than 585 cm, the carcass yield will be above BF2, so it is masked. However, when the breed is Limousine, LIVEWEIGHT has a positive effect on carcass weight in live animals with <737.22 cm (BF4, P <0.001). In addition, when the breed is Limousine, LIVEWEIGHT has a negative effect on carcass weight in live animals with <585 cm (BF7, P <0.01). When AGE> 2 there is a positive high effect on the carcass weight for the BF5 coefficient (147,262). In addition, the negative effect of the BF6 coefficient (-6.151) on the carcass weight for animals with AGE> 2 was masked. Also, for animals with AGE <2, the effect of LIVEWEIGHT on carcass weight was masked by age, considering BF8 (first order interaction term). However, for animals with AGE> 2, LIVEWEIGHT has a negative effect on carcass weight (BF8, P <0.001)

In the study, the cross validation coefficient (CVR^2) was found to be 0.959, while the highest Pearson correlation coefficient (r) between the real and predicted values was 0.997 and the results were accepted as the best model. The suitability of this model was evaluated with the criteria for the *GCV* to be minimum and R^2 to be maximum (Celik & Yılmaz 2018; Celik et al. 2020; Canga & Boga, 2019; Canga et al. 2019; Eyduran et al. 2019; Eyduran et al. 2019; Sevgenler 2019; Zaborski et al. 2019; Sengul et al. 2020).

Table 2- Coefficients of the MARS model and results of MARS analysis

	Basis Functions (BFi)	*Coefficients*	*P*
	Intercept	296.807	<2e-16 ***
BF1	max (0, LIVEWEIGHT-585)	0.492	<2e-16 ***
BF2	max (0,585-LIVEWEIGHT)	-0.444	<2e-16 ***
BF3	BREEDlimousine	9.481	4.00e-09 ***
BF4	BREEDlimousine* (0,737.22-LIVEWEIGHT)	0.554	<2e-16 ***
BF5	BREEDlimousin*max(0,AGE-2)	-99.854	<2e-16 ***
BF6	max(0,AGE-2)	147.262	1.94e-06 ***
BF7	max(0,2-AGE)	-6.151	1.69e-08 ***
BF8	BREEDlimousine*max(0,585-LIVEWEIGHT)	-0.921	4.05e-16 ***
BF9	max(0,AGE-2)*LIVEWEIGHT	-0.218	1.55e-05 ***
	GCV: 88.2 RSS: 24531 GRSq: 0.972 RSq:0.975 CVRSq:0.959		

***: P<0.001

As can be seen from Table 3, it is seen that the live weight value has the highest relative importance (100%) in the estimation of carcass weight, both the GCV value and the RSS criterion (Eyduran et al. 2019; Sevgenler 2019; Canga & Boga 2020; Celik & Boydak 2020; Eyduran & Duman 2020; Sengul et al. 2020).

Table 3- Relative importance of independent variables in the model

Variables	*Nsubsets*	*GCV*	*RSS*
LIVEWEIGHT	9	100.00	100.00
BREEDLimousine	7	22.3	22.5
AGE	6	18.5	18.7

3.1. Reducing model bias by creating training and test subsets

Model bias will be reduced with the training and test sub-sets created in the researc4. Therefore, in this study, the significance of the model was tested as follows. The model is divided into two parts as training and test set. Also, like most other things in machine learning, the split ratio in the training / test set is highly specific to your use case, making it easier to master the situation while developing more models. Here, this ratio (75:25) has been chosen; This means that 75% of the observations of the model are included in the training data set, while 25% of the observations of the model are included in the test data set. First, a model was obtained with the training data set, and then the reliability of this model was sent to the test set and the resulting values were compared. Thus, all values were checked by performing the validity process. While performing the analysis, the test set is not used until all studies related to the model are completed, it is used to test this model after the final model is decided. Using the test set more than once increases the model bias. In this case, it will not make sense to do such a partition. In other words, the training data set used in the research is the data sample used to provide an unbiased assessment of a final model that fits the data set, and the test data set provides the gold standard used to evaluate the model. It is used only after the model has been fully trained. The R codes created for the training and test sub-sets in the research and which will reduce the model bias are given in Appendix (Eyduran & Duman 2020).

The most suitable model was determined by testing the goodness of fit values of the carcass weight, which was determined as the dependent variable in the study, with the training and test set. To estimate the goodness of fit between the training set and the test set, was used in the R program of the "EhaGof" package. The most important aspect of this newly created package is that all 15 different goodness of fit criteria are calculated at the same time. After all these calculations, a comparative representation of the goodness of fit values on the training and test set of the most suitable model is given in Table 4 (Zhang & Goh 2016; Eyduran et al. 2019; Zaborski et al. 2019).

Table 4- Goodness of fit criteria for training set MARS and test set MARS algorithms

	Criterias	*Train-Set MARS results*	*Test- set MARS results*
1	Root mean square error (RMSE)	77.877	47.382
2	Relative root mean square error (RRMSE)	2.571	2.049
3	Standard deviation ratio (SDR)	**0.157**	**0.130**
4	Coefficient of variation (CV)	**2.570**	**2.060**
5	Pearson's correlation coefficients (PC)	**0.998**	**0.992**
6	Performance index (PI)	1.294	1.029
7	Mean error (ME)	0.000	0.391
8	Relative approximation error (RAE)	0.001	0.000
9	Mean relative approximation error (MRAE)	0.001	0.002
10	Mean absolute percentage error (MAPE)	**1.587**	**1.533**
11	Mean absolute deviation (MAD)	5.436	4.984
12	Coefficient of determination (Rsq)	**0.975**	**0.983**
13	Adjusted coefficient of determination (ARsq)	**0.974**	**0.981**
14	Akaike's information cCriterion (AIC)	1391.866	425.114
15	Corrected Akaike's information criterion (CAIC)	1392.590	427.455

By comparing the goodness of fit criteria of the model for the data belonging to the training and test set, it is ensured that bias is reduced by performing cross validation. In this case, goodness of fit criteria such as R^2 and *RMSE* were used. As can be seen from Table 4, the MARS model with the smallest SD_{RATIO} (0.157, 0.130) and the highest determination coefficient (R^2) (0.975, 0.983) of the training and test sets, respectively, was determined. In the model, it can be said that the bias of the model is low since the goodness of fit criteria of the training and test sets are very close to each other (Eyduran & Duman 2020). Some authors reported that the standard deviation ratio of the structured model, which fits well for regression-type problems, should be less than 0.20.

Seven different cattle breeds were used in the study, and only Limousine race among these breeds was found to be statistically significant in determining the carcass yield (P<0.001). In the report, the average obtained for hot carcass weight in Limousine breed is 319.3 ± 3.5 kg. This value is 386.76 in this study, Zahrádková et al. (2010) and Anonymous (2020) can be said to be an average value. As is known, the reliability of the obtained results depends on the selection of effective independent variables and strong statistical approaches (Eyduran et al. 2018). This study provides good evidence with the resulting results regarding the superiority of the MARS data mining algorithm. Grzesiak et al (2010); In his study on MARS analysis, he identified cows with artificial insemination difficulty using statistical and machine learning methods (classification functions, logistic regression, artificial neural networks and MARS). He also showed that the best results were obtained with artificial neural networks (ANN) and MARS methods in his study and stated that ANN and MARS gave more accurate results compared to other statistical methods in the detection of cows with artificial insemination differences with the test set. In agricultural sciences, t-test, one-way ANOVA, two-way ANOVA, multiple linear regression analysis is widely used (Efe et al. 2000; Agaoglu et al. 2007). In addition, more complex approaches, namely data mining, have recently been adopted (Grzesiak et al 2010; Aksoy et al. 2018; Akin et al. 2020). Although there are many studies on the carcass yield, no studies have been found investigating the use of the MARS algorithm in connection with the carcass yield characteristics. Therefore, no further discussion could not be made on the subject.

4. Conclusions

In the research, in order to estimate the carcass yield, MARS prediction models with first order interaction effects have been developed using the MARS algorithm. With the help of the comparison of the goodness of fit criteria of the model belonging to the data belonging to the training and test set, it was ensured that the bias was reduced by performing cross validation. It has been determined that MARS algorithms are a good determinant for the relationship between the properties used and the carcass yield. Estimating the carcass weight by data mining; It causes the determination of the animal's carcass yields in a shorter time. In this way, it is possible to have information about whether an economic animal husbandry can be done. In addition, with the development of such practices, it will be possible to access the carcass data of the animals to be obtained from slaughterhouses by looking at parameters such as race, live weight and age, and it will be possible to slaughter animals more economically. For such reasons, estimating the carcass data with the MARS method will be an important reference for the breeders.

Acknowledgements

I would like to thank the administrative and technical staff, especially the open prison director in Niğde, for their assistance in collecting the raw data used in the study.

References

Agaoglu Y S, Eyduran S P & Eyduran E (2007). Comparison of Some Pomological Characteristics of Blackberry Cultivars Growth in Ayaş Conditions. *Ankara Universitesi Tarım Bilimleri Dergisi* 13(1): 69-74. https://doi.org/10.1501/Tarimbil_0000000456

Akin M, Eyduran S P & Eyduran E (2020). Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. *Plant Cell Tiss Organ Cult* 140: 661-670. https://doi.org/10.1007/s11240-019-01763-8

Aksoy A, Erturk Y E, Eyduran E & Tariq M M (2018). 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 https://doi.org/10.17582/journal.pjz/2018.50.6.2279.2286

Ali M, Eyduran E, Tariq M M, Tirink C, Abbas F, Bajwa M A, Baloch M H, Nizamani A H, Waheed A, Awan M A, Shah S H, Ahmad Z & Jan S (2015). Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai sheep. *Pakistan Journal of Zoology* 47: 157 https://doi.org/10.17582/journal.pjz/2019.51.2.421.4319-1585

Anonymous (2020). Comparing different cattle breeds Use the 2016 across-breed EPD table. http://www.beefmagazine.com/ print/15608 (Access date 16.05.2020)

Anonymous (2021). Train/Test Split and Cross Validation in Python. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6 (Access date 01.04.2021)

Aytekin I, Eyduran E, Koksal 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 https://doi.org/10.17582/journal.pjz/2018.50.1.189.195

Canga D & Boga M (2019). Use of MARS in Livestock and an Application. *III. International Scientific and Vocational Studies Congress – Science and Health Full Text Paper Book*: 27-29 June, Nevşehir pp. 31-37

Canga D & Boga M (2020). Determination of the Effect of Some Properties on Egg Yield with Regression Analysis Met-hod Bagging Mars and R Application. *Turkish Journal of Agriculture - Food Science and Technology* 8(8): 1705-1712 https://doi.org/10.24925/turjaf.v8i8.1705-1712.3468

Canga D, Yavuz E & Efe E (2019). Use of MARS Data Mining Algorithm for Egg Weight Estimation: *The International Congress on Domestic Animal Breeding Genetics and Husbandry-19 Full Text Paper Book*; 11-13 September, Prague, Czechia, pp.127

Celik S & Boydak E (2020). Description of The Relationships Between Different Plant Characterıstıcs in Soybean Using Multivariate Adaptive Regression Splines Mars Algorıthm. *Journal of Animal And Plant Science*s 30(2): 431–441 https://doi.org/10.36899/japs.2020.2.0037

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 https://doi.org/10.17582/journal.pjz/2018.50.2.575.583

Celik S, Eyduran E, Tatliyer A, Karadas K & Kara MK (2020) Comparing Predictive Performances of Some Nonlinear Functions and Multivariate Adaptive Regression Splines (MARS) for Describing the Growth of Daera Dın Panah (DDP) Goat in Pakistan. *Pakistan Journal Zoology* 50(3): 1187-1190 https://doi.org/10.17582/journal.pjz/2018.50.3.sc2

Chou S M, Lee T S, Shao Y E & Chen I F (2004). Mining the Breast Cancer Pattern Using Artificial Neural Networks and Multivariate Adaptive Regression Splines. *Expert Systems with Applications* 27:133-142 https://doi.org/10.1016/j.eswa.2003.12.013

Craven P & Wahba G (1979). Estimating the Correct Degree of Smoothing by The Method of Generalized Cross-validation. *Numerische Mathematik* 31: 377-403 https://doi.org/10.1007/bf01404567

Deconinck E, Xu Q S, Put R, Coomans D, Massart D L & Vander H Y (2005). Prediction of Gastro-intestinal Absorption Using Multivariate Adaptive Regression Splines. *Journal of Pharmaceutical and Biomedical Analysis* 39: 1021–1030 https://doi.org/10.1016/j.jpba.2005.05.034

Devijver P A & Kittler J (1982). Pattern Recognition: A Statistical Approach. London, GB: Prentice-Hall. ISBN 0-13-654236-0

Duru S & Sak H (2017). Fattening Performance and Carcass Characteristics of Simmental, Aberdeen Angus, Hereford, Limousine and Charolais Cattle Breeds in Turkey. *Turkish Journal of Agriculture-Food Science and Technology* 5(11):1383-1388 https://doi.org/10.24925/turjaf.v5i11.1383-1388.1485

Efe E, Bek Y & Şahin M (2000). Statistical Methods with Solutions in SPS II. Rectorate of Kahramanmaraş Sütçü İmam University. Publication No: 73, Textbooks Publication No: 9, Turkey: Kahramanmaraş (in Turkish). https://doi.org/10.5152/kd.2019.44

Eyduran E, Akkus O, Kara M K, Tırınk C & Tarıq M M (2017a). 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, Nevşehir, Turkey pp.405

Eyduran E, Tirink C, Karahan AE, Türkoğlu M & Tariq M M (2017b). Comparison of Predictive Performances of MARS and CART Algorithms through R Software. *International Conference on Computational and Statistical Methods in Applied Sciences*, Samsun Turkey, pp.181

Eyduran E, M. Akin M & Eyduran S P (2019). Application of multivariate adaptive regression splines in agricultural sciences through R software. Nobel Academic Publishing pp.112. ISBN: 978-605-2149-81-2, Ankara, Turkey.

Eyduran E, Sevgenler H, Akin M & Eyduran SP (2018). Usage multivariate adaptive regression splines for predicting continuous responses. Animal and Plant Sciences. *International Agricultural Science Congress Full Text Paper Book*. 9-12 May, Van, Turkey.

Eyduran E & Duman H (2020). Statistical Applications with R Program (R Programla Dilinde İstatistik Uygulamaları). DOI: 10.13140/RG.2.2.29204.45442

Eyduran E & Gulbe A (2020). ehaGoF: Calculates Goodness of Fit Statistics. R package version 0.1.0 2019. https://CRAN.R-project.org/package=ehaGoF

Friedman J H (1991). Multivariate Adaptive Regression Splines. Annls. Stat. 19: 1-141. https://doi. org/10.1214/aos/1176347963

Galiç A & Takma Ç (2019). Estimates of Genetic Parameters for Body Weight in Turkish Holstein Bulls using Random Regression Model. *Journal of Agricultural Sciences* 25(3): 328-333. DOI: https://doi.org/10.15832/ankutbd.417422

Geisser S (1993). Predictive Inference. New York, NY: Chapman and Hall. ISBN 978-0-412-03471-8 https://doi.org/10.1016/j.compag.2010.09.001

Grzesiak W, Zaborski D, Sablik P, Żukiewicz A, Dybus A & Szatkowska I (2010). Detection of cows with insemination problems using selected classification models. *Comput. Electron. Agric* 74: 265-273.

Hastie T, Tibshirani R & Friedman J (2001). The Elements of Statistical Learning: Date mining, inference, and prediction. Springer. New York

Kayri M (2010). The Analysis of Internet Addiction Scale Using Multivariate Adaptive Regression Splines. *Iranian J Publ Health* 39: 51-63

Kibet C E (2012). A Multıvarıate Adaptıve Regressıon Splınes Approach to Predıct the Treatment Outcomes of Tuberculosıs Patıents in Kenya, PhD Thesis, Science in Biometry to The University of Nairobi, Kenya

Kohavi R (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* San Mateo, CA: Morgan Kaufmann 2(12): 1137-1143. CiteSeerX 10.1.1.48.529

Kor A & Ertuğrul M (2000). Estimation Methods of Carcass Composition in Living Animals. *Animal Production* 41: 91-10 (in Turkish with an abstract in English)

Kumlu S (2000). Livestock Organizations. Turkey Cattle Breeders Central Association Publications. Publication No: 2, Ankara (in Turkish)

Milborrow S (2018). Milborrow. Derived from mda: mars by T. Hastie and R. Tibshirani. İnternet url: https://CRAN.R-project.org/package=earth (10.10.2018)

Milborrow S (2011). Derived from MDA: MARS by T. Hastie and Tibshirani earth: Multivariate adaptive regression splines, R package.

Mukhopadhyay A & Iqbal A (2000). Prediction of mechanical property of steel strips using multivariate adaptive regression splines. *Journal of Applied Statistics* 36(1): 1-9 https://doi.org/10.1080/02664760802193252

Oğuz A (2014). Investigation of Multivariable Adaptive Regression Chains and an Application Master Thesis, Graduate School of Science, Erzincan University, Erzincan-Turkey (in Turkish) https://doi.org/10.18185/erzifbed.513981

Orhan H, Teke E Ç & Karcı Z (2018). Application of Multivariate Adaptive Regression Splines (MARS) for Modeling the Lactation Curves. *Journal of Agriculture and Nature* 21(3): 363-373. https://doi.org/10.18016/ksudobil.334237

Put R, Xu Q S, Massart D L & Vander H (2004). Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. *Journal of Chromatography A*. 1055: 11-19 https://doi.org/10.1016/j.chroma.2004.07.112

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

Ripley B D (1996) Pattern Recognition and Neural Networks, Cambridge: Cambridge University Press, p. 354

Salford Systems (2001). MARSTM User Guide. Salford Systems, San Diego, C A, USA.

Seker İ, Köseman A, Seker P & Baykalır Y (2017). Carcass Grading System Used in the United States for Beef Carcass Quality Evaluation. 15(2):192-203 https://doi.org/10.24323/akademik-gida.333676

Sengul T, Celik S, Eyduran E & Iqbal F (2020). Predicting egg production in Chukar partridges using nonlinear models and multivariate adaptive regression splines (MARS) algorithm. *European Poultry Science*

Sevgenler H (2019). Comparison of data mining algorithms (Cart, Chaid and Mars) used to determine the effects of some characteristics on body weight in goats. Master Thesis, Institute of Science, Iğdır University, Iğdır- Turkey (in Turkish)

Sevimli Y (2009). An application of multivariate adaptive regression splines to a split-mouth study. Master Thesis, Institute of Science, Marmara University, Istanbul-Turkey (in Turkish) https://doi.org/10.18016/ksudobil.334237

Tunay K B (2001). Estimation of the income velocity of money in Turkey by the MARS Method, *ODTÜ Journal of Development* 28(3-4): 431-454 (in Turkish with an abstract in English)

Xu Q S, Daeyaert F, Lewi P J & Massart D L (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 https://doi.org/10.1016/j.chemolab.2005.07.005

Yıldız G (2020). Ankara Üniversitesi, Veteriner Fakültesi, Hayvan Besleme ve Beslenme HastaliklariAnabilimDalı.https://acikders.ankara.edu.tr/pluginfile.php/46643/mod_resource/content/0/BESI-SIGIRLARININ-BESLENMESI-GULTEKIN-YILDIZ. pdf https://doi.org/10.1501/vetfak_0000002350

Zaborski D, Ali M, Eyduran E, Grzesiak W, Tariq MM, Abbas F, Waheed A & Tırınk C (2019). Prediction of Selected Reproductive Traits of Indigenous Harnai Sheep under the Farm Management System via various Data Mining Algorithms. *Pakistan Journal Zoology* 51(2):421-431 https://doi.org/10.17582/journal.pjz/2019.51.2.421.431

Zahrádková R, Bartoň L, Bureš D, Teslík V & Kudrna V (2010). Comparison of growth performance and slaughter characteristics of Limousine and Charolais heifers. *Archiv Tierzucht* 53(5): 520-528 https://doi.org/10.5194/aab-53-520-2010

Zhang W & Goh A T (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers* 7(1): 45-52 https://doi.org/10.1016/j.gsf.2014.10.003