Kafkas Univ Vet Fak Derg
25 (1): 105-110, 2019
DOI: 10.9775/kvfd.2018.20388

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

Research Article

Use of CART and CHAID Algorithms in Karayaka Sheep Breeding

Mustafa OLFAZ [1,a] Cem TIRINK [1,b] Hasan ÖNDER [1,c]

[1] Ondokuz Mayis University, Agricultural Faculty, Animal Science Department, TR-55139 Samsun - TURKEY
[a] ORCID: 0000-0002-0975-3469; [b] ORCID: 0000-0001-6902-5837; [c] ORCID: 0000-0002-8404-8700

Article ID: KVFD-2018-20388 Received: 19.06.2018 Accepted: 25.10.2018 Published Online: 28.10.2018

How to Cite This Article

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

Abstract

The aim of this study was to determine the effect of some factors (sex, birth type, farm type, birth weight and weighting time) on weaning weight through CART and CHAID data mining algorithms. The classification and regression trees are modern analytic techniques that construct tree-based data-mining algorithms. Regression trees are used for the purpose of preliminary selection of the traits affecting the continuous dependent variable. The studied data were consisted of 366 records from Karayaka sheep breed. The CHAID algorithms results revealed that; predictors such as weighting time, sex and farm type statistically influenced weaning weight Regression tree diagram constructed by CART algorithm depicted that birth type was effect the weaning weight, and in this tree weighting time of single born lambs was affected the birth type. The predicted values and original values were correlated (P<0.05). As a result, it could be suggested that CHAID algorithm was found more useful biologically than CART.

Keywords: *CART, CHAID, Karayaka, Weaning weight*

CART ve CHAID Algoritmalarının Karayaka Koyun Islahında Kullanımı

Öz

Bu çalışma, sütten kesim ağırlığı üzerime bazı faktörlerin (cinsiyet, doğum tipi, işletme tipi, doğu ağırlığı ve ölçüm zamanı) CART ve CHAID veri madenciliği algoritmaları ile belirlenmesini amaçlamaktadır. Sınıflandırma ve regresyon ağaçları veri madenciliği kapsamında olan modern analitik yöntemler sınıfında yer almaktadır. Regresyon ağaçları, sürekli bağımlı değişkeni etkileyen özelliklerin ön seçimi amacıyla kullanılmaktadır. Çalışmada Karayaka koyun ırkına ait 366 kayıt veri olarak kullanılmıştır. Sonuç olarak; CHAID algoritmasına göre ölçüm zamanı, cinsiyet ve işletme tipi sütten kesim ağırlığı üzerinde önemli derecede etkili bulunmuştur. CART algoritmasına ait sonuçlar ise sütten kesim ağırlığı üzerine doğum tipinin etkili olduğunu göstermiştir. Bu ağaçta tekiz kuzuların ölçüm zamanının doğum tipinden etkilendiği anlaşılmıştır. Tahmin edilen ve gözlenen değerler yüksek ilişkili bulunmuştur (P<0.05). Sonuç olarak, CHAID algoritmasının CART algoritmasına göre biyolojik olarak daha kullanışlı olduğu belirlenmiştir.

Anahtar sözcükler: *CART, CHAID, Karayaka, Sütten kesim ağırlığı*

INTRODUCTION

In general, the aim of animal breeding is to genetically improve populations of livestock so that they produce more efficiently under the expected future production circumstances. Genetic improvement for economic traits is achieved by selecting the best individuals of the current generation and by using them as parents of the next generation [1]. To achieve the aim of animal breeding, evaluation of the data is very important manner.

Necessary data for animal breeding is generally consisting of many factors which make the data multidimensional. When the number of factor increases, interpretation of the results become difficult. In this case, overlooking the important details will be unavoidable. Fail to satisfy of conventional analysis tools on complex data, new approaches such as data mining have begun to use. Data mining is an entire process of applying a computer-based methodology, including new technologies, to discover knowledge from data [2].

The classification and regression trees are modern analytic techniques which are member of data-mining. They allow for building graphic easily-comprehensible models used to describe and to predict the phenomenon expressed in both the nominal and the ordinal scale. The classification and regression trees can also be used for the purpose

İletişim (Correspondence)
+90 532 3059199
molfaz@omu.edu.tr

of preliminary selection of the traits which have a statistical effect on the dependent variable [3]. Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fit a simple prediction model within each partition. The partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values [4-6].

Some studies reported significant information on usability of data mining algorithms in sheep and goat breeding. In recent times, many studies have been reported on applying CART (Classification and Regression Trees), CHAID (Chi-square Automatic Interaction Detection), exhaustive CHAID and MARS (Multivariate Adaptive Regression Splines) algorithms for various animal species in animal science for regression type problems [7-10]. However, the application of the mentioned algorithms is still scarce for Karayaka sheep. Therefore, in this study, we aimed to compare the CART and CHAID algorithms of classification and regression trees to predict weaning weight by means of the significant ones among sex, birth type, farm type, birth weight and weighting time in Karayaka sheep.

MATERIAL and METHODS

In this study, to predict weaning weight through CART and CHAID tree-based algorithms, 366 records of Karayaka sheep breed were taken from three different farms in the year 2017. Sex, birth type, farm type, birth weight and weighting time (age as days) variables were used as possible explanatory variables to predict weaning weight. Descriptive statistics for dependent and explanatory variables were given in *Table 1*.

This method was modified and extended by using some algorithms to minimize an estimate of misclassification error [11].

The classification and regression tree algorithm contains three important tasks. The first task is how to segment data at each step?, the second task is when to stop segmentation?, and the last one is how to predict the value Y for each X in a segment? [5,11,12]. The classification and regression trees begin with a single root node for response variable. The tree is constructed by splitting the whole data into nodes or sub-groups by using all the independent variables. This process goes on until the requirements of homogeneity are met on any child node [13,14]. It is aimed with obtaining terminal nodes in order to increase proportion of variance among nodes [15].

The most popular algorithms used in decision trees are CART and CHAID tree-based algorithms. These algorithms are nonparametric methods which allow using nominal, ordinal and continuous variables [12]. The term "Regression tree" is used for the tree that its response variable is continuous [16].

The response variable is a continuous and explanatory variables can be continuous or categorical in CART algorithm and it creates binary split. In the CHAID algorithm, the response variable can be continuous and categorical. But, explanatory variables are categorical variables only (can be more than 2 categories) and it can create multiple splits [17,18]. The explanatory variable has continuous structure will turn into categorical structure with use of CHAID algorithm [19]. CART algorithm has a characteristic of continuation of CHAID algorithm [19,20]. The aim is to produce homogeneous data sets as much as possible. CART algorithm produce more homogeneous groups than CHAID using pruning [21,22]. Gini, Twoing and Ordered Twoing impurity or diversity measures can be used for categorical response variables, but for continuous

Table 1. *Descriptive statistics*

Parameter		N	%	Birth Weight (Mean±Std. Deviation)	Weaning Weight (Mean±Std. Deviation)	Weighting Time (Mean±Std. Deviation)
Sex	Male	198	54.1	3.42±0.48	23.28±4.58	100.49±13.76
	Female	198	45.9	3.23±0.54	22.22±4.71	101.43±13.58
	Total	366	100	3.33±0.51	22.79±4.66	100.92±13.66
Birth type	Single	332	90.7	3.45±0.35	23.62±4.05	100.53±13.96
	Twin	34	9.3	2.10±0.09	14.63±0.45	104.70±9.70
	Total	366	100	3.33±0.51	22.79±4.66	100.92±13.66
Farm type	Farm 1	95	26.0	3.12±0.75	21.3±5.37	97.87±11.67
	Farm 2	110	30.0	3.4±0.45	23.89±4.33	107.09±11.32
	Farm 3	161	44.0	3.41±0.34	22.93±4.21	98.51±14.87
	Total	366	100	3.33±0.52	22.79±4.66	100.92±13.67

variables, Least-Squared Deviation (LSD) or Least Absolute Deviation (LAD) can be used [19,23].

The best algorithm produces minimum SD_{Ratio} (Standard Deviation Ratio), MAD (Mean Absolute Deviation), RMSE (Root Mean Square Error) and coefficient of determination (R^2) criteria for goodness of fit. Standard Deviation Ratio, MAD and RMSE can be written as [24,25];

$$SD_{Ratio} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}},$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}},$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

$$R^2 = \left[1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}\right].$$

All statistical analysis was performed using IBM SPSS 20.0 via Ondokuz Mayıs University license. The best algorithm will produce the lowest goodness of fit value for mentioned statistics.

RESULTS

Decision tree diagrams to estimate weaning weight by CART and CHAID algorithms are depicted in *Fig. 1* and *Fig. 2*, respectively. CART was seen to produce more branches compared with CHAID tree-based algorithm. Both diagrams reflected environmental factors that can be effective on weight trait, more understandably.

For CART algorithm, Node 0 was split into two smaller subgroups as single birth (23.629±4.052; n=332) and twin birth (14.634±0.456; n=34) according to birth type. These results showed that the effect of single birth was higher than twins on weaning weight as expected. Nodes derived for single lambs had heavier average than those derived for twin lambs.

For single birth, the branch was divided into two smaller subgroups according to age predictor as Node 3 (the subgroup of the single lambs with the age≤91.5 [18.863±2.545 kg; n=94]) and Node 4 (the subgroup of the single lambs with the age>91.5 (25.512±2.796 kg; n=238).For twin birth lambs, binary branching was occurred with Nodes 5 and 6. Node 5 represented the subgroup of the twin born lambs with the age≤108.0 (14.372±0.362 kg; n=20), whereas the subgroup of the twin born lambs with the age>108.0 (15.008±0.283; n=14) was named Node 6. From *Fig. 1*, it was understood that the heaviest lamb average was obtained with the subgroup of the single lambs with birth weight>3.195 kg and age>112.500 days (28.344 kg).

Pearson correlation coefficient between weaning weight and predicted weaning weight for CART Algorithm was estimated as 0.938 (P<0.01).

For CHAID algorithm weaning weight 22.794±4.663 with n=366 began with Node 0 and the tree had 30 child nodes. The age variable divided into six child nodes at first depth of the tree structure. Predictive performances of the CART and CHAID algorithms are presented in *Table 2*. The influential predictors of the CHAID algorithm were found as age, birth weight and birth type, and farm. Nodes 1, 2, 7, 11, 12, 13, 15, 16, 17, 19, 20-30 were terminal nodes. Node 0 was divided by age predictor into six smaller subgroups i.e. (Nodes 1-6) at the first tree depth. Node 3 was split into three smaller subgroups i.e. (Nodes 7-9) at the second tree depth according to birth weight. Node 4 was branched into two smaller subgroups. Each of Nodes 5 and 6 was divided into four smaller subgroups (Nodes 12-15) and (Nodes 16-19) at the second tree depth according to birth weight. At the third tree depth, Nodes 8-10 were divided into two or three smaller subgroups according to farm. Nodes 14 and 18 were exposed to a binary partition at the third tree depth. At the first tree depth, Node 19 (the subgroup of those with age>114 days birth weight>3.5 kg) produced the heaviest average body weight with 29.327 kg, as expected.

Pearson correlation coefficient between weaning weight and predicted weaning weight for CHAID Algorithm was estimated as 0.937 (P<0.01).

Pearson correlation coefficient between weaning weight and predicted weaning weight for CHAID Algorithm.

DISCUSSION

Results displayed that the partition at the first depth in the tree structures was based on birth type in CART algorithm and weighting age in CHAID algorithm. So, the most predictor variables were different according to the specified algorithms. CART algorithm produced five levels of branching where CHAID algorithm produced three levels of branching. It could be said that CHAID algorithm could be interpreted more easily [12].

For these algorithms, the estimation of SD_{ratio} value smaller than 0.40 was an indicator of the good fit means [12,26,27]. Both MAD and RMSE produced nearly equal values, coefficient of determination were found equal for both algorithms.

To predict weaning weight from sex, birth type, farm type,

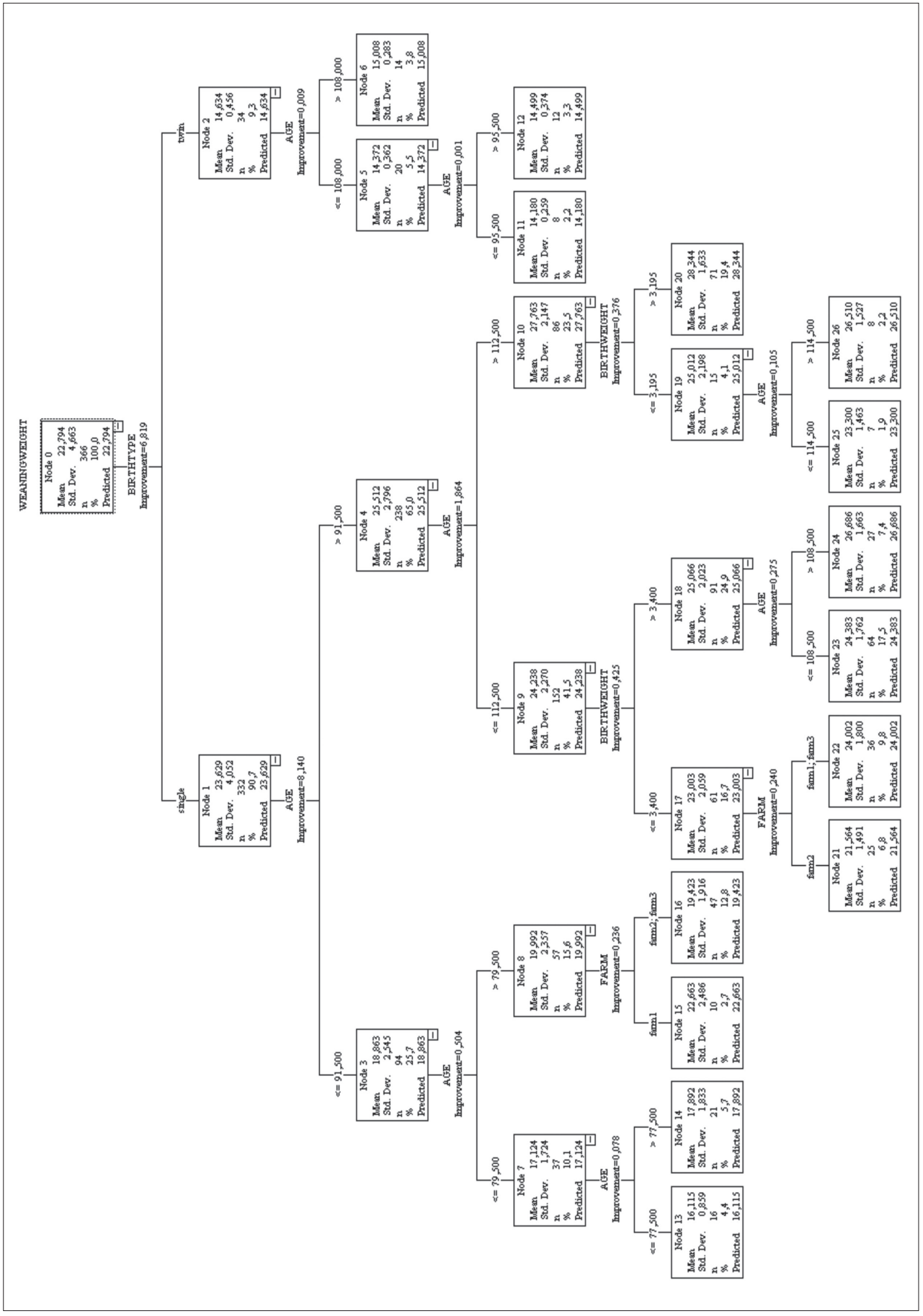

Fig 1. The decision tree diagram obtained by CART

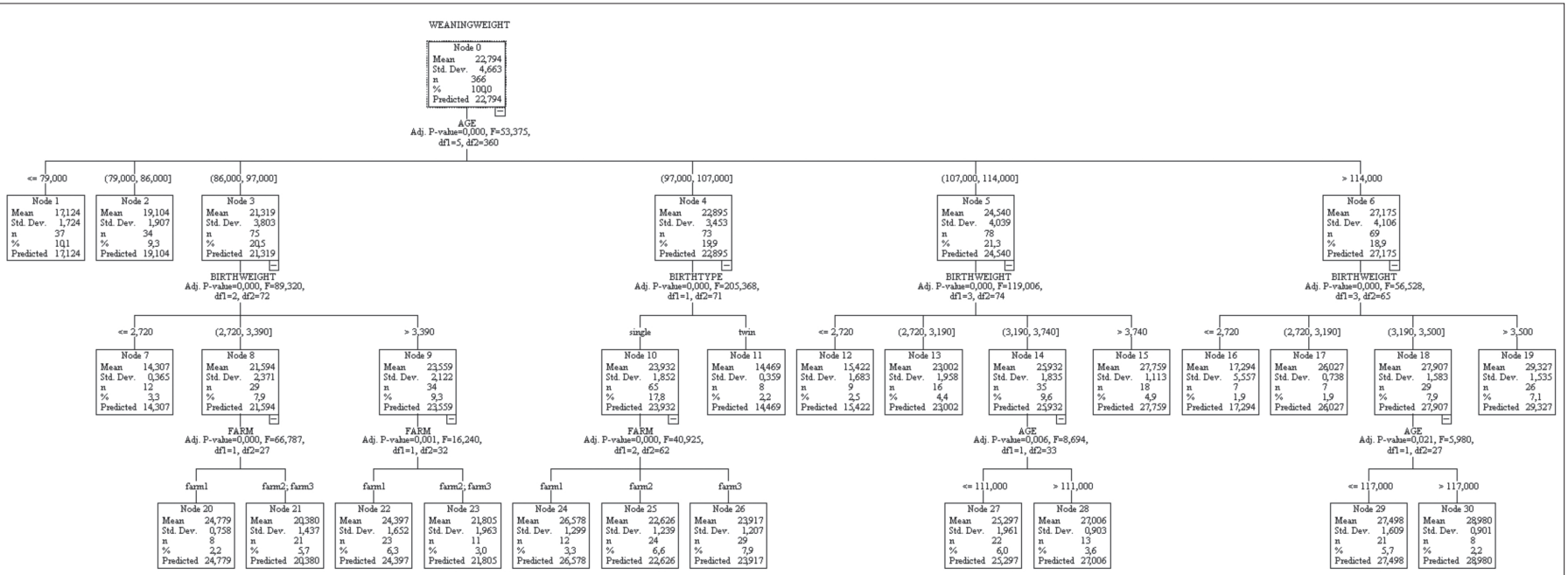

Fig 2. The decision tree diagram obtained by CHAID algorithm

Table 2. *Comparison criteria for CART and CHAID algorithms*

Algorithm	**SD_{ratio}**	**MAD**	**RMSE**	**R^2**
CART	0.347	1.195	1.612	0.88
CHAID	0.348	1.192	1.623	0.88

birth weight and weighting time (age) variables, CART and CHAID algorithms can be used interchangeable. The sex variable was excluded from the model in both algorithms. This result was biologically interesting for sheep breeding. Both two methods had similar fitting criteria; however, interpretation of CHAID algorithm because of less branching was more user friendly than CART algorithm. CART algorithm had sub branching over the same variable, which obstruct the interpretation. Similar findings also supported the statements of some earlier authors [10,12,15,25].

Also, the-tree based algorithms could be applied as a remarkable alternative for the data of response surface designs [10].

Regression tree aims to repeatedly partitioning the population into different child nodes where the variation of response variable is minimum within and maximum between the child nodes. Also, it aims to balance predictive accuracy and complexity with interpretation of model. Advantage of easy interpretation for both response and predictor variables with visual diagrams are superiority of regression tree. Due to nonparametric properties, it does not require any parametric assumptions. Results of this study showed that CART and CHAID algorithms can be used interchangeable.

Acknowledgements

The short summary of this work has been published at the abstract book of International Conference "Agriculture for Life, Life for Agriculture" in Bucharest at June 7-9 2018.

REFERENCES

1. Önder H, Abacı SH: Path analysis for body measurements on body weight of Saanen kids. *Kafkas Univ Vet Fak Derg*, 21 (3): 351-354, 2015. DOI: 10.9775/kvfd.2014.12500

2. Kantardzic M: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley&Sons, Inc. Hoboken, New Jersey, 2011.

3. Dariusz P: Using classification trees in statistical analysis of discrete sheep reproduction traits. *J Cent Eur Agric*, 10 (3): 303-309, 2009.

4. Kayri M, Boysan M: Assesment of relation between cognitive vulnerability and depression's level by using classification and regression tree analysis. *Hacet Üniv Eğit Fak Derg,* 34, 168-177, 2008.

5. Loh WY: Classification and regression trees. *Wires Data Min Knowl,* 1, 14-23, 2011. DOI: 10.1002/widm.8

6. Speybroeck N: Classification and regression trees. *Int J Public Health*, 57, 243-246, 2012. DOI: 10.1007/s00038-011-0315-z

7. Eyduran E, Karakus K, Keskin S, Cengiz F: Determination of factors ınfluencing birth weight using Regression Tree (RT) Method. *J Appl Anim Res*, 34 (2): 109-112, 2008. DOI: 10.1080/09712119.2008.9706952

8. 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). *Pakistan J Zool*, 48, 957-960, 2016.

9. Eyduran E, Zaborski D, Waheed A, Celik Ş, Karadas K, Grzesiak W: Comparison of the predictive capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the indigenous Beetal Goat of Pakistan. *Pakistan J Zool*, 49, 273-282, 2017.

10. Akin M, Eyduran E, Reed, BM: Use of RSM and CHAID data mining algorithm for predicting mineral nutrition of hazelnut. *Plant Cell Tiss Organ Cult*, 128, 303-316, 2017. DOI: 10.1007/s11240-016-1110-6

11. Moghadam MPA, Pahlavani P, Naseralavi: Prediction of car following behavior based on the ınstantaneous reaction time using an ANFIS-CART based model. *Int J Transport Eng*, 4 (2): 109-126, 2016.

12.Ali M, Eyduran E, Tariq MM, Tirink C, Abbas F, Bajwa MA, Baloch MH, Nizamani AH, Waheed A, Awan MA, Shah SH, Ahmad Z, Jan S: 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 J Zool*, 47, 1579-1585, 2015.

13. Mendeş M, Akkartal E: Regression tree analysis for predicting slaughter weight in broilers. *Ital J Anim Sci*, 8, 615-624, 2009. DOI: 10.4081/ijas.2009.615

14. Oruçoğlu O: Determination of envitronmental factors affecting 305-day milk yield of holstein cows by regression tree method. *MSc Thesis,* Süleyman Demirel University, Enstitue of Applied and Natural Science, 2011.

15. Koc Y: Application of regression tree method for different data from animal science. *MSc Thesis*, Igdir University, 58, 2016.

16. Eyduran E, Karakus K, Keskin S, Cengiz F: Determination of factors influencing birth weight using regressiontree (RT) method. *J Appl Anim Res,* 34, 109-112, 2008. DOI: 10.1080/09712119.2008.9706952

17. Celik S, Yilmaz O: Comparison of different data mining algorithms for prediction of body weight from several morphological measurements in dogs. *J Anim Plant Sci,* 27, 57-64, 2017.

18. Jarošík V: CART andrelated methods. **In,** Simberloffand D, Rejmánek M (Eds): Encyclopaedia of Biological Invasions. 104-108, University of California Press, Berkeley and Los Angeles, 2011.

19. Gevrekci Y, Takma C: A Comparative study for egg production in layers by decision tree analysis. *Pakistan J Zool*, 50 (2): 437-444, 2018.

20. Cimenli S: Churn analysis and prediction with decision tree and artificial neural network. *Graduate Thesis,* Kadir Has Unıversity Graduate School of Science and Engineering, 2015.

21. Alkhasawneh MS, Ngah UK, Tay LT, Isa NAM, Al-Batah MS: Modeling and testing landslide hazard using decision tree. *J Appl Math*, 2014:929768, 2014. DOI: 10.1155/2014/929768

22. Kurt I, Ture M, Kurum AT: Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl*, 34 (1): 366-374, 2008. DOI: 10.1016/j.eswa.2006.09.004

23. Yadav SK, Bharadwaj B, Pal S: Data mining applications: A comparative study for predecting students' performance. *Int J Inn Tech Crea. Eng*, 1, 13-19, 2011.

24. Celik S, Eyduran E, Karadas S, Tariq MM: Comparison of predictive performance of data mining algorithms in predicting body weight in Mengali rams of Pakistan. *R Bras Zootec*, 46 (11): 863-872, 2017. DOI: 10.1590/s1806-92902017001100005

25. Koc Y, Eyduran E, Akbulut O: Application of regression tree method for different data from animal science. *Pakistan J Zool*, 49 (2): 599-607, 2017. DOI: 10.17582/journal.pjz/2017.49.2.599.607

26. Grzesiak W, Lacroix R, Wójcik J, Blaszczyk P: A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. *Can J Anim Sci*, 83, 307-310, 2003. DOI: 10.4141/A02-002

27. Grzesiak W, Zaborski D: Examples of the use of data mining methods in animal breeding. **In,** Karahoca A (Ed): Data Mining Applications in Engineering and Medicine *InTech*, 303-324, Rijeka, Croatia. DOI: 10.5772/50893

28. Akin M, Hand C, Eyduran E, Reed BM: Predicting minor nutrient requirements of hazelnut shoot cultures using regression trees. *Plant Cell Tiss Organ Cult*, 132, 545-559, 2018. DOI: 10.1007/s11240-017-1353-x