ISBN 978-989-9121-39-3
36th AMSTERDAM Int'l Conference on "Chemical, Biological, Environmental & Medical Sciences" (ACBEMS-24)
Amsterdam (Netherland) June 26-28, 2024

Estimation of Body Weight from Body Measurements with CART Algorithm for Mexican Hair Sheep

Alfonso J. Chay-Canul[1], Cem Tırınk[2*]

[1]División Académica de Ciencias Agropecuarias, Universidad Juárez Autónoma de Tabasco, Carr, Villahermosa-Teapa, km 25, Villahermosa CP 86280, Tabasco, Mexico.

[2]Igdir University, Faculty of Agriculture, Department of Animal Science, TR76000, Iğdır, Türkiye.

***Abstract*:** *This study aimed to estimate body weight from body measurements. For this aim, forty animals were used in the study. CART data mining algorithm was used to estimate the body weight from various features. To evaluate the model, several goodness-of-fit criteria, such as determination of coefficients, mean absolute error and root mean square error, were used. The results of this study showed that the CART algorithm may help the sheep breeders improve characteristics of great importance. In this way, the sheep breeders can get an exclusive population and determine which characteristics are significant for estimating the body weight of the herd in Mexico.*

Keywords: *Data Mining, CART, Sheep, Body Weight, Mexican Hair Sheep*

1. Introduction

Sheep play an important role in obtaining animal products and developing the rural economy among civilizations [1,2]. In this context, when evaluated in terms of products obtained from animal breeding, sheep breeding is a multi-purpose ruminant animal that is very valuable not only in developing healthful societies but also in the development of rural economies obtaining from animal products such as milk, meat and fleece [3,4].

Increasing the animals' genetic potential is important not only for animal breeders but also for increasing the potential of products of animal origin for the whole country. In this context, many breeding types of research that will raise the genetic potential to be accomplished gain great importance. In this regard, various statistical methods can be used in stating the breeding studies. In sheep breeding, knowing the live weights of the animals in the flock is very important in determining the breeding strategy and herd management [5]. Considering that multivariate statistical methods can be used to estimate factors that are difficult to measure from factors that are easier to measure, it would be a correct application to accurately determine sheep's live weight.

This study aimed to define the advantages of the CART algorithm for predicting body weight from body measurements for Mexican hair sheep.

2. Material Methods

The animals used in this study were treated following the guidelines and regulations for ethical animal experimentation of the División Académica de Ciencias Agropecuarias of the Universidad Juárez Autónoma de Tabasco (ID UJAT-CIEI-2023-084). The experiment was conducted at the Southeastern Centre for Ovine Integration (Centro de Integración Ovina del Sureste [CIOS]; 17° 78" N, 92° 96" W; 10 masl), located 25 km from the Villahermosa-Teapa road in the town of Alvarado Santa Irene 2nd Section, Tabasco State, Mexico.

The trial population consisted of 40 Hair sheep animals. The ewes were grouped and housed in a cage system with raised slatted floors and group feeding (ten animals per cage) for 40 days. Animals received a diet consisting of 12 MJ/kg dry matter (DM) and 10% crude protein (CP) with 66% forage and 34% concentrate according to AFRC [6]. Dietary ingredients included cereal grains (19 % ground corn, 11% soybean meal, 66% star grass hay, 3% molasses, and 1 %premix of vitamins and minerals.

The BW and BM were recorded in 40 animals that ranging from two to three years in age. The BM was measured using a flexible tape fiberglass (Truper®, Truper S.A. de C.V., San Lorenzo, Mexico) and a 65-cm calliper (Haglof®, Sweden). The BM registered were as follows: heart girth (HG), abdominal girth (AG), body length (BL), diagonal BL (DBL), withers height (WH), rump height (RH), and hip width (HW), thorax width (TW), abdominal width (AW) and rib depth (RD) according to described by [7].

Decision tree procedures such as CART, Exhaustive-CHAID, and CHAID can be used for modelling the quantitative characteristics [8-12]. In this context, the examined algorithm in the current study, the classification and regression tree (CART) algorithm proposed by [13]. The CART algorithm aims to create a binary tree structure by iteratively dividing a node into two child nodes. In this regard, this algorithm includes a series of splits like tree branches that occur in the significant explanatory variables until many homogeneous nodes are obtained that provide the minimum error variance covered by the data set.

In addition, the CART algorithm is an algorithm that continues the process until new and homogeneous two-piece partitions are obtained. In this algorithm, each split is for only one explanatory variable, and in obtaining the tree structure. During one of these processes, the pruning criterion was specified depending on the variance. For this study, the stopping criterion was specified such that the minimum size of the tree node was 5. It depends on the size of the sample. Additionally, a 10-fold cross-validation procedure with standard error correction was used to obtain a suitable tree structure, ensuring that no overfitting problem occurred for the CART algorithm.

The following goodness-of-fit criteria used to evaluate the performances of CART algorithm [14-17]:

1. Coefficient of determination (R^2):

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

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

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

3. Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

where n is the number of sample size for training set, k is the number of parameters of the model, yi is the actual value of the response variable, $\hat{y}_i$ is the predicted value for response variable, In the evaluation to be made within the scope of goodness-of-fit criteria, RMSE, MAE and R^2 used. In addition, the relationship between actual and predicted response variable evaluated with Pearson's correlation coefficient.

All statistical processes were accomplished by using R software [18]. For information about the sheep data set, the descriptive statistics are utilized by using the "psych" package in R software [19]. The "caret" package was used for performing the CART (Decision tree) algorithm [20]. The model performances were shown by using the "ehaGoF" package [21].

3. Results

Response and explanatory characteristics of Hair sheep breed are presented in Table 1. This table contains statistical summaries of BW and various body measurements (HG, AG, BL, etc.) obtained in a study of 40 sample sizes. The mean value for BW is 33.38, the standard deviation is 4.69 and the values range from 22.3 to 42.7. Means, standard deviations, and ranges for other body measurements vary depending on the type of measurement. For example, HG has a mean value of 75.69 and a standard deviation of 5.29, while AG has a mean value of 78.9 and a standard deviation of 7.27. These data show the central tendency and variability of body weight and measurements of the creatures examined within the scope of the research. Having a similar number of samples for each variable provides a statistically consistent basis for comparisons between variables. This information is especially valuable for developing a general understanding of the physical characteristics of living things and their distribution.

TABLE I: Descriptive statistics of response and explanatory variables

	n	Mean	Std. Deviation	Min	Max
BW	40	33.38	4.69	22.3	42.7
HG		75.69	5.29	64	85
AG		78.9	7.27	63	92
BL		42.17	2.18	38	49
DBL		50.98	3.16	44	56
WH		61.99	4.00	54	71
RH		62.79	3.62	57	71
HW		15.91	1.67	13	19
TW		17.36	1.45	13.5	20
AW		22.03	2.53	14	27
CD		26.43	2.76	22	32

The relationship between response and explanatory variables are presented with the Pearson's correlation coefficients in Figure 1. According to Figure 1, This correlation chart provides a detailed statistical analysis showing how different body measurement variables relate to each other. The correlation between BW and HG has the highest value (0.841), indicating a strong and positive relationship between them. That is, when the value of one increase, the other tends to increase similarly. While correlation coefficients vary for each pair of variables, some (e.g., between BW and HG, AG, WH, and CD) exhibit quite high and significant relationships. On the other hand, there is a very low correlation (0.110) between RH (presumably a circumference measurement) and HW (possibly a height measurement), indicating a weak relationship between them. While these coefficients express the strength and direction of the linear relationship of the variables with each other, asterisks (*, **, ***) indicating statistical significance are ranked according to the low p-values. Histograms showing the unique distribution of each variable and scatter plots showing the distribution of each pair of variables make these correlations more understandable by visualizing the distribution shape and density of the data. This contains valuable information, especially for researchers who want to develop an in-depth understanding of the nature of the relationship between two variables and how one variable may potentially influence the other.

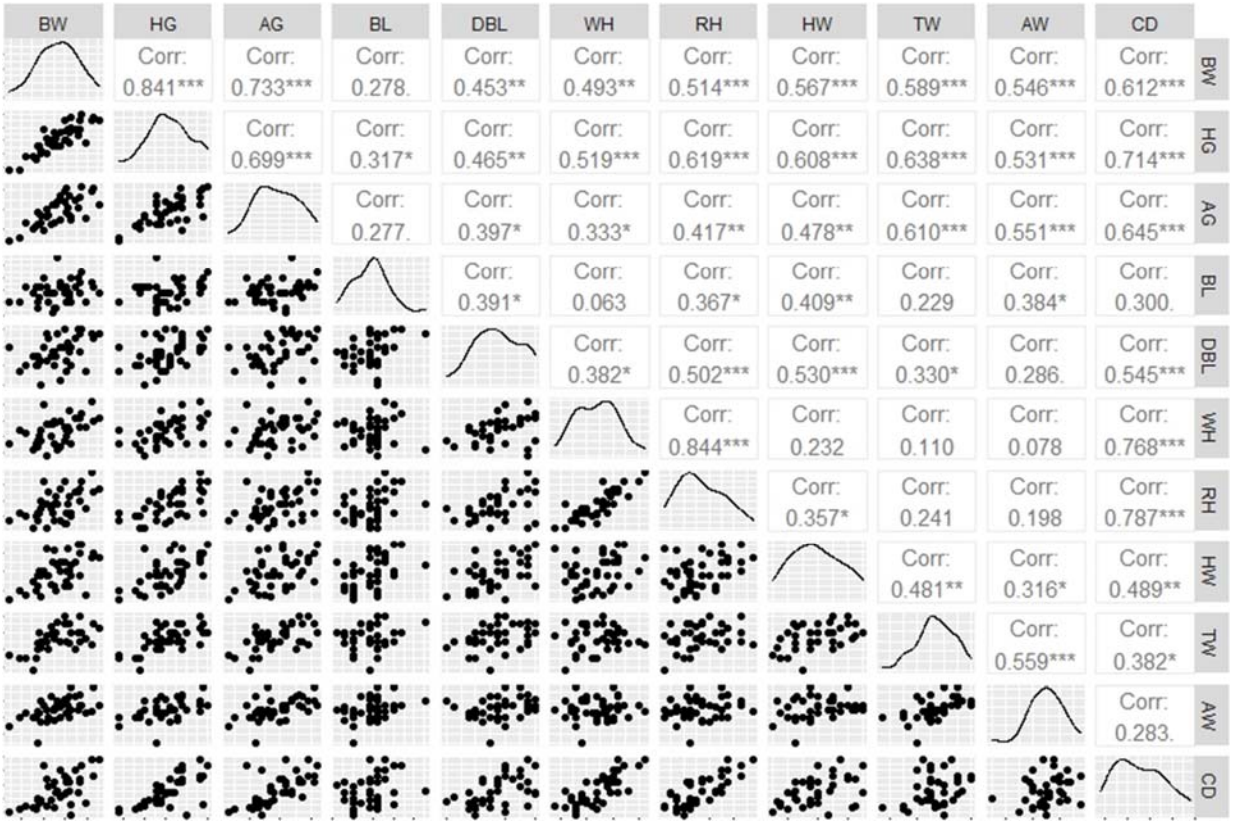

Fig. 1. The correlation coefficients of whole variables.

In Table 2, it contains the complexity parameter (CP), tree branches (nsplit), relative error (rel error), cross-validation error (xerror) and cross-validation standard deviation (xstd) values obtained from a Classification and Regression Tree (CART) analysis. Lower values of CP indicate that the model is less complex; As nsplit increases, that is, as the tree branches more, the model becomes more complex. Relative error indicates how well the model fits the training data set; lower values mean a better fit. xerror and xstd are used to evaluate the performance of the model on new data and the consistency of that performance. Ideally, as the CP value increases, both the relative error and the cross-validation error should decrease. However, here it seems that xerror values increase or decrease in some cases as nsplit increases; This indicates to us that the tree may show overfitting after a certain point. This can be observed when the minimum xerror value reaches 7 branches (nsplit) and the error rate increases or remains constant with more branches. This is important in determining the point at which the model has optimal complexity and can be used to prune the tree. For example, after the 7th branch, the decrease in xerror value is very low or increases, indicating that further branches do not improve the model. This also provides insight into what level of branching should be chosen to prevent possible overfitting of the tree.

Figure 2 reflects the relative importance of various predictor variables within a decision tree model. As the dominant predictor in the model, the AG variable is observed to have a relative importance of around 18%, indicating that it has the greatest impact on the predictive ability of the model. Secondly, the CD variable has a relative importance of 16%, and these two variables appear to have a more significant weight than all other predictors in the model. On the other hand, WH, AW, HW and TW variables are moderately important in estimating the model and contribute with importance percentages varying between approximately 8% and 10%. DBL, BL and RH variables have relatively low importance in the model, and the RH variable stands out as the least important predictor of the model with a very low percentage of 2%. This distribution of relative importance can play a critical role in determining the variables that need to be addressed in further optimizing the model.

These findings are important in terms of the understandability and interpretability of the model and shed light on which variables should potentially be weighted more in future studies.

TABLE II: Complexity parameter of the CART algorithm.

	CP	nsplit	rel error	xerror	xstd
1	0.520571	0	1	1.075071	0.216128
2	0.118202	1	0.479429	0.546281	0.103232
3	0.086517	2	0.361228	0.670305	0.115439
4	0.043003	3	0.274711	0.656931	0.126407
5	0.040931	5	0.188705	0.695296	0.152101
6	0.030787	6	0.147774	0.724866	0.151261
7	0.021741	7	0.116986	0.620804	0.140634
8	0.013825	8	0.095246	0.567489	0.106225
9	0.013738	10	0.067596	0.578173	0.107209
10	0.01	11	0.053858	0.57751	0.107292

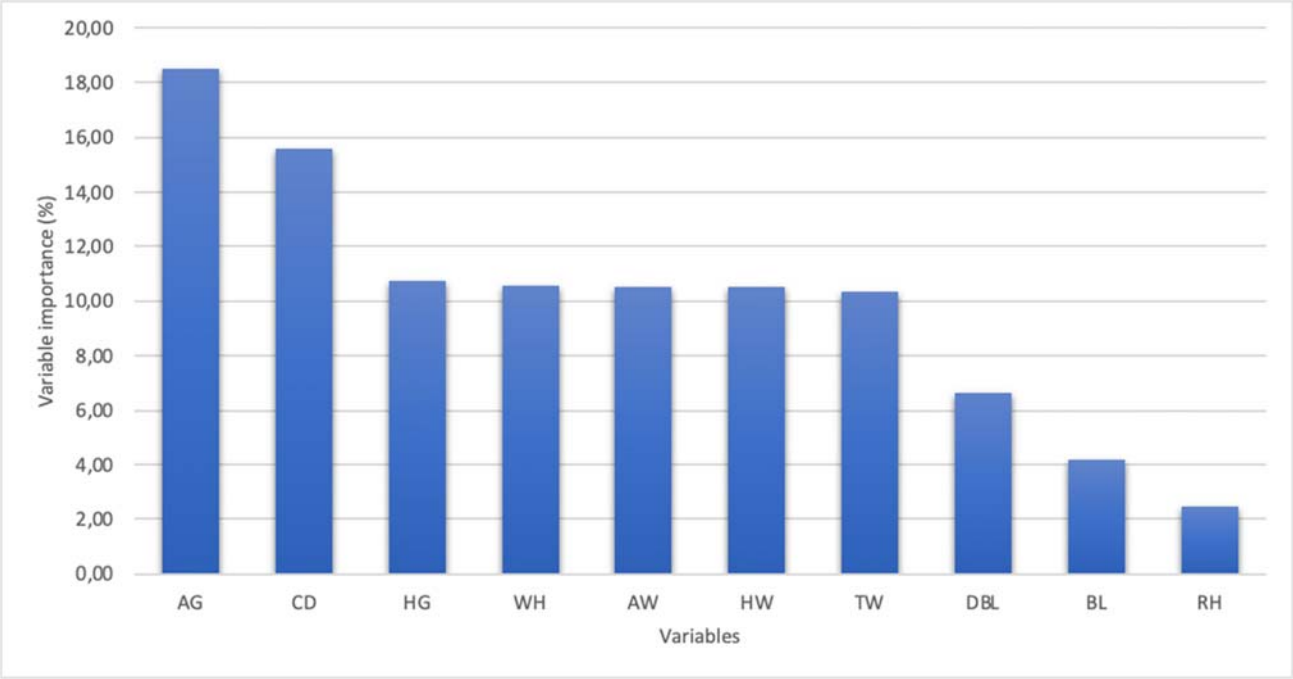

Fig. 2. Variable importance

Figure 3 represents a Classification and Regression Tree (CART) model created for Body Weight (BW) estimation. The tree is branched to predict categorized values of BW using the predictor variables (HG, TW, AG, DBL, AW, WH) in the dataset. At the root node, the division begins with whether the HG variable is less than 74 units, representing 100% of the observations. A yes answer (HG < 74) leads to the left branch of the tree, and a no answer (HG ≥ 74) leads to the right branch. Each split shows the percentage of the number of observations in the data set and the proportion of observations within the branch. The left branch branches further depending on whether the variable TW is less than 16 units, while the right branch branches based on whether HG is less than 82 units. This branching process continues until we reach more homogeneous subsets of the data set for each group, subdivided according to the threshold of a particular variable. The final leaf nodes provide the estimated percentile distribution of BW for a subset of observations in the data set. For example, observations satisfying the conditions HG < 66 and TW < 16 constitute the smallest subset of 5% of the predicted value of BW, while observations satisfying the conditions HG < 74, AG < 74, and WH < 65 constitute the larger subset of 15%.

TABLE III: Goodness of fit criteria for CART algorithm

Criterion	value
Root mean square error (RMSE)	1.075
Pearson's correlation coefficients (PC)	0.973
Mean absolute percentage error (MAPE)	2.578
Coefficient of determination (R^2)	0.946

According to the results of this study, BW predictions made by the CART model have a root mean square error (RMSE) value of 1.075, a Pearson correlation coefficient (*r*) value of 0.973, a mean absolute error (MAPE) value of 2.578 and a coefficient of determination (R^2) The value was evaluated as 0.946 in Table 3. High *r* and R^2 values indicate that the model's predictions have a strong positive correlation with the actual values and predicted values of the response variable in the data. Low RMSE and MAPE values also indicate that the model's predictions are generally accurate and reliable. These results indicate that the CART model is successful in BW predictions and fits the data.

4. Discussion

In the present study, the correlation between BW and HG has the highest value (0.841), indicating a strong and positive relationship between them ($P < 0.01$). Sabbioni et al. [22] reported to be the significant influence of sex, HW, ChC, BL, HCr, ChW, ChD, and CrW on live body weight in Cornigliese sheep and produced greater correlation coefficient of 0.871 between BL and LBW for female sheep in comparison with those found for males. Kunene et al. [23] reported that the correlation coefficients were determined significant for BW-HG (0.75) and BW-WH (0.67) in one pair of incisors as well as ($P < 0.01$) for BW-HG (0.65) and BW-WH (0.45) in three and four pairs of incisors in Nguni female sheep.

Compared with our present study's R^2 value, Ali et al. [7] have estimated lower R^2 of CHAID (0.8377 Exhaustive CHAID (0.8421), CART (0.82644), and ANN (0.81999) for BW prediction of yearling Harnai sheep, and Yakubu [24] estimated much lower R^2 (0.62) of CART for BW prediction with CG and FL in Uda sheep of Nigeria. Similary, Khan et al. [25] reported the lower R^2 value of 0.844 with Exhaustive CHAID algorithm for Harnai sheep in comparison with our present R^2 values reported here for Thalli sheep.

In breeding Mexican Hair sheep, it is vital to determine body sizes that contribute to increased body weight (BW) to strengthen the rural economy and improve selection processes. Many studies on the relationship between BW and body measurements have been reported in the literature, and these measurements are effective in defining breed characterization and detecting body measurements affecting BW as indirect selection criteria with the goal of increasing meat productivity. For this purpose, advanced statistical methods such as data mining and Artificial Neural Networks (ANN) can be used reliably instead of classical statistical techniques, which are unreliable if basic assumptions are violated.

Consequently, the use of CART predictive modeling may allow to obtain an elite population of Mexican Hair sheep adapted to regional conditions and to identify body measurements that contribute to increasing BW as indirect selection criteria. This modeling not only explains breed characterization and improves herd management standards, but also supports sustainable meat production and regional development.

5. References

[1] 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*, 863-872. https://doi.org/10.1590/s1806-92902017001100005

[2] Ağyar, O.; Özköse, E.; Ekinci, M.S. and Akyol, İ. (2020). Investigation of live weight measurements of morkaraman lambs according to various times in terms of different variables. *Black Sea Journal of Agriculture*, *3*, 193–199.

[3] 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. *Brazilian Journal of Animal Science*, 46(11), 863-872.

https://doi.org/10.1590/s1806-92902017001100005

[4] Tırınk, C., Tosun, R., Saftan, M., Kaya, E. and Atalay, A. İ. (2022). Prediction of Birth Weight from Body Measurements with the CART Algorithm in Morkaraman Lambs. *Large Animal Review*, *28*(4), 187-192.

[5] Sakar, Ç.M., Ünal, İ., Okuroğlu, A., Coşkum, M.İ. and Zülkadir, U. (2020). Prediction of live weight from chest girth from birth to 12 months of age in Yerli Kara cattle, *Black Sea Journal of Agriculture* 3, 200–204.

[6] AFRC. Agricultural and Food Research Council. 1993. Technical Committee on Responses to Nutrients. Energy and Protein Requirements of Ruminants. Wallingford, UK: CAB International.

[7] Bautista-Díaz, E., Mezo-Solis, J.A., Herrera-Camacho, J., Cruz-Hernández, A., Gomez-Vazquez, A., Tedeschi, L.O., Lee-Rangel, H.A., Bello-Pérez, E.V., Chay-Canul, A.J., 2020. Prediction of carcass traits of hair sheep lambs using body measurements, Animals, 10, e1276.

https://doi.org/10.3390/ani10081276

[8] Ali M., Eyduran E., Tariq M.M., Tirink C., Abbas F., Bajwa M.A., Baloch M.H., Nizamani A.H., Waheed A., Awan M.A., Shah S.H., Ahmad Z., Jan S., 2015. Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai sheep, Pakistan Journal of Zoology, 47, 1579-1585.

[9] Eyduran E., Zaborski D., Waheed A., Celik S., Karadas K., Grzesiak W., 2017. Comparison of the Predictive Capabilities of Several Data Mining Algorithms and Multiple Linear Regression in the Prediction of Body Weight by Means of Body Measurements in the Indigenous Beetal Goat of Pakistan, Pakistan Journal of Zoology, 49(1), 257-265.

https://doi.org/10.17582/journal.pjz/2017.49.1.257.265

[10] Akin, M., Eyduran, E., Niedz, R.P., Reed, B.M., 2017. Developing hazelnut tissue culture free of ion confounding, Plant Cell, Tissue and Organ Culture 13(3), 483–494.

https://doi.org/10.1007/s11240-017-1238-z

[11] Akin, M., Eyduran, E., 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(2), 303–316.

https://doi.org/10.1007/s11240-016-1110-6

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

https://doi.org/10.1007/s11240-017-1180-0

[13] Breiman L., Friedman J.H., Olshen R.A., Stone C.J., 1984. Classification and regression trees. Chapman and Hall, WadsworthInc., New York, NY, USA.

[14] Grzesiak, W., Zaborski, D., 2012. Examples of the use of data mining methods in animal breeding. In: Data mining applications in engineering and medicine (ed. A Karahoca). InTech, Rijeka, Croatia, in IntechOpen 303–324.

https://doi.org/10.5772/50893

[15] Eyduran E., Akin M., Eyduran S.P., 2019. Application of Multivariate Adaptive Regression Splines through R Software, Nobel Academic Publishing, Ankara.

[16] Olfaz M., Tirink C., Onder H., 2019. Use of CART and CHAID algorithms in Karayaka sheep Breeding, Kafkas Universitesi Veteriner Fakultesi Dergisi, 25(1), 105-110.

[17] Zaborski, D., Ali, M., Eyduran, E., Grzesiak, W., Tariq, M.M., Abbas, F., Waheed, A., Tirink, C., 2019. Prediction of selected reproductive traits of indigenous Harnai sheep under the farm management system via various data mining algorithms, Pakistan Journal of Zoology, 51, 421–431.

https://doi.org/10.32614/CRAN.package.ehaGoF

[18] R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[19] Revelle, W., 2022. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.9.

[20] Kuhn, M., 2022. caret: Classification and Regression Training_. R package version 6.0-93, <https://CRAN.R-project.org/pack age=caret >.

[21] Eyduran, E., 2020. ehaGoF: Calculates Goodness of Fit Statistics_. R package version 0.1.1,https://CRAN.R-project.org/package=ehaGoF

https://doi.org/10.32614/CRAN.package.ehaGoF

[22] Sabbioni, A., Beretti, V., Superchi, P., Ablondi, M., 2020. Body weight estimation from body measures in Cornigliese sheep breed, Italian Journal of Animal Science, 19:1, 25-30.

https://doi.org/10.1080/1828051X.2019.1689189

[23] Kunene, N.W., Nesamvuni, A.E., Nsahlai, I.V., 2009. Determination of prediction equations for estimating body weight of Zulu (Nguni) sheep. Small Ruminant Research, 84, 41-46.

https://doi.org/10.1016/j.smallrumres.2009.05.003

[24] Yakubu, A., 2012. Application of regression tree methodology in predicting the body weight of Uda sheep. Animal Science and Biotechnologies, 45 (2), 484-490.

[25] Khan, M.A., Tariq, M.M., Eyduran, E., Tatliyer, A., Rafeeq, M., Abbas, F., Rashid, N., Awan, M.A., Javed, K., 2014. Estimating body weight from several body measurements in Harnai sheep without multicollinearity problem, The Journal of Animal & Plant Sciences, 24(1), 120-126.