Kafkas Universitesi Veteriner Fakultesi Dergisi
ISSN: 1300-6045 e-ISSN: 1309-2251
Journal Home-Page: http://vetdergikafkas.org

Kafkas Univ Vet Fak Derg
28 (3): 401-410, 2022
DOI: 10.9775/kvfd.2022.27163

Research Article

Phenotypic Characterization of Hair and Honamli Goats Using Classification Tree Algorithms and Multivariate Adaptive Regression Spline (MARS)

Yasin ALTAY [1,a (*)]

[1] Eskisehir Osmangazi University, Faculty of Agriculture, Department of Animal Science, Biometry and Genetics Unit, TR-26160 Eskişehir - TÜRKİYE
[a] ORCID: 0000-0003-4049-8301

Article ID: KVFD-2022-27163 Received: 21.01.2021 Accepted: 30.05.2022 Published Online: 31.05.2022

Abstract: Some morphological and physiological data are needed to scientifically describe animals and distinguish breeds from one another. Except for those who are not experts in the field, it is difficult to distinguish goat breeds from each other. Using data mining algorithms, this study aimed to develop a new phenotypic characterization for Honamli and Hair goats via some body measurement characteristics. In the study, some body characteristics of the Hair goat (65 animals) and the Honamli goat (83 animals) were used as independent variables. The dependent variable of the data mining algorithms, on the other hand, was defined as the binary response variable of Honamli and Hair breeds. The success of the CHAID, Exhaustive CHAID, CART, QUEST, and MARS algorithms in breed discrimination was determined at 87.80%, 85.80%, 87.80%, 77.00%, and 88.51%, respectively, while the area under the ROC curve was detected 0.880, 0.853, 0.868, 0.784, and 0.942, respectively, and Cohen's Kappa coefficient (κ) 0.755, 0.711, 0.749, 0.549 and 0.739, respectively. As a result, the phenotype characterization of Honamli and Hair goats, whose morphological distinctions could not be made exactly, in MARS and CHAID algorithms, achieved with high success compared to other methods. The present study showed that Honamli and Hair goats may be distinguished by suitable statistical algorithms based on morphological data, which can be integrated with goat breeding studies to detect the origin of breeding animals.

Keywords: *CART, CHAID, Classification, Exhaustive CHAID, MARS, QUEST*

Sınıflandırma Ağacı Algoritmaları ve Çok Değişkenli Uyarlanabilir Regresyon Uzanımları (MARS) Kullanılarak Kıl ve Honamlı Keçilerinin Fenotipik Karakterizasyonu

Öz: Hayvanları bilimsel olarak tanımlamak ve ırkları birbirinden ayırt etmek için bazı morfolojik ve fizyolojik verilere ihtiyaç vardır. Alanında uzman olmayanlar dışında keçi ırklarını birbirinden ayırt etmek güçtür. Bu çalışma, veri madenciliği algoritmaları kullanılarak bazı vücut özellikleri üzerinden Honamlı ve Kıl keçileri için yeni bir fenotipik karakterizasyon geliştirmeyi amaçlamıştır. Çalışmada, Kıl keçisi (65 hayvan) ve Honamlı keçisinin (83 hayvan) bazı vücut özellikleri bağımsız değişkenler olarak kullanılmıştır. Veri madenciliği algoritmalarının bağımlı değişkeni ise Honamlı ve Kıl ırkları ikili yanıt değişkeni olarak tanımlanmıştır. CHAID, Exhaustive CHAID, CART, QUEST ve MARS algoritmalarının ırk ayrımındaki başarısı sırasıyla %87.80, %85.80, %87.80, %77.00 ve %88.51 iken, ROC eğrisi altında kalan alan ise sırasıyla 0.880, 0.853, 0.868, 0.784 ve 0.942 ve Cohen's Kappa katsayıları (κ) 0.755, 0.711, 0.749, 0.549 ve 0.739 olduğu tespit edilmiştir. Sonuç olarak, morfolojik ayrımları tam olarak yapılamayan Honamlı ve Kıl keçilerinin MARS ve CHAID algoritmalarında fenotip karakterizasyonu diğer yöntemlere göre yüksek başarı ile gerçekleşmiştir. Bu çalışma, Honamlı ve Kıl keçilerinin morfolojik verilere dayalı uygun istatistiksel algoritmalarla ayırt edilebileceğini ve damızlık hayvanların kökenini tespit etmek için keçi ıslahı çalışmaları ile entegre edilebileceğini göstermiştir.

Anahtar sözcükler: *CART, CHAID, Sınıflama, Exhaustive CHAID, MARS, QUEST*

Introduction

Approximately 97% of the existing goats in Turkey consist of the Hair goats [1]. Hair and Honamli goats have some morphological similarities. Therefore, the breeds were not separated until the 2000s, and total numbers were evaluated as if all were the same breeds [2]. However, it is stated that the Honamli goat breed has higher productivity in terms of birth weight, live weight, lactation milk yield, and reproduction [3,4]. The lack of scientific research on

How to cite this article?
Altay Y: Phenotypic characterization of hair and Honamli goats using classification tree algorithms and multivariate adaptive regression spline (MARS). *Kafkas Univ Vet Fak Derg,* 28 (3): 401-410, 2022.
DOI: 10.9775/kvfd.2022.27163

(*) Corresponding Author
Tel: +90 222 239 3750 Fax: +90 222 324 2990
E-mail: yaltay@ogu.edu.tr

the Honamli goat breed is possibly due to the continuous transhumance of the Turkish Yoruks (nomads) [5]. The Honamli goat, which is defined as a new goat breed in animal genetic resources, was taken under protection by the Turkish General Directorate of Agricultural Research and Policies in 2015 [6].

To date, the morphological characteristics of Honamli have been defined and numerous studies have aimed to compare Honamli and Hair goats. Generally, these studies indicate that the phylogenetic similarity of Honamli and Hair goats is over 85%, and these breeds cannot be distinguished via microsatellite markers [2,7,8]. Therefore, the phenotypic characterization may be useful to separate these breeds. Because of Honamli and Hair goats morphologically having body clours are similar to each other, they can be separated subjectively by experienced breeders [4]. The reliability of this separation should be tested by robust quantitative methods and these methods could fill an important gap in the literature. Considering both the conservation of genetic resources and economical aspects, discrimination against the Honamli breed should be beneficial for goat breeders. Due to their very different morphological and physiological characteristics, goat breeders in Turkey have preferred Honamli goats in recent years, which are reared within the scope of the improvement projects in breeder conditions, to Hair goats [9].

Identification and classification of breeds within a certain species according to phenotypic characteristics play a key role in the basis of breeding and conservation program strategies. The identification and classification of breeds are of great importance to separate and define the breeds. For this purpose, comparisons of some local goat genotypes reared worldwide in terms of morphology traits and breed discrimination have been made using multivariate statistical methods (MANOVA), principal components analysis (PCA), canonical analysis, hierarchical, k-means clusters and step-wise, linear and nonlinear discriminate analysis [10-14]. To use these traditional statistical methods, the data must be multivariate normally distributed, the covariance matrix must be equal across all groups, the independent variables must be independent of each other, and the number of observations must be at least 10 times the number of independent variables included in the model [15]. The most significant distinction between data mining algorithms and traditional statistical methods is that they are not subject to any preconditions and can control classification through cross-validation using the sampling method [16]. As a result, it will be possible to classify with much greater accuracy and reliability [17].

This study, it was aimed to determine the phenotypic characterization of Hair and Honamli goats using data mining algorithms and MARS algorithm together on morphological characteristics.

Material and Methods

Ethical Statement

Ethical rules were considered by following all applicable international, national, and institutional guidelines for the care and use of animals. In the study, there is no need for ethical approval due to the lack of blood sampling from the animals and the absence of any surgical procedures. All data were collected with the approval of the breeder.

Animals

The animal material of the study consists of 65 Hair goats (45 female, 20 male) and 83 Honamli goats (73 female, 10 male) at different ages (1, 2, 3, 4, 5, and 6 years and over), which were reared extensively on a private farm in the Çalca district of Kütahya province, Turkey. In 2015, Honamli goats with pedigree records were brought from a farm that is a member of the Antalya Sheep and Goat Breeding Association. Hair goat breeding has been carried out on the farm, where the study was conducted since 2008. Pedigree records of the breeds were checked through the TurkVet system. During the heavy winter season or in adverse weather conditions, goats were fed in the barn. Concentrated feed was given for one month before the breeding season for flushing, and in addition to pasture, straw, alfalfa, and fescue grass are given as roughage sources.

Measurement of Morphological Characteristics of Breeds

A special scale designed for weighing small ruminants was used in determining the live weight of animals. All body characteristics of the goats were measured as described by [18], and live weight (LW) was taken with 0.1 kg precision. All body measurements were taken after the animals had adapted to the environment on a flat platform and the stress factors were minimized. Withers height (WH), back height (BH), rump height (RH), body length (BL), and chest depth (CD) were taken using a measuring stick and body circumference measurements (chest girth (CG), and leg girth (LG)) were taken using a measuring tape. Head length (HL), nose length (NL), ear length (EL), and tail length (TL) values were measured using calipers. All animals were measured by the same expert.

Statistical Analysis

Regression Tree-Based Data Mining Algorithms

The data structure created by using all arguments and dividing the data into subgroups is termed a classification tree. The root node, which does not contain any fragmentation and contains only the dependent variable, is at the top of the classification tree. First, this root node is divided into two or more parts. These separated parts are called parent branches. The breaking up of parent branches created child nodes or subsets [19]. The node, which is appeared when the fragmentation is complete in additional nodes and there is no more branching is called a terminal node [20]. By testing the independent variables in the model, the cut-off values of the explanatory variable are determined in a way to provide the specified category in the new node to be formed [21].

In all algorithms of this study, LW, WH, BH, RH, CD, BL, CG, LG, HL, NL, EL, TL variables and sex, age, and ear type factors were taken, while the dependent variable was binary goat breeds such as Honamli and Hair. In literature, many algorithms were used to create classification trees. In the study, tree-based data mining algorithms were used Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), Exhaustive CHAID, Quick Unbiased, Efficient Statistical Tree (QUEST), and Multivariate Adaptive Regression Spline (MARS) algorithm, which is a not tree-based algorithm. The main reason for using these algorithms is that they were simple to understand and allow for the determination of the cut-off points of the independent variables. In the classification trees, the maximum tree depth was used as CHAID (3), Exhaustive CHAID (3), CART (5), and QUEST (3), respectively. In the formation of classification trees, the minimum number of parent and daughter (child) nodes were taken as 10 and 5. Also, the multicollinearity problem was not detected to exist due to the pearson correlation coefficients and variance inflation factor (VIF) values between the independent variables used by the classification tree algorithms being smaller than the critical values specified in the literature [22].

CART (Classification and Regression Trees)

The CART algorithm is a non-parametric regression method developed by some researchers [23]. The CART is a tree-based data mining algorithm that reveals the relationship between the dependent variable and the independent variable, as well as the relationships between the independent variables. The branching in the tree structure is based on the division into two sub-homogeneous groups. As the split criterion in the CART algorithm, impurity and Gini index variability are taken into account in the selection of the best independent variable in the classification. The Gini index takes values between 0 and 1 and provides assignments to classes. The Gini index is calculated by subtracting the sum of the squares of the probabilities of each class from one and is obtained using equation 1.

$$\text{Gini Index (L)}=1-\sum_{i=1}^{j} p_i^2 \qquad (1)$$

j: number of class; L: a data set with j th class; pi: relative frequency if class 'i' in 'L'

CHAID (Chi-squared-Automatic-Interaction-Detection)

Some researchers [24] developed the CHAID algorithm, which is a non-parametric regression method in the tree structure created by taking statistical significance ratios and cross tables into account. Branching in the tree structure is based on the split of two or more sub-homogeneous groups. CHAID algorithm with merge, split, and stop stages iteratively creates homogeneous nodes starting from the root node, increasing/decreasing variance between/within nodes [25]. Because the whole population can be split into stable sub-nodes using a strong translation algorithm, a regression equation to be obtained is kept independent of classical assumptions (normality, linearity, homogeneity, etc.) in CHAID analysis. With this process, normality and homogeneity can be achieved in the distribution of the data.

Exhaustive CHAID (Exhaustive Chi-squared-Automatic-Interaction-Detection)

It is a modified version of the CHAID algorithm that looks into all possible splits for each predictor developed by some researchers [26]. As a result, transactions take longer than with the CHAID algorithm. Exhaustive CHAID keeps combining the prediction variable's categories until only two supercategories remain. It identifies the category with the strongest relationship to the dependent variable and computes the adjusted p-value for these. Although it depends on the data, it can be said that there will be no significant difference between the results of the CHAID algorithms.

QUEST (Quick Unbiased, Efficient Statistical Tree)

Some researchers [27] created QUEST as a classification algorithm, and the branching in the tree structure, like the CART algorithm, is based on the separation of two sub-homogeneous groups. Unlike CHAID and CART, it handles variable selection and split point selection separately. In the QUEST algorithm, the association between each independent and dependent variable for each separation is found by calculating the F test, Levene test, or Pearson Chi-square value. In the algorithm, the variable with a small p-value is selected for explanatory variable selection.

MARS (Multivariate Adaptive Regression Splines)

The MARS algorithm, developed by some researchers [28], is a non-parametric regression technique used to examine complex relationships between the dependent variable and a set of independent variables. In order to apply this non-linear technique, there is no need for any assumptions between the dependent variable and predictor variables. The MARS algorithm, which is a modified version of the CART algorithm, makes better predictions than binary logistic regression thanks to the hinges function in its structure [29]. Using appropriate transformation techniques, the MARS technique converts nonlinear relationships between dependent and independent variables into linear ones. The MARS method can calculate the best transformations and interactions of variables, as well as analyze complex relationships in high-dimensional data. To prevent these relationships from causing multicollinearity problems in the MARS algorithm, it is recommended that a model be created in the earth package of the R software with penalty=2 [17].

k-Fold Cross-Validation

Cross-validation is a popular method for assessing the effectiveness of a machine learning model. This method is used for small datasets and is based on the resampling procedure. It can also be done using cross-validation instead of dividing the data into training and test sets because it works based on validation in machine learning algorithms [30]. Cross-validation is used to train and validate the model by dividing all the data into k multiples, also known as subsamples. In this way, it reduces overfitting and determines the model's hyperparameters. For this purpose, usually 10-fold or 5-fold cross-validation is used. In this study, after all data set (148 records) was randomly divided into 10 parts, nine parts of the training set of the models were created, while the model was validated 5 times with the remaining part in this study.

Model Evaluation Criteria

CART, CHAID, Exhaustive CHAID, QUEST, and MARS data mining algorithms were utilized to compare in terms of accuracy, sensitivity, specificity, Matthews correlation (Phi), Cohen's Kappa coefficient (κ), and area under Receiver Operating Characteristics (ROC) curve. The Phi coefficient was used to determine the relationship between the real classes and the classes estimated by the algorithms, and Cohen's Kappa coefficient was used to determine the concordance. Accuracy is the proportion at which a classification algorithm correctly separates Honamli and Hair goats. Sensitivity is the proportion at which the algorithm correctly classifies Honamli goats, while specificity is the proportion at which the algorithm correctly classifies Hair goats. *Table 1* presents the confusion table for classifying algorithms.

The expressions T+, T-, F+, and F- used in the accuracy, sensitivity and specificity equations represent numbers of true positive, true negative, false positive, and false negative, respectively. The formula is used below to determine the area under the ROC curve (AUC) and the area under the ROC curve's standard error (AUCse) as developed by [31].

Table 1. *Confusion table for the classifier algorithms*

Observed	Predicted as Breeds	
	Honamli	Hair
Honamli	T^+	F^+
Hair	F^-	T^-

Accuracy = $(T^+ + D)/(T^+ + F^+ + F^- + T^-)$

Sensitivity = $T^+/(T^{+-} + F^+)$

Specificity = $T^- / (F^- + T^-)$

Error proportion = 1 - Accuracy

$$se_{AUC} = \sqrt{\frac{AUC(1-AUC)+(n_A-1)(q1-AUC^2)+(n_B-1)(q2-AUC^2)}{n_A n_B}} \quad (2)$$

$n_A = T^+ + F^-$ and $n_B = F^+ + T^-$

$q1 = \frac{AUC}{2-AUC}$ and $q2 = \frac{2AUC^2}{1+AUC}$

Statistical analyses of the classification trees, Phi and Cohen's Kappa (κ) coefficients were performed in IBM SPSS 23 package program [32]. Earth (v5.1.2) [33] and caret (v60.86) [34] packages of R software were used for MARS analysis [35]. The trial version (19.5.1) of the MedCalc package program was used to determine the areas under the ROC and to compare (z test) the area under the ROC curve of the algorithms.

Results

Categorical variables belonging to Honamli and Hair goats in the study are given in *Table 2*. Descriptive statistics of continuous variables obtained from Honamli and Hair goats are given in *Table 3*. Honamli females were larger than males because females in the herd were older *(Table 3)*. Young billy goats were preferred in the herd to reduce generation intervals.

Table 2. *Categorical variables belonging to Honamli and Hair goats*

Factors	Levels		N	Percentage (%)
Breed-Sex	Honamli	Female	73	49.32%
		Male	10	6.76%
	Hair	Female	45	30.41%
		Male	20	13.51%
Breed-Age	Honamli	1	16	10.81%
		2	6	4.06%
		3	14	9.46%
		4	4	2.70%
		5	2	1.35%
		6	41	27.70%
	Hair	1	32	21.62%
		2	2	1.35%
		3	4	2.70%
		4	6	4.06%
		5	2	1.35%
		6	19	12.84%
Breed-Ear	Honamli	Comuk (native terms)	27	18.24%
		Lop	56	37.84%
	Hair	Comuk (native terms)	21	14.19%
		Lop	44	29.73%

Table 3. Descriptive statistics on live weight and some body measurements in Honamli and Hair goats of different age

Traits	Breed	Sex	N	Minimum	Maximum	Mean±SE	StdDev	CoefVar
LW	Honamli	Female	73	27.10	84.70	60.76±1.54	13.13	21.60
		Male	10	37.20	63.10	48.02±2.77	8.76	18.24
	Hair	Female	45	27.30	72.20	48.40±1.76	11.77	24.33
		Male	20	29.00	43.60	36.15±0.95	4.25	11.74
WH	Honamli	Female	73	51.90	89.50	75.78±1.04	8.90	11.75
		Male	10	66.00	87.00	76.25±2.37	7.51	9.84
	Hair	Female	45	51.70	85.00	70.63±1.20	8.05	11.39
		Male	20	63.50	77.50	68.70±0.94	4.22	6.15
BH	Honamli	Female	73	64.50	90.50	79.14±0.71	6.05	7.65
		Male	10	66.50	86.00	76.60±1.97	6.23	8.13
	Hair	Female	45	60.00	79.50	71.09±0.69	4.66	6.55
		Male	20	61.00	75.50	68.40±1.01	4.50	6.58
RH	Honamli	Female	73	65.50	89.00	79.14±0.66	5.63	7.12
		Male	10	70.00	86.50	77.45±1.69	5.36	6.92
	Hair	Female	45	60.50	81.00	71.82±0.72	4.80	6.68
		Male	20	61.00	77.00	68.08±0.98	4.39	6.45
CD	Honamli	Female	73	16.50	29.50	24.66±0.29	2.47	10.00
		Male	10	20.50	27.50	24.45±0.76	2.41	9.85
	Hair	Female	45	18.50	27.50	22.36±0.28	1.86	8.31
		Male	20	18.50	23.50	21.13±0.32	1.44	6.82
BL	Honamli	Female	73	46.00	91.50	80.57±0.97	8.25	10.24
		Male	10	62.50	88.00	75.65±2.70	8.54	11.29
	Hair	Female	45	60.00	92.00	74.63±1.12	7.50	10.05
		Male	20	58.00	76.50	68.30±0.83	3.69	5.41
CG	Honamli	Female	73	70.00	104.00	91.04±0.82	6.96	7.65
		Male	10	81.00	95.00	86.25±1.36	4.30	4.99
	Hair	Female	45	73.00	102.50	86.64±1.05	7.02	8.11
		Male	20	72.50	90.50	81.50±0.98	4.38	5.37
LG	Honamli	Female	73	35.00	67.00	51.21±0.74	6.29	12.28
		Male	10	46.00	65.00	55.90±2.05	6.48	11.59
	Hair	Female	45	39.00	77.50	50.87±1.13	7.58	14.90
		Male	20	46.50	61.00	55.00±0.75	3.35	6.09
HL	Honamli	Female	73	17.00	24.00	21.23±0.18	1.49	7.03
		Male	10	19.50	22.50	20.80±0.31	0.98	4.70
	Hair	Female	45	17.00	23.00	19.67±0.22	1.45	7.39
		Male	20	17.50	22.50	20.13±0.29	1.30	6.44
NL	Honamli	Female	73	11.00	21.00	14.26±0.25	2.13	14.95
		Male	10	12.00	16.00	13.95±0.46	1.46	10.48
	Hair	Female	45	11.00	17.00	13.37±0.24	1.58	11.79
		Male	20	11.50	23.00	14.03±0.53	2.37	16.86
EL	Honamli	Female	73	8.00	22.50	16.97±0.44	3.74	22.02
		Male	10	9.50	21.00	15.50±1.34	4.22	27.24
	Hair	Female	45	13.00	28.00	17.79±0.45	3.00	16.88
		Male	20	7.00	20.50	14.68±0.86	3.86	26.31
TL	Honamli	Female	73	13.00	27.50	19.11±0.39	3.37	17.61
		Male	10	15.00	29.00	20.35±1.35	4.28	21.04
	Hair	Female	45	11.00	22.00	15.97±0.35	2.31	14.50
		Male	20	14.00	20.00	16.43±0.42	1.88	11.44

Live weight (LW), Withers height (WH), Back height (BH), Rump height (RH), Chest Depth (CD), Body length (BL), Chest girth (CG), Leg girth (LG), Head length (HL), Nose length (NL), Ear length (EL), and Tail length (TL)

The MARS algorithm, which provided one of the best classifications of Honamli and Hair goats, takes the form of body characteristics "LW", "BH", "CD", "HG", and "HL" as independent variables in the prediction model. In addition, the model also includes "Age" variable and "Sex" factors that do not have body characteristics. The remaining characteristics were not included in the MARS model because they were found to be statistically non-significant (P<0.05). In the MARS model which was given below, GLM indicates the general linear model, while max denotes the basic function of the MARS.

$GLM_{HONAMLI}$ = -0.5232799 - 3.033782 * SexMale + 1.5192 * max(0, 4 - Age) - 1.068315 * max(0, 35.4 - LW) + 0.4609831 * max(0, BH - 72) + 25.86152 * max(0, BH - 82) - 0.6795643 * max(0, 25 - CD) + 1.559002 * max(0, 77.5 - HG) - 0.5741605 * max(0, 21.5 - HL). The probability of any goat being Honamli can be estimated by PHONAMLI = exp$GLM_{HONAMLI}$/(1+ exp $GLM_{HONAMLI}$). The "exp" value used in the equation refers to the base of the natural logarithm of 2.718. Using the basic MARS model, it is possible to derive a new prediction equation used in the classification of females. If the goats used in breed discrimination estimation are female animals older than four years old, the following equation can be used.

$GLM_{HONAMLI}$ = -0.5232799 - 1.068315 * max(0, 35.4 - LW) + 0.4609831 * max(0, BH - 72) + 25.86152 * max(0, BH - 82) - 0.6795643 * max(0, 25 - CD) + 1.559002 * max(0, 77.5 - HG) - 0.5741605 * max(0, 21.5 - HL).

For example, a 4-year-old female Honamli goat with body characteristics which was LW = 40 kg, BH = 78 cm, CD = 25 cm, HG = 75 cm, and HL = 20 cm in the dataset could be estimated discrimination of breed. As follows by the MARS estimation equation;

1- $GLM_{HONAMLI}$ = -0.5232799 - 3.033782 * SexMale (Female=0) + 1.5192 * max(0, 4 - 4) - 1.068315 * max(0, 35.4 - 40) + 0.4609831 * max(0, 78 - 72) + 25.86152 * max(0, 78 - 82) - 0.6795643 * max(0, 25 - 25) + 1.559002 * max(0, 77.5 - 75) - 0.5741605 * max(0, 21.5 - 20)

2- $GLM_{HONAMLI}$ = -0.5232799 + 0.4609831 * max(0, 78 - 72) + 1.559002 * max(0, 77.5 - 75) - 0.5741605 * max(0, 21.5 - 20)

3- $GLM_{HONAMLI}$ = -0.5232799 + 0.4609831 *6 + 1.559002 * 2.5 - 0.5741605 * 1.5

4- $GLM_{HONAMLI}$ = 5.27888295

5- $P_{HONAMLI}$ = exp$GLM_{HONAMLI}$/(1+ exp $GLM_{HONAMLI}$)

6- $P_{HONAMLI}$ = 2.7185.27888295/(1+ 2.7185.27888295)

$P_{HONAMLI}$ = 0.994924974

The estimated goat with a probability of 99.49% belongs to the Honamli breed.

Classification performances of data mining algorithms used for breed discrimination are shown in *Table 4*. The areas under the ROC curve (AUC) were statistically significant for all algorithms for breed discrimination (P<0.01).

The sensitivity and specificity values of the model's criteria were close to each other and the AUC values were close to 1, which indicated the accuracy of the classification *(Fig. 1)*. Models compared statistically with the z-test in terms of AUC could be mathematically expressed as MARS = CHAID = CART> = Exhaustive CHAID> = QUEST *(Table 4)*. When all data mining algorithms were compared among themselves in terms of AUC performance criteria, it was determined that the most successful algorithm used in breed discrimination were MARS, CHAID, CART, and Exhaustive CHAID. The performance of the MARS algorithm was determined as 0.916, 0.846, and 0.937 in terms of sensitivity, specificity, and general accuracy rate respectively. The MARS algorithm was able to classify 75 of 83 Honamli goats, 55 of 65 Hair goats, and 88.50% of all goats correctly. The MARS algorithm was found to have the highest breed discrimination diagnostic test with the area under the ROC curve of 0.942. Also, the concordance (κ) and correlation (Phi) coefficients between the breeds estimated by the MARS algorithm and the actual breeds were found to be 0.739. It was determined that the CHAID algorithm had the best diagnostic test performance and other performance criteria among the classification tree algorithms. The discrimination performances made by the CHAID algorithm had the values of sensitivity, specificity and accuracy respectively as 0.911, 0.841, and 0.878. The CHAID algorithm allocated 11 of 83 Honamli incorrectly and 72 correctly, while it separated 58 of 65 Hair goats correctly. CHAID has the second-largest AUC value as 0.880 after the MARS algorithm. In addition, among the

Table 4. *Classification performances of the data mining algorithms for each diagnosis test*

Algorithm	Sensitivity	Specificity	Matthews Correlation (Phi)	Cohen's Kappa Coefficient (κ)	AUC±SE	Accuracy of Model	Correctly Classify of Honamli Breed	Correctly Classify of Hair Breed	P-Value
MARS	0.916	0.846	0.739	0.739	0.942±0.028^{a}	0.885	0.894	0.892	<0.001
CHAID	0.911	0.841	0.756	0.755	0.880±0.027^{a}	0.878	0.867	0.892	<0.001
CART	0.849	0.927	0.756	0.749	0.868±0.023^{a}	0.878	0.952	0.785	<0.001
Exhaustive CHAID	0.861	0.855	0.711	0.711	0.853±0.030^{ab}	0.858	0.892	0.815	<0.001
QUEST	0.889	0.682	0.569	0.549	0.784±0.032^{b}	0.770	0.675	0.892	<0.001

^{a,ab,b} *The difference between AUC with letter in any data mining algorithm column is significant (P<0.05)* tree-based data mining algorithms, the CHAID algorithm had the highest concordance with a Kappa (κ) value of 0.755, while it had the same correlation coefficient (Phi) as the CART algorithm with a value of 0.756. Although the CART algorithm correctly classified Honamli goats with a high rate (95.20%), the correct classifying percentage of Hair goats (78.50%) remained low. The performances of CART were determined as 0.849, 0.927, and 0.848 for sensitivity, specificity, and accuracy rate respectively. The CART algorithm estimated 79 of 83 Honamli goats, 51 of 65 Hair goats, and 87.80% of all goats by classifying them correctly. Moreover, the CART algorithm had the third-largest AUC (0.868), and the coefficient of agreement between actual breeds and breeds estimated by the CART algorithm was 0.749. The Exhaustive CHAID algorithm had performance values as 0.861 for sensitivity, 0.855 for specificity 0.855, and 0.858 for accuracy rate. While the Exhaustive CHAID algorithm classified 74 of 83 Honamli goats correctly, this algorithm assigned 12 of 65 Hair goats incorrectly. Exhaustive CHAID had the 4th largest area under ROC among algorithms, with an AUC of 0.853. The coefficient of concordance (κ) and correlation (Phi) between the predicted values of this algorithm and the actual values were 0.711. Although the Exhaustive CHAID algorithm correctly separated both breeds in close percentages, their performance values were a little low compared to other algorithms (MARS, CHAID, and CART). The QUEST algorithm correctly separated Hair goats with a high rate (89.20%) but, the separation percentage of Honamli goats (67.50%) remained quite low. It was also the algorithm with the worst prediction performance in terms of other performance criteria.

It has been determined that the CHAID algorithm was one of the best classifiers among classification trees for Honamli and Hair goat discrimination *(Table 4)*. When the CHAID diagram is examined, it was determined that the first order effective independent variable on breed discrimination was RH (Adj. P-value = 0.000, χ2 = 59.332), second order was Age (Adj. P-value = 0.014, χ2 = 9.981), and BH (Adj. P-value = 0.036, χ2 = 6.313), and third-order independent variables were LG (Adj. P-value = 0.045, χ2 = 13.362) and CD (Adj. P-value = 0.003, χ2 = 12.577) *(Fig. 2)*. Branches generated by independent variables in the entire tree structure were statistically significant (P<0.05).

All goats considered in the study were divided into 3 sub-groups (nodes) in terms of RH variable. In the first node, 39 (83%) of the goats with RH = <71.00 cm shorter were Hair and 8 (17%) of them were Honamli. In the second node, 25 Hair (43.1%) and 33 Honamli (56.9%) of 58 goats were classified in a range of 71.0 <RH = <79.0. In the third node, it was determined that 42 of the goats (71.9 <RH) with RH traits more than 79 cm were Honamli (97.7%) and only one of them was Hair goat.

Goats (3rd node) with RH characteristics greater than 79 cm formed the 6th and 7th nodes in terms of GH characteristics. In the 6th node, 83.30% of the goats with the BH trait less or equal to 79.50 cm were classified as Honamli and 16.70% as Hair goat. All of the goats with the BH trait values greater than 79.50 belong to the Honamli breed (7th node).

While the 3rd and 4th nodes of the CHAID algorithm diagram showed a division according to the age variable, it did not have a direct effect on breed discrimination.

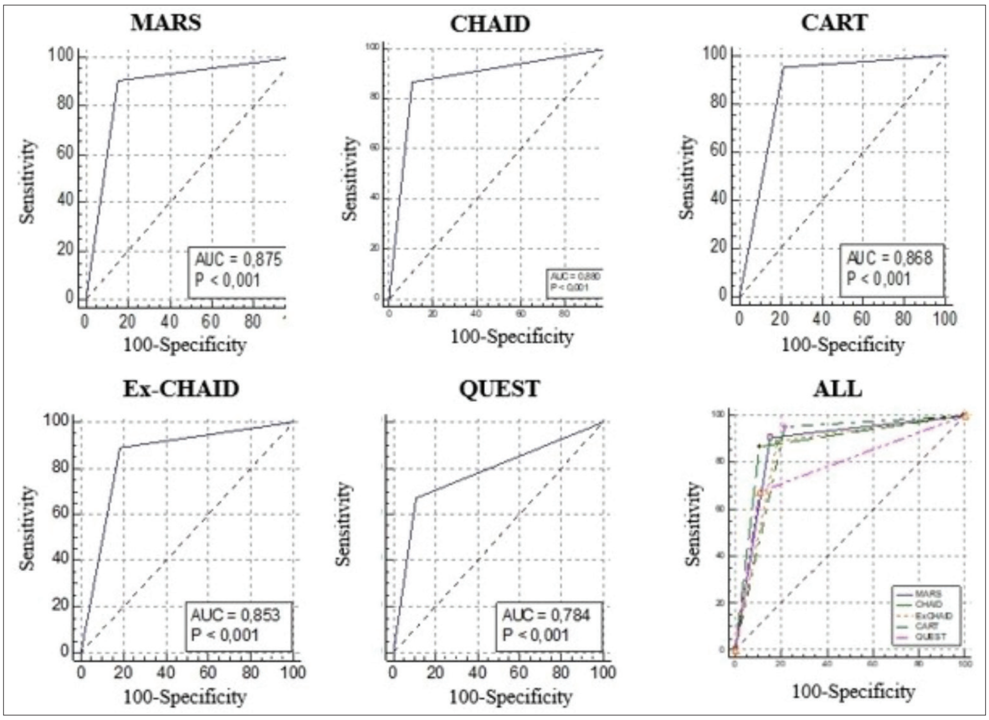

Fig 1. All and individual ROC curves of classifying algorithms for diagnostic tests of breed discrimination

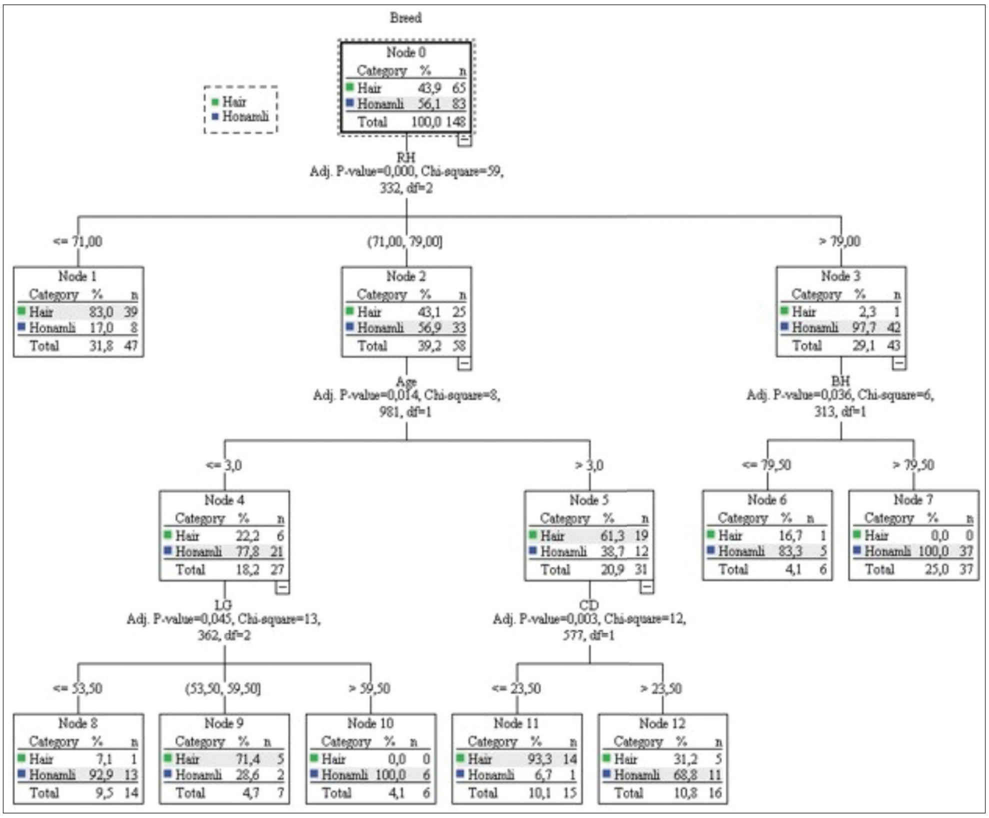

Fig 2. CHAID classification tree diagram of the diagnosis test of breed discrimination

Accordingly, it was determined that 13 of the goats aged 3 and under were Honamli (92.9%), and one of them was Hair goat (7.1%) (LG = <53.50) in terms of LG (8th node). In 9th node, the goats between 53.50 <LG = <59.50 were classified as 2 Honamli (28.6%) and 5 as Hair (71.4%). At the 10th node, 100% of all goats with the LG trait larger than 79 cm belong to the Honamli breed. In the 11th node, when the CD of goats older than three years was less than and equal to 23.50 cm, (CD = <23.50), 93.30% of goats are classified as Hair and 6.7% as Honamli. If the CD trait is greater than 23.50 cm (23.50 <CD), the probability of finding a Honamli goat is 68.80% and 31.20% is a Hair goat (12th node).

Discussion

The most successful data mining algorithms used in the phenotypic characterization of Honamli and Hair goats were MARS and CHAID. While the MARS algorithm used "LW", "BH", "CD", "CG" "Sex", "Age", and "HL" traits as independent variables in breed discrimination, the CHAID algorithm used "RH", "Age", "BH", "LG", and "CD". The reason why these two algorithms use different independent variables was their different working principles. While the CHAID algorithm created a more homogeneous subset, the MARS algorithm reveals the independent variables and coefficients of regression that affect the prediction model. Also, the CHAID algorithm determines the independent variables by using the Chi-square statistics and the Bonferroni corrected P-value after categorizing the independent variables and converting them into binary crosstabs [36,37]. The MARS algorithm, on the other hand, selects the independent variable using the generalized cross-validation error (GCV) method. The non-significant variables in the model are eliminated using the penalty function (λ) in the GCV term, and in this way, the multicollinearity problem is avoided [33].

It is claimed that the distinction between Honamli and Hair goats will be made by experienced breeders using HL, NL, and TL characteristics. This, however, is a subjective statement that has not been scientifically proven by any literature. Based on the findings of our study, it is understood that this is just a discourse with no scientific value. BH and CD features are common independent variables used by MARS and CHAID algorithms in the current study. It can be used in both algorithms to discriminate between these two breeds. However, because the CHAID algorithm uses fewer explanatory variables than the MARS algorithm, it may be preferred by breeders or researchers in terms of time and labor.

Essentially the same key variables can be used to describe closely related animal species [38]. Nsoso et al.[10] reported that the effect of age is important for the phenotypic characterization of indigenous Tswana goats reared in Bostwana. Body length (BL) and chest girth (CG) characteristics were reported to differ significantly in the distinction between Brown and Gray Bengal goats [39]. It was emphasized that cannon bone circumference (CBC), chest girth (CG), chest depth (CD), rump height (RH), rump length (RL), and withers height (WH) traits are important for the distinction of five different indigenous goat breeds in Spain [12]. Gonzalez-Martinez et al.[40] reported that the chest depth (CD) and rump height (RH) characteristics of the Murciano-Granadina and Malagueña dairy goat breeds reared in Spain are important in breed discrimination. In Jordan, four indigenous breeds and crossbred goats were separated by simple, cluster, canonical, and stepwise discriminant analysis by using the morphological characteristics. The independent variables used in this distinction were reported as nose shape, withers height (WH), live weight (LW), ear type (ET), color, teat placement, chest width (CW), withers depth (WD), and rump width (RW) [41]. Although the statistical methods used in these studies were different, the goat breeds were reared in different environmental conditions, and their genetic structures are different, they were partially compatible with our study results.

Bourzat et al.[42] proposed two different indices for a simple classification of goats. The first index is the difference between withers height (WH), and chest depth (CD), while the other is the difference between ear length (EL) and chest depth (CD). Based on a univariate statistic for such a classification, it should not be discussed how successful these methods can be when all body characteristics of goats are considered together. The classification should be determined by a multivariate method and the methods used should be quite powerful. In the literature, they classified goats by using multivariate analysis methods such as discriminant, clustering, canonical, principal components analysis (PCA), multivariate statistical methods (MANOVA), etc. [10,11,13,41,43-45]. However, a strong classification could not be made since the multivariate classification methods have some prerequisites and the methods used do not have calibration (validation) capabilities [16]. In this context, it would be a more accurate approach to use data mining algorithms that are more powerful than the methods used and that can control the algorithm by cross-validation [33].

Orhan et al.[46] reported that there was a statistical difference in terms of strength, thickness, cuticle, medulla, and cortex characteristics of the hair structure of Honamli and Hair goats ($P<0.05$), but there was not any difference in terms of bulbous pili and scapus pili characteristics ($P>0.05$). Although they are phenotypically different and the individual comparison of the hair structure characteristics of goats is an important finding, it is not known what the result will be for breed discrimination when all the features are examined together. In our current study, although the data set has a very heterogeneous structure, it is seen that the characteristics of goats that are important for breed discrimination can be successfully made using data mining algorithms.

In this study results, showed that a new phenotypic characterization successfully allows distinguishing of Honamli and Hair goat breeds by using some body measurements and factors by data mining algorithms. Considering the successful performances of three different classification trees and MARS in breed distinction, CHAID and MARS methods can be used to make a more accurate classification. Moreover, data mining algorithms enable the discrimination of breed, which is phenotypically similar. In this way, the separation of phenotypically similar animals with powerful classification tools can be used as a preliminary step in selection programme. The results suggest that data mining algorithms could contribute to future studies about breed distinction of animals and might have a good potential for the protection of animal genetic resources. In addition, there is a need for studies that will be used in different species of animals by using data mining algorithms with both genetic and phenotype data. In this way, it is hoped that by doing so, a new quantitative method for the supply of breeding material can be developed.

Ethical Statement

In the study, there is no need for ethical approval due to the lack of blood sampling from the animals and the absence of any surgical procedures. All data were collected with the approval of the breeder.

Availability of Data and Materials

The author declares that data supporting the study findings are also available to the corresponding author.

Acknowledgments

I would like to thank Ramazan Karayel and Mehmet Toscu goat farm in Kütahya for obtaining body characteristics and providing the records. In addition, the author acknowledge that this work is original and that no part of or complete content of this has been published in any other journal and congress.

Funding Support

This study was not funded by any organization.

Competing Interests

The author declares no competing interests.

References

1. **Atay O, Gökdal Ö, Eren V:** Reproductive characteristics and kid marketing weights of hair goat flocks in rural conditions in Turkey. *Cuba J Agric Sci*, 44 (4): 353-358, 2010.

2. **Karadag O:** A study on the investigation of Honamli through some morphological characteristics fertility and casein genes polymorphism. *PhD Thesis*. Namık Kemal University Graduate School of Natural and Applied Sciences, 2016.

3. Varol M: Definition of morphological traits of Hair goats in Denizli province. *MSc Thesis,* Adnan Menderes University Graduate School of Natural and Applied Sciences, 2014.

4. Akbas AA, Saatci M: Growth, slaughter, and carcass characteristics of Honamlı, Hair, and Honamlı x Hair (F1) male goat kids bred under extensive conditions. *Turk J Vet Anim Sci*, 40 (4): 459-467, 2016. DOI: 10.3906/vet-1511-5

5. Taskaya H, Kale M: Investigation of Caprine Arthritis Encephalitis Virus (CAEV) infection in Honamlı goat breed. *MAE Vet Fak Derg*, 5 (2): 58-63, 2020. DOI: 10.24880/maeuvfd.682590

6. Elmaz Ö, Akbaş AA, Saatcı M: Effects of birth type on growth, fattening performance and carcass characteristics in Honamlı male kids. *Kafkas Univ Vet Fak Derg*, 23 (5): 749-755, 2017. DOI: 10.9775/kvfd.2017.17703

7. Bulut Z, Kurar E, Ozsensoy Y, Altunok V, Nizamlioglu M: Genetic diversity of eight domestic goat populations raised in Turkey. *BioMed Res Int*, 2016:2830394, 2016. DOI: 10.1155/2016/2830394

8. Karsli T, Demir E, Fidan HG, Aslan M, Karsli BA, Arik IZ, Semerci ES, Karabag K, Balcioglu MS: Determination of genetic variability, population structure and genetic differentiation of indigenous Turkish goat breeds based on SSR loci. *Small Ruminant Res*, 190:106147, 2020. DOI: 10.1016/j.smallrumres.2020.106147

9. Aytekin I: Effects of two different rearing systems on kid growth in Honamli goat. **In,** *Proceedings of International Human and Nature Sciences: Problems and Solution Seeking Congress*, 7-9 October, Sarajevo, Bosnia and Herzegovina, 2016.

10. Nsoso SJ, Podisi B, Otsogie E, Mokhutshwane BS, Ahmadu B: Phenotypic characterization of indigenous Tswana goats and sheep in Botswana: Continuous traits. *Trop Anim Health Prod*, 36 (8): 789-800, 2004. DOI: 10.1023/B:TROP.0000045979.52357.61

11. Jordana J, Ribo O, Pelegrin M: Analysis of genetic relationships from morphological characters in Spanish goat breeds. *Small Ruminant Res*, 12, 301-314, 1993. DOI: 10.1016/0921-4488(93)90065-P

12. Herrera M, Rodero E, Gutierrez MJ, Peña F, Rodero JM: Application of multifactorial discriminant analysis in the morphostructural differentiation of Andalusian caprine breeds. *Small Ruminant Res*, 22, 39-47, 1996. DOI: 10.1016/0921-4488(96)00863-2

13. Capote J, Delgado JV, Fresno M, Camacho ME, Molina A: Morphological variability in the Canary goat population. *Small Ruminant Res*, 27, 167-172, 1998. DOI: 10.1016/S0921-4488(97)00047-3

14. Herrera PSJI, Rodero E, Sànchez MD, Luque M: Raza caprina Moncaina. 1. Caracteres cuantitativos morfoestructurales. **In,** *Proceedings of Actas XXVIII Congreso de la SEOC,* 8-10 October, Granada, Spain, 2003.

15. Alpar C: Uygulamalı Çok Değişkenli Istatistiksel Yöntemler. 4th Ed., Detay Yayıncılık, Ankara, Turkey, 2017.

16. Maimon OZ, Rokach L: Data mining with decision trees: Theory and applications. **In,** Rocach L, Maimon O (Eds): Series In Machine Perception and Artificial Intelligence 2nd ed., World Scientific, Danvers, USA, 2014.

17. Akın M, Eyduran SP, Eyduran E: Mars algorithm in solving regression and classification type problems in agricultural sciences with R software. 1-264, Nobel Academic Publishing, Ankara, Turkey, 2020.

18. Ertugrul M: Ovine Breeding Practices. 2nd ed., Ankara University Faculty of Agriculture Publications Number: 1446, Ankara, Turkey, 1996.

19. Eyduran E, Keskin I, Erturk YE, Dag B, Tatliyer A, Tirink C, Aksahan R, Tariq MM: Prediction of fleece weight from wool characteristics of sheep using regression tree method (Chaid Algorithm). *Pak J Zool*, 48 (4): 957-960, 2016.

20. Orucoglu O: Determination of environmental factors affecting 305-day milk yield of Holstein cows by regression tree method. *MSc Thesis.* Suleyman Demirel University, Graduate School of Natural and Applied Sciences, 2011.

21. Aksahan R, Keskin I: Determination of the some body measurements effecting fattening final live weight of cattle by the regression tree analysis. *Selcuk J Agr Food Sci*, 2 (1): 53-59, 2015.

22. Kim SH, Kim CY, Seol DH, Choi JE, Hong SJ: Machine learning-based process-level fault detection and part-level fault classification in semiconductor etch equipment. *IEEE Trans Semicond Manuf,* 35, 174-185, 2022. DOI: 10.1109/TSM.2022.3161512

23. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and regression trees. 1-368, Chapman & Hall/CRC, Wadsworth, New York, USA, 1984.

24. Kass GV: An exploratory technique for investigating large quantities of categorical data. *J R Stat Soc Ser C Appl Stat*, 29 (2): 119-127, 1980. DOI: 10.2307/2986296

25. Karakaya E, Çelik Ş, Taysi MR: CHAID algoritması ile balıketi tüketimini etkileyen faktörlerin incelenmesi. *JAFAG*, 35 (2): 85-93, 2018. DOI: 10.13002/jafag4381

26. Biggs D, De Ville B, Suen E: A method of choosing multiway partitions for classification and decision trees. *J Appl Stat*, 18 (1): 49-62, 1991. DOI: 10.1080/02664769100000005

27. Loh WY, Shih YS: Split selection methods for classification trees. *Stat Sin*, 7, 815-840, 1997.

28. Friedman JH: Multivariate adaptive regression splines. *Ann Statist*, 19, 1-67, 1991. DOI: 10.1214/aos/1176347963

29. Tırınk C, Eyduran E, Faraz A, Waheed A, Tauqir NA, Nabeel MS, Tariq MM, Sheikh IS: Use of multivariate adaptive regression splines for prediction of body weight from body measurements in Marecha *(Camelus dromedaries)* camels in Pakistan. *Trop Anim Health Prod*, 53 (3): 1-10, 2021. DOI: 10.1007/s11250-021-02788-y

30. Huma ZE, Iqbal F: Predicting the body weight of Balochi sheep using a machine learning approach. *Turk J Vet Anim Sci*, 43 (4): 500-506, 2019. DOI: 10.3906/vet-1812-23

31. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36, 1982. DOI: 10.1148/radiology.143.1.7063747

32. IBM Corp. Released: IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp., 2015.

33. Eyduran E, Akin M, Eyduran SP: Application of multivariate adaptive regression splines in agricultural sciences through R software. 1-112, Nobel Academic Publishing, Ankara, Turkey, 2019.

34. Kuhn M: Caret: Classification and regression training. Retrieved from. https://CRAN.R-project.org/package=caret; *Accessed:* 07.12.2020.

35. R Core Team: R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. https://www.Rproject.org/; *Accessed:* 05.07.2020.

36. Altay Y, Boztepe S, Eyduran E, Keskin İ, Tariq MM, Bukhari FA, Ali I: Description of factors affecting wool fineness in Karacabey Merino sheep using chaid and mars algorithms. *Pak J Zool*, 53 (2): 691, 2021. DOI: 10.17582/journal.pjz/20190329150359

37. Olfaz M, Tırınk C, Önder H: Use of CART and CHAID algorithms in Karayaka sheep breeding. *Kafkas Univ Vet Fak Derg*, 25 (1): 105-110, 2019. DOI: 10.9775/kvfd.2018.20388

38. FAO: Phenotypic Characterization of Animal Genetic Resources. FAO Animal Production and Health Guidelines No: 11. Rome, Italy, 2012.

39. Mukeherjee DK, Singh CSP, Mishra HR: Anote on some phenotypic parameters in grey and brown Bengal goats. *Indian J Anim Sci*, 49, 671-671, 1979.

40. Gonzalez-Martinez A, Herrera M, Luque M, Rodero E: Influence of farming system and production purpose on the morphostructure of Spanish goat breeds. *Span J Agric Res*, 12 (1): 117-124, 2014. DOI: 10.5424/sjar/2014121-4673

41. Zaitoun IS, Tabbaa MJ, Bdour S: Differentiation of native goat breeds of Jordan on the basis of morphostructural characteristics. *Small Ruminant Res*, 56, 173-182, 2005. DOI: 10.1016/j.smallrumres.2004.06.011

42. Bourzat D, Souvenir Zafindrajoana P, Lauvergne JJ, Zeuh V: Comparaison morpho-biometrique de chevres au Nord Cameroun et au Tchad (Morphological and biometric comparison of goats in Northern Cameroon and Chad). *Rev Élev Méd Vét Pays Trop,* 46, 667- 674, 1993.

43. Dossa LH, Wollny C, Gauly M: Spatial variation in goat populations from Benin as revealed by multivariate analysis of morphological traits. *Small Ruminant Res*, 73, 150-159, 2007. DOI: 10.1016/j.smallrumres.2007.01.003

44. Chacón E, Macedo F, Velázquez F, Paiva SR, Pineda E, McManus C: Morphological measurements and body indices for Cuban Creole goats and their crossbreds. *R Bras Zootec*, 40, 1671-1679, 2011. DOI: 10.1590/S1516-35982011000800007

45. El Moutchou N, González Martínez AM, Chentouf M, Lairini K, Rodero E: Approach to morphological characterization of northern Morocco goat population. *Cah Options Mediterr*, 108, 427-432, 2014.

46. Orhan I, Duzler A, Alan A, Elmaz O, Ozgel O: Morphological investigations on the hairs of the Honamli and the Hair goat (Black goat). *Kocatepe Vet J*, 11 (2): 173-179, 2018. DOI: 10.30607/kvj.407473