Revista Brasileira de Zootecnia
Brazilian Journal of Animal Science

ISSN 1806-9290
www.sbz.org.br

R. Bras. Zootec., 46(11):863-872, 2017

Comparison of predictive performance of data mining algorithms in predicting body weight in Mengali rams of Pakistan

Senol Celik[1*], Ecevit Eyduran[2], Koksal Karadas[3], Mohammad Masood Tariq[4]

[1] Bingol University, Agricultural Faculty, Department of Animal Science, Biometry Genetics Unit, Bingol, Turkey.
[2] Igdir University, Agricultural Faculty, Department of Animal Science, Biometry Genetics Unit, Igdir, Turkey.
[3] Igdir University, Agricultural Faculty, Department of Agricultural Economics, Igdir, Turkey.
[4] University of Balochistan, Center for Advanced Studies in Vaccinology and Biotechnology, Quetta, Pakistan.

ABSTRACT - The present study aimed at comparing predictive performance of some data mining algorithms (CART, CHAID, Exhaustive CHAID, MARS, MLP, and RBF) in biometrical data of Mengali rams. To compare the predictive capability of the algorithms, the biometrical data regarding body (body length, withers height, and heart girth) and testicular (testicular length, scrotal length, and scrotal circumference) measurements of Mengali rams in predicting live body weight were evaluated by most goodness of fit criteria. In addition, age was considered as a continuous independent variable. In this context, MARS data mining algorithm was used for the first time to predict body weight in two forms, without (MARS_1) and with interaction (MARS_2) terms. The superiority order in the predictive accuracy of the algorithms was found as CART > CHAID ≈ Exhaustive CHAID > MARS_2 > MARS_1 > RBF > MLP. Moreover, all tested algorithms provided a strong predictive accuracy for estimating body weight. However, MARS is the only algorithm that generated a prediction equation for body weight. Therefore, it is hoped that the available results might present a valuable contribution in terms of predicting body weight and describing the relationship between the body weight and body and testicular measurements in revealing breed standards and the conservation of indigenous gene sources for Mengali sheep breeding. Therefore, it will be possible to perform more profitable and productive sheep production. Use of data mining algorithms is useful for revealing the relationship between body weight and testicular traits in describing breed standards of Mengali sheep.

Key Words: ANN, artificial intelligence, data mining, decision tree, MARS algorithm

Introduction

Sheep is a small ruminant, utilized in rural development and in most civilizations with multiple purposes (Karadas et al., 2017). In sheep breeding programs, it is fundamental to affirm the relationship between target traits and the connected traits due to the fact that the target traits are correlated genetically with the related traits. For example, body weight, as a target characteristic, is predicted through morphological (body length, withers height, punch girth, hearth girth, and chest depth etc.) and testicular (testicular length, scrotal circumference, and scrotal length) characteristics, in which weighing bridges are unavailable (Birteeb and Ozoje, 2012). Many researchers are willing to predict live body weight from its connected body characteristics in breed characterization of sheep breeds and in the recognition of the ideal drug dose, feed amount, and price of the sheep (Khan et al., 2014; Eyduran et al., 2016). The prediction plays a fundamental role in flock management and, hence, to obtain more profit in a flock (Birteeb and Ozoje, 2012).

Practical approaches to predict live body weight in sheep use a variety of regression analysis techniques, viz., simple regression, multiple regression (Birteeb and Ozoje, 2012), ridge regression and application of factor (Eyduran et al., 2009), and principle component analyses in multiple linear regressions (Eyduran et al., 2016). More flexible in terms of the classical assumptions, CART (classification and regression tree) and CHAID (chi-square automatic interaction detection) data mining approaches have recently been implemented to perfectly indicate morphological traits connected with the live body weight in sheep breeding studies (Yakubu, 2012; Ali et al., 2015). Additionally, ANN (artificial neural network) types such as MLP (multilayer perceptron) and RBF (radial basis function) have seldom been applied in the prediction of body weight in sheep. On the other hand, MARS (multivariate adaptive regression spline) has not been fit so far about the prediction of body

Received: November 11, 2016
Accepted: June 8, 2017
***Corresponding author:** senolcelik@bingol.edu.tr
http://dx.doi.org/10.1590/S1806-92902017001100005

How to cite: Celik, S.; Eyduran, E.; Karadas, K. and Tariq, M. M. 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.

weight in sheep husbandry, although Grzesiak and Zaborski (2012) applied CART, CHAID, and some ANN algorithms on animal data for lactation milk yield and dystocia in dairy cattle. When the above-mentioned advantages are considered in the scope of sheep breeding, there is growing interest on data mining and artificial neural networks with respect to the weight prediction from morphological traits.

Karadas et al. (2017) tested predictive capabilities of CHAID, Exhaustive CHAID, CART, and MLP algorithms in lactation milk yield prediction of Akkaraman sheep. There are many studies on CHAID (Dogan, 2003; Bakir et al., 2010; Eyduran et al., 2013a), CART (Topal et al., 2010), and ANN (Grzesiak et al., 2003; Grzesiak et al., 2006) algorithms in the prediction of milk yield in dairy cattle and dairy goat (Eyduran et al., 2013b). Eyduran et al. (2008) applied CHAID algorithm for predicting birth weight by means of non-genetic factors in Norduz and Karakas sheep. In the prediction of live body weight from withers height, body length, and chest circumference for Mengali, Balochi, Harnai, Beverigh, and Rakshani sheep breeds at yearling age, Mohammad et al. (2012) used CHAID algorithm. Yakubu (2012) preferred to use CART algorithm for the body weight prediction of Uda rams from their body measurements. Although there is some earlier research on the body weight prediction from biometrical characteristics in sheep via Exhaustive CHAID (Khan et al., 2014) and CART, CHAID, and ANN algorithms (Ali et al., 2015), such studies are still rare; however, MARS data mining algorithm has not been handled thus far in the prediction of the body weight in farm animals such as sheep, goat, and cattle. The success in sheep breeding and selection programs depends on high genetic correlations between the target trait and the biometrical traits within the scope of indirect selection criteria. In this respect, we aimed to compare predictive performances of several data mining algorithms with one another in the prediction of live body weight by main predictors such as morphological and testicular traits in sheep.

Material and Methods

Previous data published by Tariq et al. (2012) were employed to compare data mining algorithms, i.e., CART, CHAID, Exhaustive CHAID, MARS, MLP, and RBF. The data were gathered from 107 Mengali male lambs born in Quetta, Pakistan. The body weight was predicted by body length (BL), withers height (WH), and heart girth (HG), testicular length (TL), scrotal length (SL), and scrotal circumference (SC) at varying ages ranging from 12 to 48 months. The animals were exposed to the same feeding system. Descriptive statistics of the input and output variables are given in Table 1.

CART, CHAID, and Exhaustive CHAID are visual algorithms that create regression tree structures and analyze qualitative and quantitative data simultaneously. CHAID (Kass, 1980) and Exhaustive CHAID (Biggs et al., 1991) three-stage-data mining algorithms (merging, partitioning, and stopping) are tree-based algorithms that recursively use multi-way splitting to form homogenous subsets on the basis of Bonferroni adjustment until the differences between the actual and the predicted values in output variable are minimal (Orhan et al., 2016; Akin et al., 2016; Akin et al., 2017; Eyduran et al., 2016). A quantitative input variable in CHAID algorithms is converted into an ordinal variable (Orhan et al., 2016).

CART, developed by Breiman et al. (1984), uses a recursively binary-splitting that produces homogenous subsets in regression tree structure until finding the least differences (Karadas et al., 2017). Exhaustive CHAID has an exhaustive procedure in merging stage to merge any similar pair until merely a single pair remains. However, for both algorithms, splitting and stopping stages are the same.

MARS, developed as a-nonparametric regression technique by Jerome Friedman in 1991, is a data mining algorithm specifying piecewise basis functions for describing an output variable and a set of input variables and it automatically selects knot locations. Prediction equation of the MARS algorithm can be written as follows:

$$f_M(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x),$$

in which β_0 and β_m are the basis function parameters of the MARS algorithm used on the basis of the least squares criterion. The spline basis function $B_m(x)$ can be implemented as:

$$B_m(x) = \prod_{k=1}^{k_m} [s_{km}(x_{v(k,m)} - t_{k,m})],$$

Table 1 - Description of the input and output variables (means and standard deviations)

	N	Mean	Standard deviation
Body weight	107	50.411	13.200
Age	107	26.070	10.139
Testicular length	107	13.144	2.757
Scrotal circumference	107	24.280	5.995
Scrotal length	107	15.490	2.112
Withers height	107	71.145	10.036
Body length	107	83.830	14.341
Heart girth	107	86.383	12.198

in which *k*m is the number of knots, s_{km} takes either 1 or −1 and presents the right/left regions of the related step function, v(k,m) is the label of the input variable, and $t_{k,m}$ is the knot location (Friedman, 1991).

The generalized cross validation (GCV) is approved to eliminate the redundant basis functions (Craven and Wahba, 1979):

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_1 - \hat{f}(x_i)]^2}{\left[1 - \frac{c(B)}{N}\right]^2},$$

in which N is the number of data and *C(B)* is a complexity penalty increasing with the number of basis function in the model, which is expressed as:

$$C(B) = (B+1) + dB,$$

in which d is a penalty for each basis function entered into the model and B is the number of the basis functions (Friedman, 1991).

In the study, MARS algorithm was built in two forms: MARS_1 (no interaction) and MARS_2 (with interaction). In the MARS algorithm, we used 100 as the initial number of basis functions considered in the model construction process.

ANN is a processing system that implements activities that are similar to those of the human brain. MLP typically comprises of three connected feed-forward layers of neurons (Rumelhart and McClelland, 1986). The hyperbolic tangent function and the linear activation function are employed for the hidden and output layers and they are illustrated by means of the following functions:

Hyperbolic tangent:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

Linear:

$$f(x) = x,$$

in which x represents the weighted sum of inputs to the neuron and $f(x)$ denotes the outputs obtained from the neuron.

MLP, consisting of the input, hidden, and output layers, is a type of ANN. In MLP, each of the layers is linked to the earlier layer. In this study, we specified one hidden layer, hyperbolic tangent, as an activation function and an identity activation function for output layer when obtaining the least differences in goodness of fit criteria between training and testing sets.

As a precious alternative to MLP, RBF neural network is another type of ANN employing radial basis functions on the scope of activation functions and consists of three layers: input, hidden, and a linear output layer. The hidden layer of RBF, which is a feed forward network, has a nonlinear activation function. Radial basis function activates a function using a network of Gaussian functions in the hidden layer and functions regarding linear activation in the output layer. The Gaussian function is expressed as follows:

$$f(x) = e^{-x^2/2\sigma^2},$$

in which x characterizes the weighted sum of inputs, σ is the sphere of effect or the width of the basis function, and $f(x)$ is defined as the corresponding output from neurons [see Kaewtapee et al. (2011) for more detailed information].

In IBM SPSS 23 statistical software, the number of tree depth is 3 in CART and 5 in CHAID algorithms, by default. The least animal numbers for parent and child subsets were taken as 4 and 2 for getting the highest predictive ability of the algorithms. As in MLP, RBF was also applied at training (80%) and testing (20%) sets in the study. Since we had small sample size of 107 rams, we obtained better results for the 10-fold cross-validation compared with results of training and testing sets. In this regard, we prepared the 10-fold cross-validation in the case of MARS and regression trees. A V-fold cross-validation is employed to assess the overall accuracy of the data mining algorithms (Kovalchuk et al., 2017). As a specification rule of IBM SPSS program, we used the cross-validation using a training (80%) and testing (20%) set. Epoch number was set at 1000 for MLP and RBF algorithms, also known as two types of ANN. At least 30 different networks were considered for the algorithms during the search procedure. The criterion we used was minimal RMS. To comparatively test the predictive performance of CART, CHAID, Exhaustive CHAID, and MARS in ten cross-validations, the following goodness of fit criteria were calculated (Willmott and Matsuura, 2005; Takma et al., 2012):

Pearson correlation coefficient (r) between the actual and predicted body weight (BW) values;

Akaike information criterion (AIC), calculated as:

$$AIC = nln\left[\frac{1}{n}\left(y_i - y_{ip}\right)^2\right] + 2k, \quad if\ \frac{n}{k} > 40$$

$$\text{or } AIC_c = nln\left[\frac{1}{n}\left(y_i - y_{ip}\right)^2\right] + 2k + \frac{2k(k+1)}{n-k-1}, \text{ otherwise;}$$

Root-mean-square error (RMSE) given by the following formula:

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

Mean error (ME) given by the following equation:

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

Mean absolute deviation (MAD):

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

Standard deviation ratio (SD_{ratio}):

$$SD_{ratio} = \frac{s_m}{s_d}$$

Global relative approximation error (RAE):

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

Mean absolute percentage error (MAPE):

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

Coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{ip})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Adjusted coefficient of determination:

$$Adj.R^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - y_{ip})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

in which n is the number of cases in a set, k is the number of model parameters, y_i is the actual (observed) value of an output variable (BW), y_{ip} is the predicted value of an output variable (BW), s_m is the standard deviation of model errors, and s_d is the standard deviation of an output variable (BW).

Adjusted coefficient of determination as a goodness of fit criterion can be used for algorithms having different input numbers. In this regard, we estimated that it makes adjustment for this differentness. In MARS algorithm, k is described as number of terms.

Statistical evaluations on CART, CHAID, Exhaustive CHAID, MLP, and RBF algorithms were carried out using IBM SPSS 23, but MARS algorithm was specified by STATISTICA program (8.0 trial version).

Results

We aimed to confirm linear body measurements strongly connected with live body weight in Mengali ram breeding with the support of robust statistical techniques. Because of this, some data mining algorithms allowed testing of the predictive accuracy of live body weight. All the algorithms produced very fit results in body weight prediction (Table 2). The superiority order in the predictive accuracy of the algorithms was found as CART > CHAID ≈ Exhaustive CHAID > MARS_2 > MARS_1 > RBF > MLP.

The regression tree structure for CHAID algorithm had AGE, TL, HG, and WH, which were found to be significant input variables in the live body weight prediction of Mengali rams. All the rams were divided into six subgroups (Nodes 1-6) according to age (Figure 1). The weight order among Nodes 1-6 was found to be Node 1 < Node 2 < Node 3 < Node 4 < Node 5 < Node 6 (Adjusted P = 0.000), because of significant differences in BW.

Node 1 was a subgroup of Mengali rams whose age was 12 months or younger (BW = 32.864 kg). Node 2 was a subgroup of Mengali rams with 12 < AGE ≤ 19 months among all the rams (BW = 39.250 kg). The subgroup of Mengali rams with an age of 19 < AGE ≤ 26 months was entered into Node 3 in the decision tree construction of CHAID algorithm (BW = 47.353 kg). The subgroup of those having 26 < AGE ≤ 35.5 months was included in Node 4 through CHAID algorithm (BW = 55.933 kg). Node 5 consisted of rams falling into 35.5 < AGE ≤ 36 subgroup (BW = 62.182 kg). Node 6 was the subgroup of rams older than 36 months (BW = 74.000 kg). The rams incorporated into Node 5 were split into smaller subgroups (Nodes 7-8) according to TL trait. The rams (TL ≤ 14.5 and 35.5 < AGE ≤ 36) in Node 7 were lighter in weight compared with those (with TL > 14.5 and 35.5 < AGE ≤ 36) in Node 8 (Adjusted P = 0.001; 54.500 vs. 63.889 kg).

Node 6 (ramos older than 36 months) was divided into two smaller subgroups (Nodes 9 and 10) in terms of HG

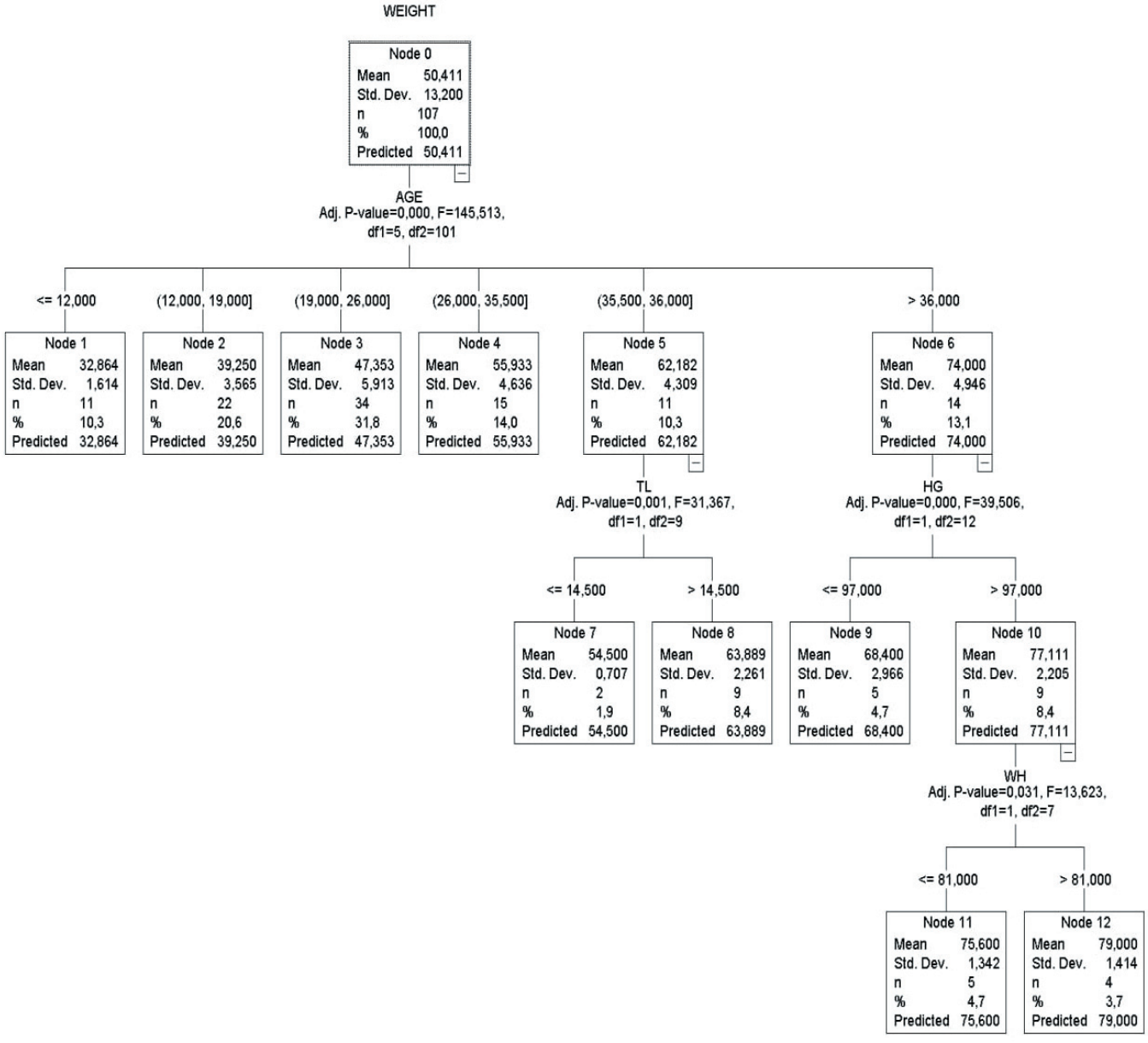
Table 2 - Predictive performance of MARS, CART, CHAID, and ANN types

	MARS_1 (no interaction)	MARS_2 (with interaction)	CHAID	Exhaustive CHAID	CART	MLP	RBF	Average
AIC	333.397	329.862	310.464	312.338	290.835	360.809	340.899	325.515
AICc	334.866	331.331	310.856	312.571	291.429	362.279	342.369	326.500
RMSE	4.527	4.450	4.144	4.138	3.704	4.995	4.551	4.358
ME	2.39×10^{-14}	6.18×10^{-15}	0.002	0.002	−0.001	−0.100	0.299	0.029
MAD	3.373	3.355	2.991	2.980	2.664	3.746	3.545	3.236
SD ratio	0.345	0.339	0.315	0.315	0.282	0.380	0.346	0.332
RAE	0.087	0.085	0.080	0.079	0.071	0.096	0.087	0.084
MAPE	7.142	7.145	6.486	6.471	5.797	7.856	7.572	6.924
R^2	0.881	0.885	0.901	0.901	0.920	0.857	0.880	0.889
Adjusted R^2	0.877	0.880	0.899	0.899	0.917	0.847	0.871	0.884
r	0.939	0.941	0.949	0.949	0.959	0.926	0.938	0.943

MARS - multivariate adaptive regression spline; CART - classification and regression tree; CHAID - chi-square automatic interaction detection; ANN - artificial neural network; MLP - multilayer perceptron; RBF - radial basis function; AIC - Akaike information criterion; AICc - corrected Akaike information criterion; RMSE - root-mean-square error; ME - mean error; MAD - mean absolute deviation; SD - standard deviation; RAE - relative approximation error; MAPE - mean absolute percentage error; R^2 - coefficient of determination; r - correlation coefficient.

(Adjusted P = 0.000). Mean body weight of rams with HG ≤ 97 cm in Node 9 was estimated as 68.400 kg. Also, mean body weight of rams with HG>97 cm in Node 10 was found as 71.111 kg. Node 10 was branched by WH into smaller subgroups (Nodes 11 and 12) (Adjusted P = 0.031; 75.600 vs. 79.000 kg in BW).

The decision tree structure of the Exhaustive CHAID (Figure 2) contained significant input variables (AGE, TL, and BL) in predicting the live body weight of Mengali rams. All the Mengali lambs were divided into smaller subgroups (Nodes 1 and 6) with respect to age (input) variable as also formed in tree-based structure of the CHAID algorithm when trees of the CHAID algorithms were analyzed in the first depth. Node 5 (a subgroup of the rams with 35.5< AGE ≤ 36 months) was divided into smaller subgroups (Nodes 7 and 8) according to TL trait. The rams (TL ≤ 14.5 and 35.5 < AGE ≤ 36) in Node 7 were found lighter in BW than those (with TL > 14.5 and 35.5 < AGE ≤ 36) in Node 8 (Adjusted P = 0.001; 54.500 vs. 63.889 kg). As a subgroup of rams with AGE > 36 months, Node 6 (74.000 kg) was exposed to a division according to TL trait and partitioned into three smaller subgroups (Nodes 9-11), whose live body weight averages were estimated at 73.000, 67.750, and 77.375 kg, respectively (Adjusted P = 0.000). Node 9

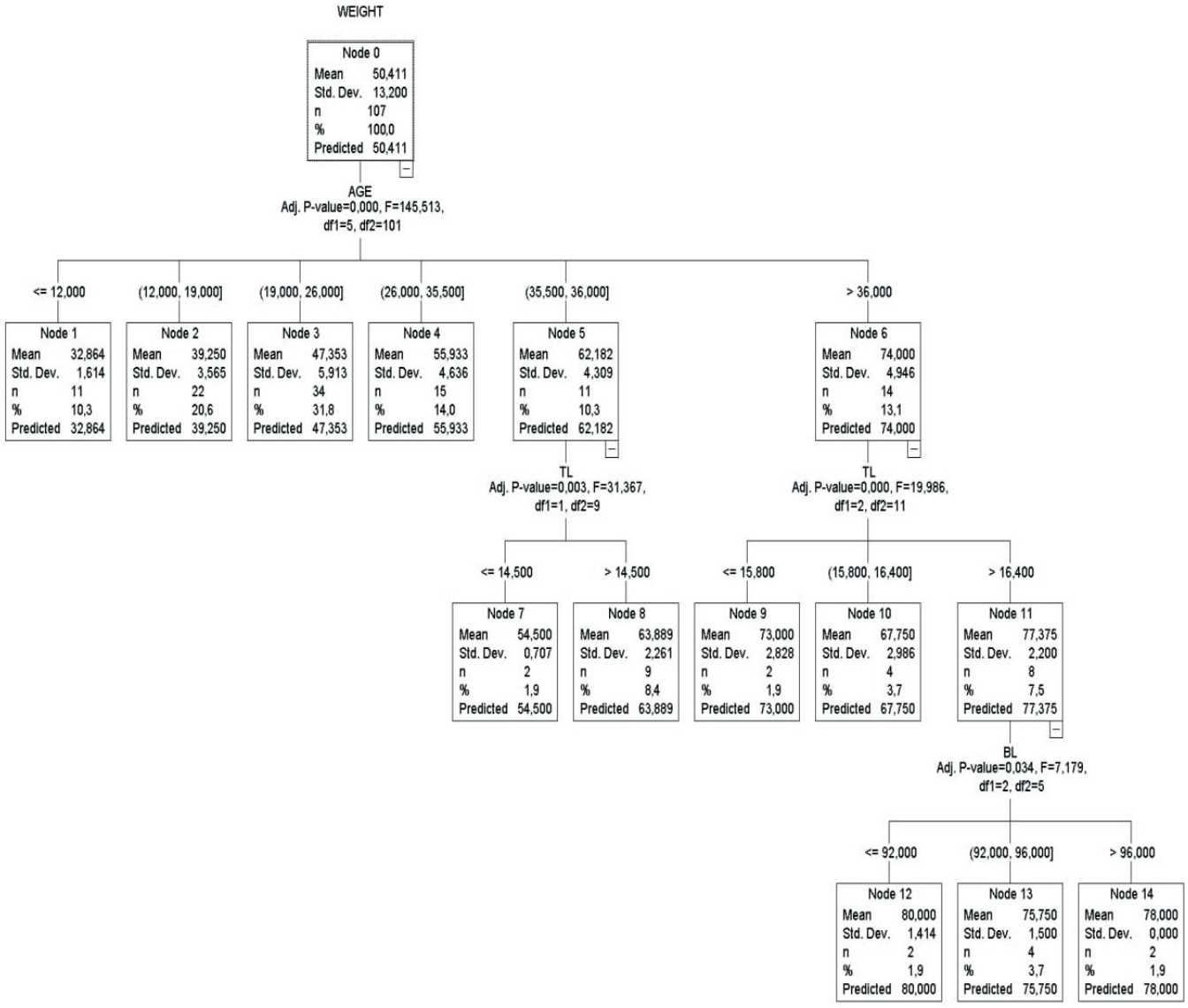


TL - testicular length; HG - heart girth; WH - withers height; CHAID - chi-square automatic interaction detection.

Figure 1 - Regression tree structure of CHAID algorithm.

was characterized through the subgroup of rams with TL ≤ 15.800 and AGE > 36 months. Node 10 was described as the subgroup of rams with 15.800 < TL ≤ 16.400 and AGE > 36 months, and rams having TL > 16.400 and AGE > 36 months were grouped into Node 11, which were entered into three smaller groups, Node 12 (80.000 kg), Node 13 (75.750 kg), and Node 14 (78.000 kg), in third tree depth respectively.

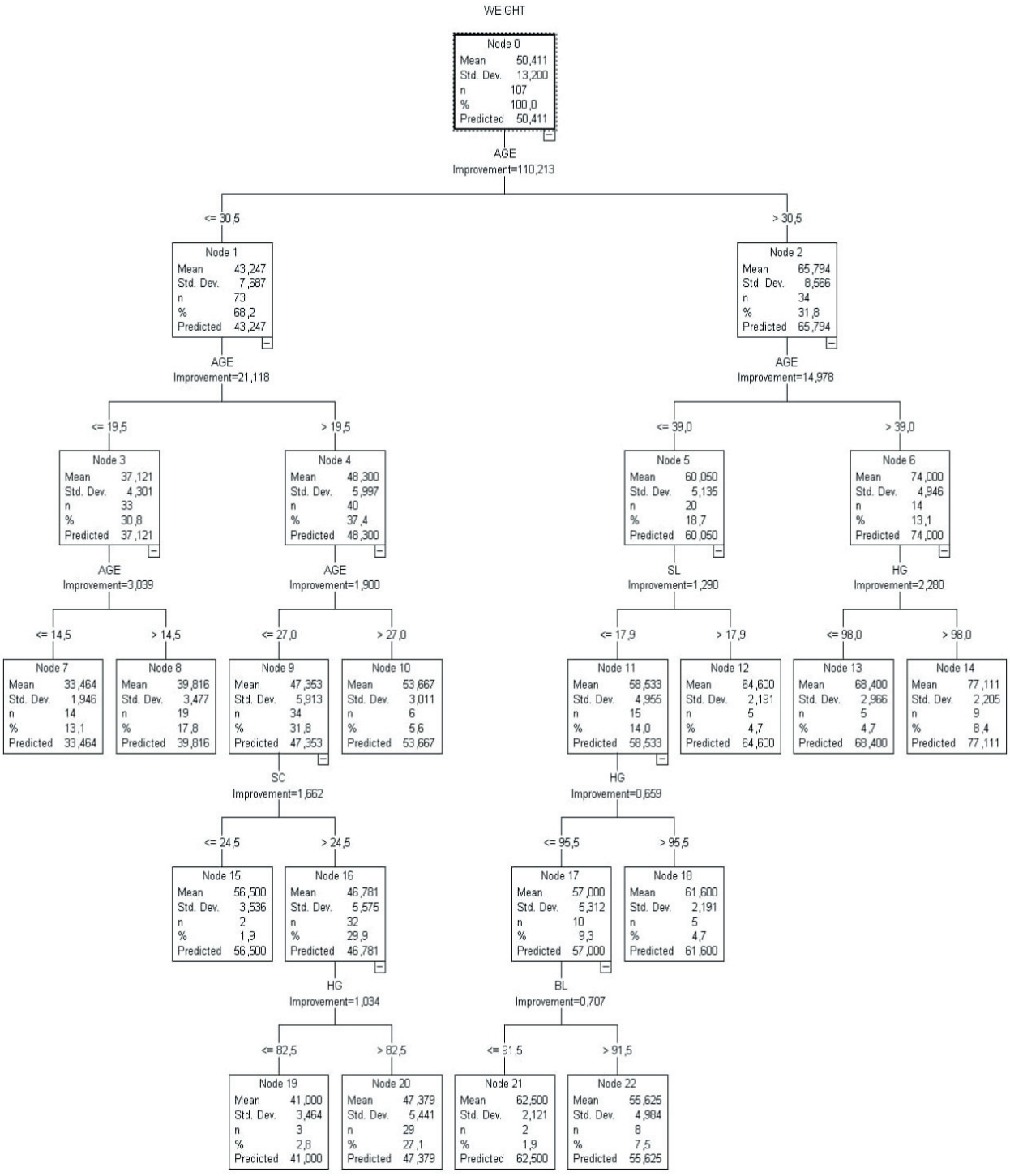
In the first tree depth, Node 0 was divided into two subgroups, Node 1 (43.247 kg – the subgroup of rams with AGE ≤ 30.5 months) and Node 2 (65.794 kg – the subgroup of rams with AGE > 30.5 months) according to AGE (Figure 3). In the second tree depth, Nodes 1 and 2 were further divided by means of AGE input variable and Nodes 3 and 4 (37.121 – the subgroup of rams with AGE ≤ 19.5 months vs. 48.300 kg – the subgroup of rams with 19.5 < AGE ≤ 30.5 months) along with Nodes 5 and 6 (60.050 kg – the subgroup of rams with 30.5 < AGE ≤ 39 months vs. 74.000 kg – the subgroup of rams with AGE > 39 months) were generated. Nodes 3 and 4 were partitioned into Nodes 7 (the subgroup of rams with AGE ≤ 14.5 months) and 8 (33.464 – the subgroup of rams with AGE ≤ 14.5 months vs. 39.816 kg – the subgroup of rams with 14.5 < AGE ≤ 19.5 months) and Nodes 9 and 10 (47.353 kg – the subgroup of rams with 19.5<AGE ≤ 27 months vs. 53.667 kg – the subgroup of rams with AGE > 27 months) again. Node 9 was split into Nodes 15 and 16 (56.500 vs. 46.781 kg) with respect to SC input trait as discriminator.



TL - testicular length; BL - body length; CHAID - chi-square automatic interaction detection.

Figure 2 - Regression tree structure of Exhaustive CHAID algorithm.

Node 15 was characterized with rams having SC ≤ 24.5 and 19.5 < AGE ≤ 27.5 months, whereas Node 16 was obtained from rams having SC > 24.5 and 19.5 < AGE ≤ 27.5 months and discriminated into two subgroups (Nodes 19 and 20) as a result of the variability in HG (41.000 vs. 47.379 kg). The rams having SC > 24.5, HG ≤ 82.5, and 19.5 < AGE ≤ 27.5 months were included in Node 19. Node 20 comprised of rams having SC > 24.5, HG > 82.5, and 19.5 < AGE ≤ 27.5 months. Node 5 was separated by SL into two subgroups, Nodes 11 and 12, whose means were 58.533 vs. 64.600 kg in the subgroups SL ≤ 17.9 in 30.5 < AGE ≤ 39 months and SL > 17.9 in 30.5 < AGE ≤ 39 months. Node 11 was



SL - scrotal length; HG - hearth girth; SC - scrotal circumference; BL - body length; CART - classification and regression tree.

Figure 3 - Regression decision tree diagram for body weight in sheep using CART algorithm.

discriminated into two subgroups (Nodes 17 and 18) due to the variability in HG for the rams with SL ≤ 17.9 in 30.5 < AGE ≤ 39 months (57.000 vs. 61.600 kg). The rams in Node 17 comprised of two subgroups (Nodes 21 and 22) according to BL linear body measurement (62.500 vs. 55.625 kg). Node 6 was exposed to a division in HG and it produced Nodes 13 and 14 (68.400 vs. 77.111 kg). Node 13 was the subgroup of rams having HG ≤ 98 and AGE > 39 months. In addition, Node 14 was generated from the rams having HG > 98 and AGE > 39 months.

The morphological data was exposed to MLP and RBF as ANN types in the live weight prediction at training-testing set proportion of 80:20. The importance order of independent variables for MLP algorithm was AGE (100%) > SD (46.3%) > SC (43.2%) > HG (37.1%) > BL (14%) > SL (12.3%) > TL (9.6%) > WH (6.8%). The significance order for RBF algorithm was AGE (100%) > WH (68.3%) > TL (63%) > SD (54.7%) > SD (54.4%) > SL (43.3%) > BL (40.6%) > HG (32.1). Prediction equation of MARS data mining algorithm when the degree of interaction was used as 1 (no interaction) was written for MARS_1 form as:

$$BW = 50.13968 + 1.28971*\max(0, AGE-30) - 0.89417*\max(0, 30-AGE) + 0.23338*\max(0, HG-77)$$

The prediction equation obtained by MARS had approximately 90% R^2 with the support of two significant input variables, AGE and HG. When AGE > 30 months and HG > 77 cm, an increase in ram weight is expected. For example, if a ram has AGE = 31 months and HG = 80 cm, then max (0, 31-30) = 1 and max (0, 80-77) = 3, that is, its weight prediction was BW = 50.13968 + 1.28971*max(0, 31-30) – 0.89417*max(0, 30-31) + 0.23338*max(0, 80-77) = 50.13968 + 1.28971*1 – 0.89417*0+0.23338*3 = 52.13 kg.

Prediction equation of MARS data mining algorithm when the degree of interaction was used as 2 was expressed for MARS_2 form as:

$$BW = 55.55369 + 1.40170*\max(0, AGE-30) - 1.212*\max(0, 30-AGE) - 1.80805*\max(0, TL-11)*\max(0, SL-16.5) + 2.74558*\max(0, TL-11)*\max(0, SL-17).$$

If a ram has AGE > 30 months in the above equation, an increase in live body weight is expected. Interaction of SL and TL had a significant effect on live body weight.

Discussion

One of the most important targets of the breeders in sheep breeding is to determine the relationship between body weight and the linked traits such as linear body measurements as indirect selection criteria in describing breed traits of the evaluated sheep breeds and predicting live body weight. Rams play an important role for obtaining genetic improvement in sheep breeding studies. In this respect, we aimed to predict body weight of Mengali rams from body and especially testicular measurements.

We found better goodness of fit criteria (SD ratio, R^2, and Adjusted R^2) in the predicting body weight by means of testicular and body measurements in Mengali rams for CHAID (0.31542, 0.9006, and 0.89869), Exhaustive CHAID (0.315, 0.900601, and 0.89869), CART (0.28197, 0.91968, and 0.91653), MLP (0.38011, 0.85748, and 0.84740), and RBF (0.34566, 0.87984, and 0.87135) in comparison with those recorded for Harnai sheep such as Ali et al. (2015), who estimated CART (0.403, 0.82644, and 0.82199), CHAID (0.397, 0.8377, and 0.83354), Exhaustive CHAID (0.417, 0.8421, and 0.83805), and MLP (0.4230, 0.81999, and 0.81537). The present goodness of fit findings were in agreement with the statement of Ali et al. (2015), who reported that data mining algorithm whose SD ratio is equal to the value less than 0.40 had a good predictive capability in the prediction.

Ali et al. (2015) concluded for Exhaustive CHAID algorithm that WH was found as a good and significant predictor for male lambs and for female lambs, LBE (length between ears) and FL (face length) were significant input variables in the prediction of live body weight of Harnai sheep aged six to nine months. The heaviest weight (31.733 kg) was obtained from Harnai male lambs having WH ≤ 47 cm, whereas female lambs with LBE ≤ 10 cm produced the heaviest weight (24.273 kg) for female lambs. Compared with the present R^2 estimated for CHAID algorithm, Mohammad et al. (2012) estimated a lower R^2 value of 0.72 for only CHAID algorithm in the body weight prediction of WH, CG, BL, and breed in indigenous Pakistan sheep. Yakubu (2012) recorded the lowest R^2 estimate of 0.62 for CART algorithm in the body weight estimation of Uda rams in comparison with that found by Mohammad et al. (2012), Ali et al. (2015), and the results of the present study. The variation might be ascribed to breed, gender, age, managerial and agro-climatic conditions, and statistical methods. Yakubu (2012) informed that face length and chest circumference were good predictors for CART algorithm in the live body estimation of Uda rams. However, we observed that AGE, SL, SC, HG, and BL were significant predictors for the same algorithm in live body weight prediction of Mengali rams under investigation and the heaviest live weight was provided by the rams with HG > 98 cm and AGE > 39 months. Khan et al. (2014) obtained 0.844 R^2 for Exhaustive CHAID algorithm in the prediction of the Harnai sheep from significant predictors FL, WH,

CG, and BL. The current goodness of fit results gave better fit compared with Khan et al. (2014), who obtained the heaviest body weight (49.166 kg) from the sheep with FL > 24.750 cm.

There are a limited number of studies on predicting the live body weight by using neural network algorithms. In our study, it is evident (Table 1) that RBF neural network algorithm showed better fit compared with MLP algorithm, which was in agreement with the results found by Kaewtapee et al. (2011) for predicting body weight in Cherry Valley ducks. The goodness of fit criteria results for MLP and RBF neural network algorithms in the current study were much better than those reported by Ruhil et al. (2013), who calculated roughly 0.65 SD ratio value at 6-12 months of age for the training (75%) and testing (25%) sets in female Attappady Black goats. The prediction of body weight in rabbits by means of breed, sex, HG, BL, and WH input variables was reported by Salawu et al. (2014) in the scope of ANN modeling (0.68-0.71 R^2) for the training (75%) and testing (25%) sets. The results of ANN and other data mining algorithms were superior to those estimated by Salawu et al. (2014). The training and testing set proportions might be significant source of variation affecting the prediction operations (Dongre et al., 2012).

MARS data mining algorithm has been used typically in the classification of the binary output variable in literature (Grzesiak and Zaborski, 2012). However, in our study, MARS data mining algorithm was used for the first time in the body weight prediction of the Mengali rams. When first order of interaction in MARS modeling was considered, we obtained that age and interaction of some testicular traits (TL and SL) were significant predictors in the body weight estimation of Mengali rams. The prediction equation of MARS can be an important reference in improving fertility in selection studies due to the fact that testicular traits were linked genetically with ovulation proportion in females (Bilgin et al., 2004) and related with lamb production (Karakus et al., 2010). The MARS modeling provided very high predictive performance with few input variables.

The present goodness of fit criteria for data mining and neural network algorithms were found to be highly predictive compared with those obtained by Tariq et al. (2012) using factor scores in multiple linear regression analysis used previously for the same data set in prediction of body weight of the Mengali rams. Data mining algorithms were more advantageous about the violation of the assumption regarding input variables (Mendes and Akkartal, 2009).

Conclusions

Data mining algorithms is useful for revealing the relationship between body weight and testicular traits in describing breed standards of Mengali sheep.

References

Akin, M.; Eyduran, E. and Reed, B. M. 2016. Using the CHAID data mining algorithm for tissue culture medium optimization. In Vitro Cellular & Developmental Biology - Animal 52:S66-S66. Springer, New York, NY, USA.

Akin, M.; Eyduran, E. and Reed, B. M. 2017. Use of RSM and CHAID data mining algorithm for predicting mineral nutrition of hazelnut. Plant Cell Tissue and Organ Culture 128:303-316.

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. and 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 47:1579-1585.

Bakir, G.; Keskin, S. and Mirtagioglu, H. 2010. Determination of the effective factors for 305 days milk yield by regression tree (RT) method. Journal of Animal and Veterinary Advances 9:55-59.

Biggs, D.; De Ville, B. and Suen, E. 1991. A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics 18:49-62.

Bilgin, O. C.; Emsen, E. and Davis, M. H. 2004. Comparison of non-linear models for describing the growth of scrotal circumference in Awassi male lambs. Small Ruminant Research 52:155-160.

Birteeb, P. T. and Ozoje, M. O. 2012. Prediction of live body weight from linear body measurement of West African long-legged and West African dwarf sheep in Northern Ghana. Online Journal of Animal and Feed Research 2:425-434.

Breiman, L.; Friedman, J. H.; Olshen, R. A. and Stone, C. J. 1984. Classification and regression trees. Chapman and Hall, Wadsworth Inc., New York, NY, USA.

Craven, P. and Wahba, G. 1979. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31:377-403.

Dogan, I. 2003. Investigation of the factors which are affecting the milk yield in Holstein by CHAID analysis. Ankara University Journal of Veterinary Faculty 50:65-70.

Dongre, V. B.; Gandhi, R. S.; Singh, A. and Ruhil, A. P. 2012. Comparative efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation 305-day milk yield in Sahiwal cattle. Livestock Science 147:192-197.

Eyduran, E.; Karakuş, K.; Keskin, S. and Cengiz, F. 2008. Determination of factors influencing birth weight using regression tree (RT) method. Journal of Applied Animal Research 34:109-112.

Eyduran, E.; Karakus, K.; Karakus, S. and Cengiz, F. 2009. Usage of factor scores for determining relationships among body weight and some body measurements. Bulgarian Journal of Agricultural Science 15:373-377.

Eyduran, E.; Yilmaz, I.; Kaygisiz, A. and Aktas, Z. M. 2013a. An investigation on elationship between lactation milk yield, somatic cell count and udder traits in first lactation Turkish Saanen goat using different statistical techniques. The Journal of Animal and Plant Sciences 23:956-963.

Eyduran, E.; Yilmaz, I.; Tariq, M. M. and Kaygisiz, A. 2013b. Estimation of 305-d milk yield using regression tree method in Brown Swiss cattle. The Journal of Animal and Plant Sciences 23:731-735.

Eyduran, E.; Keskin, I.; Erturk, Y. E.; Dag, B.; Tatliyer, A.; Tirink, C.; Aksahan, R. and Tariq, M. M. 2016. Prediction of Fleece weight from wool characteristics of sheep using regression tree method (Chaid Algorithm). Pakistan Journal of Zoology 48:957-960.

Friedman, J. H. 1991. Multivariate adaptive regression splines. The Annals of Statistics 19:1-141.

Grzesiak, W.; Lacroix, R.; Wójcik, J. and Bxlaszczyk, P. 2003. A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. Short Communication. Canadian Journal of Animal Science 83:307-311.

Grzesiak, W.; Blaszczyk, P. and Lacroıx, R. 2006. Methods of predicting milk yield in dairy cows-Predictive capabilities of Wood's lactation curve and artificial neural networks (ANNs). Computer and Electronics in Agriculture Journal 54:69-83.

Grzesiak, W. and Zaborski, D. 2012. Examples of the use of data mining methods in animal breeding. InTech Open Science/Open Minds, Zagreb.

Kaewtapee, C.; Khetchaturat, C. and Bunchasak, C. 2011. Comparison of growth models between artificial neural networks and nonlinear regression analysis in Cherry Valley ducks. The Journal of Applied Poultry Research 20:421-428.

Karadas, K.; Tariq, M.; Tariq, M. M. and Eyduran, E. 2017. Measuring predictive performance of data mining and artificial neural network algorithms for predicting lactation milk yield in Indigenous Akkaraman Sheep. Pakistan Journal of Zoology 49:1-7.

Karakus, K.; Eyduran, E.; Aygun, T. and Javed, K. 2010. Appropriate growth model describing some testicular characteristics in Norduz male lambs. The Journal of Animal and Plant Sciences 20:1-4.

Kass, G. V. 1980. An exploratory technique for ınvestigating large quantities of categorical data. Applied Statistics 29:119-127.

Khan, M. A.; Tariq, M. M.; Eyduran, E.; Tatliyer, A.; Rafeeq, M.; Abbas, F.; Rashid, N.; Awan, M. A. and Javed, K. 2014. Estimating body weight from several body measurements in Harnai Sheep without multicollinearity problem. The Journal of Animal and Plant Sciences 24:120-126.

Kovalchuk, I. Y.; Mukhitdinova, Z.; Turdiyev, T.; Gulnara Madiyeva, G.; Akin, M.; Eyduran, E. and Reed, B. M. 2017. Modeling some mineral nutrient requirements for micropropagated wild apricot shoot cultures. Plant Cell Tissue and Organ Culture 129:325-335.

Mendes, M. and Akkartal, E. 2009. Regression tree analysis for predicting slaughter weight in broilers. Italian Journal of Animal Science 8:615-624.

Mohammad, M. T.; Rafeeq, M.; Bajwa, M. A.; Awan, M. A.; Abbas, F.; Waheed, A.; Bukhari, F. A. and Akhtar, P. 2012. Prediction of body weight from body measurements using regression tree (RT) method for indigenous sheep breeds in Balochistan, Pakistan. The Journal of Animal and Plant Sciences 22:20-24.

Orhan, H.; Eyduran, E.; Tatliyer, A. and Saygici, H. 2016. Prediction of egg weight from egg quality characteristics via ridge regression and regression tree methods. Revista Brasileira de Zootecnia 45:380-385.

Ruhil, A. P.; Raja, T. V. and Gandhi, R. S. 2013. Preliminary study on prediction of body weight from morphometric measurements of goats through ANN models. Journal of the Indian Society of Agricultural Statistics 67:51-58.

Rumelhart, D. E. and McClelland, J. L. 1986. Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1. Foundations. MIT Press/Bradford Books, Cambridge, MA.

Salawu, E. O.; Abdulraheem, M.; Shoyombo, A.; Adepeju, A.; Davies, S.; Akinsola, O. and Nwagu, B. 2014. Using artificial neural network to predict body weights of rabbits. Open Journal of Animal Sciences 4:182-186.

Takma, C.; Atil, H. and Aksakal, V. 2012. Comparison of multiple linear regression and artificial neural network models goodness of fit to lactation milk yields. Kafkas Universitesi Veteriner Fakultesi Dergisi 18:941-944.

Tariq, M. M.; Eyduran, E.; Bajwa, M. A.; Waheed, A.; Iqbal, F. and Javed, Y. 2012. Prediction of body weight from testicular and morphological characteristics in indigenous Mengali sheep of Pakistan: Using factor analysis scores in multiple linear regression analysis. International Journal of Agriculture and Biology 14:590-594.

Topal, M.; Aksakal, V.; Bayram, B. and Yaganoglu, M. 2010. An analysis of the factor affecting birth weight and actual milk yield in swedish red cattle using regression tree analysis. The Journal of Animal and Plant Sciences 20:63-69.

Willmott, C. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30:79-82.

Yakubu, A. 2012. Application of regression tree methodology in predicting the body weight of Uda sheep. Animal Science and Biotechnologies 45:484-490.