Pakistan J. Zool., vol. 47(6), pp. 1579-1585, 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

Muhammad Ali,[1] Ecevit Eyduran,[2] Mohammad Masood Tariq,[1]* Cem Tirink,[2] Ferhat Abbas,[1] Masroor Ahmad Bajwa,[1] Muhammad Haroon Baloch,[3] Abdul Hussain Nizamani,[4] Abdul Waheed,[4] Muhammad Arif Awan,[1] Syed Haider Shah,[5] Zafar Ahmad[1] and Saadullah Jan[1]

[1]Center for Advanced Studies in Vaccinology and Biotechnology, University of Balochistan, Quetta, Balochistan, Pakistan

[2]Biometry Genetics Unit, Department of Animal Science, Faculty of Agriculture, Igdir University, Igdir, Turkey

[3]Faculty of Veterinary Science, Sindh Agriculture University, Tandojam, Pakistan

[4]Faculty of Veterinary Science, Bahauddin Zakariya University, Multan, Pakistan

[5]Department of Statistics, University of Balochistan, Quetta, Balochistan, Pakistan

Abstract.- The present investigation deals with comparing performance of different data mining algorithms *viz*., CHAID, EXHAUSTIVE CHAID, CART and ANN, in order to predict body weight from biometrical characteristics for a total of 161 Harnai lambs changing in age between 6-9 months. Minimum numbers of lambs in parent and child nodes were set 20:10 for first three algorithms in order to improve their predictive ability. Considered as an outcome variable, body weight were predicted from explanatory variables, linear body measurements; namely, body length (BL), withers height (WH), chest girth (CG), paunch girth (PG), face length (FL), length between ears (LBE), length of ears (EARL), width (FTW) and length (FTL) of tail, and sex factor. To find out the best one among aforenamed four algorithms, model quality criteria like coefficient of determination (R^2%), adjusted coefficient of determination (Adj-R^2%), coefficient of variation (CV%), SD ratio (the ratio of SD of errors to SD of the actual dependent variable values), Root Mean Square Error (RMSE), Relative Approximation Error (RAE) and the Pearson correlation between actual and predicted values were estimated. All the qualified algorithms were very sufficient; therefore, the use of Exhaustive CHAID in the prediction of body weight presented the best fit in both model quality criteria and more proper decision tree diagram visually, which may provide some advantageous in exhibition of some breed standards of the Harnai sheep.

Keywords: Artificial neural network, CART, CHAID, exhaustive CHAID, Harnai sheep, weight prediction.

INTRODUCTION

In sheep production systems, it is exceedingly difficult to predict body weight playing an important role for deciding suitable feed amounts and drug doses of sheep in village conditions where scale and routine records are unavailable (Mahmud *et al.*, 2014). In sheep breeding, the illumination of the association between live weight and morphological characteristics is gaining importance for determining standards of sheep breeds and for choosing more productive sheep in respect of amending meat yield level of sheep and performing conservation objectives. In the genetic characterization of indigenous sheep breeds, the attainment of selection studies conducted to select better sheep at growth period is dependent on the existence of the genetic correlations between live weight and morphological characteristics; in fact, which are postulated as indirect selection criteria in sheep breeding studies (Jahan *et al.*, 2013; Mohammad *et al.*, 2013). For justifiable reasons, several studies in the characterization of indigenous sheep populations have emphasized the identification of original breed standards (Mohammad *et al.*, 2012).

To find the prediction equation of live weight from morphological linear characteristics; some common statistical techniques (simple linear regression, multiple linear regression, ridge

* Correspondence address: tariqkianiraja@hotmail.com
0030-9923/2015/0006-1579 $ 8.00/0

regression and path analysis) have been mostly applied. In these classic statistic techniques, if there were very strong correlations greater than 0.80-0.90 among the independent variables, the biased parameter estimation is available due to multicollinearity problem for each morphological characteristic and it is extremely difficult to accurately interpret the influence of morphological characteristics on the weight (Yakubu, 2009). In recent years, notwithstanding, the most adopted techniques (namely, robust regression (Cankaya, 2009), use of factor scores in multiple linear regression (Cankaya *et al.,* 2009; Eyduran *et al.,* 2009; Khan *et al.,* 2014), use of principal component scores in multiple linear regression (Khan *et al.,* 2014) and regression tree method (Mohammad *et al.,* 2012; Khan *et al.,* 2014) have been made allowance in the prediction with a great accuracy due to many desirable features as investigators take an eager interest in choosing the most influential statistical techniques eradicating the multicollinearity problem and thus giving the best prediction of the body weight as a target characteristic, of great economic magnitude (Eyduran *et al.,* 2009).

In addition to using factor and principal component scores in multiple regressions, data mining methods such as, CART (Yakubu, 2009), CHAID and Exhaustive CHAID (Khan *et al.,* 2014) have been come prominence for predicting the body weight from morphological linear characteristics in sheep. Data mining algorithms obtaining homogenous sub-groups as soon as possible are not affected by multicollinearity problem existing strong correlations between morphological characteristics, outliers and missing data (Mendes and Akkartal, 2009). Referred as a powerful data mining algorithm, artificial neural network (ANN) structurally resembles human brain and are manipulated for ordinal, nominal, scale dependent variables, like in the former data mining algorithms.

In the literature, there were several described studies for eliminating multicollinearity problem about the prediction of the weight from morphological characteristics through applying CHAID and Exhaustive CHAID algorithms (regression tree methods) (Mohammad *et al.,* 2012; Yakubu, 2012; Khan *et al.,* 2014) in addition to using factor and principal component scores in the multiple regression analysis (Khan *et al.,* 2014), in the recent years. In fact, implementation of Artificial Neural Network among the data mining algorithms is still insufficient for the weight prediction (Grzesiak *et al.,* 2014), in contravention of 305-d milk yield prediction (Grzesiak *et al.,* 2006; Gorgulu, 2012; Grzesiak and Zaborski, 2012, Takma *et al.,* 2012; Shahinfar *et al.,* 2012). Therefore, in order to ideally exhibit the causal relationship of body weight by morphological characteristics in sheep, there is lack of knowledge about evaluating performance of several data mining methods. Due to these reasons, the goal of the present research was to assess performance of some data mining algorithms; namely, CART, CHAID, Exhaustive CHAID and Artificial Neural Network, with a view to predict the post weaning live weight with their morphological linear characteristics for the Harnai lambs at 6-9 months of age, which was genetically associated with reproduction characteristics. The current results may be proven helpful in the classification of superior of Harnai lambs for early selection and in describing the breed standards of Harnai sheep.

MATERIALS AND METHODS

Animal data

In the present work, 161 Harnai indigenous lambs (71 male lambs + 90 female lambs) at 6-9 months of age were provided for predicting live weight (Kg) via the morphologically obtained quantitative characteristics (cm), body length (BL), withers height (WH), chest girth (CG), paunch circumference (PC), face length (FL), length between ears (LBE), ear length (EARL), fat tail length (FTL) and fat tail width (FTW), respectively. For comparing four algorithms, the current data were taken from a very limited part of the data of Khan *et al.* (2014).

The data mining methods mostly establish homogenous subgroups in visual decision trees as far as possible from the presented data set. CART algorithm constructs a binary decision tree by splitting a node into two child nodes recursively. As a recursive partitioning method, CHAID algorithm selects explanatory variables powerfully interacted

by outcome (response=dependent) variable and constructs the decision trees with multi way node splitting, like also in Exhaustive CHAID algorithm. Analyzing all conceivable splits in terms of each explanatory variable, Exhaustive CHAID algorithm is a modified form of CHAID data mining algorithm. In addition, Artificial Neural Network (ANN) with one hidden layer on the basis of Multilayer Perceptron, also addressed as a feed-forward neural network, is applied to probe non-linear causal relationships between explanatory variables of complex data sets, which is a powerful data mining algorithm inspired from human brain and in the structure of the ANN, there are three layers of input, hidden, output (Gorgulu, 2012). The data were at random split into three clusters, the training set (70%), the verification set (20%) and the test set (10%).

Multi-way node splitting is available in CHAID (Chi-Squared Automatic Interaction Detection) and Exhaustive CHAID, but binary node splitting is only available for CART (Classification and Regression Trees) algorithm. Bonferroni method is employed to obtain adjusted significance values for merging and splitting criteria. By default, minimum tree depths were specified for CHAID (3), Exhaustive CHAID (3) and CART (5) algorithms, respectively. Minimum numbers of animals for parent and child nodes were assigned as 20:10 with the target to construct a proper decision tree with more nodes. A measurement of within node variance in a decision tree constructed via any data mining algorithm, risk estimate is an indicator of the predictive accuracy of the decision tree.

No special assumption on the distribution of the data to be evaluated for any study is available for CHAID (Chi-Squared Automatic Interaction Detection), Exhaustive CHAID, CART (Classification and Regression Trees) and Artificial Neural Network algorithms, all of which may be used for ordinal, nominal and scale outcome variables. As a regression task, a great effort for the study was made on predicting body weight (scale outcome variable) from biometrical (measurable=continuous) variables and sex factor (male and female), within the scope of General Linear Models (GLMs).

Model quality criteria

To choice the best algorithm, model quality criteria were calculated by means of SPSS 22 program. Formulas of the quality criteria as defined by Grzesiak and Zaborski (2012) are shown below:

Quality criteria	Formula
Coefficient of Determination (%)	$R^2(\%) = \left[1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}\right] * 100$
Adjusted Coefficient of Determination (%)	$Adj\text{-}R^2(\%) = \left[1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}\right.$
Coefficient of Variation (%)	$CV(\%) = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}}{\bar{Y}} * 100$
Standard Deviation Ratio	$SD_{ratio} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$
Relative Approximation Error	$RAE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} Y_i^2}}$
Root Mean Square Error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}}$
Pearson correlation between actual and predicted BW values	$r_{Y_i\hat{Y}_i}$

where, *Yi:* the actual body weight value of i[th] lamb, $\hat{Y}_i$: the predicted body weight value of i[th] lamb, $\bar{Y}$: mean of the actual body weight values, ε_i: the residual value of i[th] lamb, $\bar{\varepsilon}$: mean of the residual

values, *n:* total sample size and *k*: number of explanatory variables included remarkably in the model.

The value of R^2 still increases when adding new variables to the model. Therefore, the so-called "adjusted" coefficient of determination is used, which takes into account the number of degrees of freedom.

The best algorithm should have the highest value for coefficient of determination (%), adjusted coefficient of determination (%), but the lowest value for coefficient of variation (%), standard deviation ratio, relative approximation error and root mean square error.

All the statistical analyses were executed via SPSS 22 software program.

RESULTS AND DISCUSSION

Results of performance quality criteria of data mining algorithms in the present work are summarized in Table I. The Pearson correlation coefficients (r) between actual and predicted body weight values for CHAID, exhaustive CHAID, CART and ANN algorithms were found as: 0.915, 0.918, 0.909 and 0.906, respectively. Estimates of SD ratio as the rate of SD error to SD of body weight were 0.403, 0.397, 0.417 and 0.423, respectively. It could be suggested that the algorithm whose SD ratio was less than 0.40 or between 0 and 0.10 had a good fit or a very good fit (Grzesiak and Zaborski, 2012). Coefficients of variation (CV %) ((SD error/the mean of body weight)*100) for corresponding algorithms was determined to be 5.711, 5.633, 5.906 and 5.990 respectively. With the same order, coefficients of determination (R^2 %) were 83.770%, 84.210%, 82.644% and 81.999% respectively; adjusted coefficients of determination were 83.354%, 83.805%, 82.199% and 81.537%, respectively; relative approximation error (RAE) estimates were 0.0564, 0.0556, 0.0583 and 0.0594, respectively and the estimates of root mean square error were 1.509, 1.488, 1.560 and 1.589, respectively.

All the examined algorithms provided nearly similar results. However, results of the quality criteria displayed that exhaustive CHAID algorithm was better in comparison to CHAID, CART and ANN algorithms owing to the fact that the exhaustive CHAID estimated the highest r, R^2(%) and Adj-R^2 (%) values, but the lowest SD ratio, CV(%), RAE and RMSE values. Similarly, the biological superiority of CHAID algorithms in the studies concerning body weight prediction was emphasized (Mohammad *et al.,* 2012; Khan *et al.*, 2014). Also, it was determined in the study that CART algorithm could not build the proper decision tree structure, biologically.

Due to the fact that exhaustive CHAID algorithm was the most appropriate algorithm, a regression decision tree was constructed for exhaustive CHAID algorithm. Figure 1 reveals the regression tree diagram for the prediction of body weight from the measured characteristics (BL, WH, CG, PC, FL, LBE, EL and FTW) and sex factor. With an examination of the decision tree formed via exhaustive CHAID in visual form, the predominant factor most significantly affecting body weight was sex factor (F=673.553, df1=1, df2=159, Adj-P=0.000). Of measurable biometrical characteristics, WH, LBE and FL along with the sex factor explained, with an ideal accuracy of % 84.210 R^2 % and 83.805 Adj-R^2 %, the variation of the body weight on Harnai lambs. Besides, WH (Adj-P=0.034, F=9.177, df1=1, df2=69) and LBE (Adj-P=0.000, F=12.563, df1=1, df2=87) were observed to be the second-degree significant biometrical characteristics, followed by FL, (Adj-P=0.034, F=13.582, df1=2, df2=55), thirdly.

The general BW average of 26.503 (S=3.757) kg was predicted from Node 0 where 161 Harnai lambs was found at the top of regression tree diagram. Node 0, root node, was branched into two new child nodes (Nodes 1 and 2) according to sex factor, respectively. As a subgroup of only male Harnai lambs, Node 1 had the average body weight of 30.296 (S=2.187) kg. Among the assigned morphological characteristics, only WH was a major characteristic influenced on BW of the male lambs. Whereas, Node 2 with the weight average of 23.511 (S=1.041) kg was monitored as a subgroup of only female Harnai lambs and was effected by LBE. The finding means that, at the first stage, WH for male lambs and LBE for female lambs should be taken into account as also understood from Figure 1.

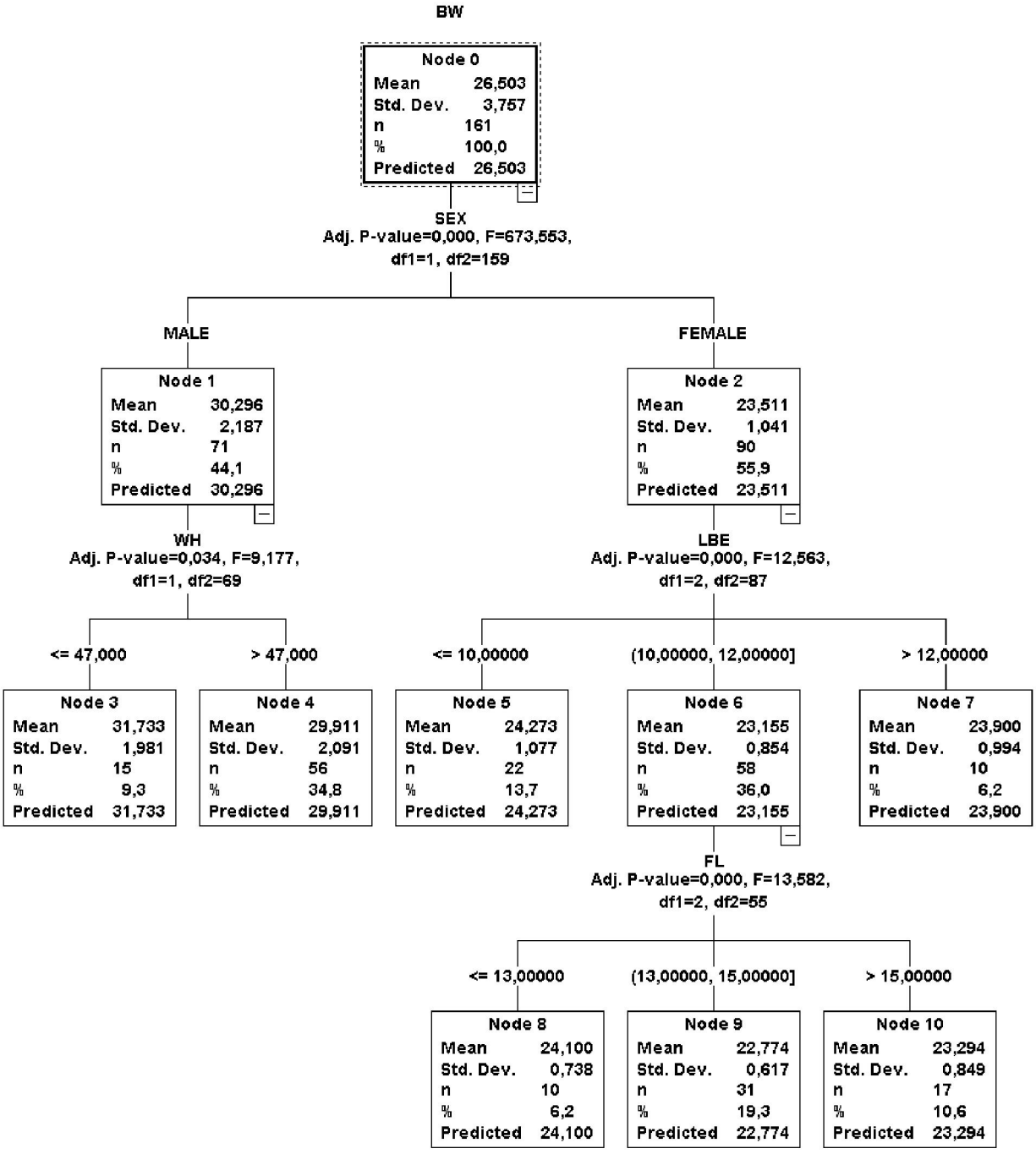

Fig. 1. Decision tree constructed for exhaustive CHAID algorithm

Node 1 was re-branched into Nodes 3 and 4, with respect to WH respectively. With the average BW of 31.733 (S=1.981) kg, Node 3 represented a sub-group of male lambs with WH $\leq$ 47 cm, however was manifestly heavier in BW than Node 4, as a sub-group of male lambs with WH > 47 cm, which produced 29.911 (S=2.091) kg (Adj-P=0.034, F=9.177, df1=1, df2=69).

Table I.- Results of performance quality criteria for data mining algorithms.

Algorithm	R	SDratio	CV%	R^2(%)	Adj-R^2 (%)	RAE	RMSE
CHAID	0.915	0.403	5.711	83.770	83.354	0.0564	1.509
EX. CHAID	0.918	0.397	5.633	84.210	83.805	0.0556	1.488
CART	0.909	0.417	5.906	82.644	82.199	0.0583	1.560
ANN	0.906	0.423	5.990	81.999	81.537	0.0594	1.589

In terms of LBE, Node 2 (being a sub-group of all the female lambs) was split into 3 new child Nodes 5, 6 and 7, respectively. Nodes 5 and 7 are terminal nodes in which splitting were stopped at next stages. Described as a sub-group of female lambs with LBE $\leq$ 10 cm, Node 5, gave the average BW of 24.273 (S=1.077) kg. Average BW for the sub-group of female lambs with LBE > 12 cm, which was also labeled as Node 7, was 23.900 (S=0.994) kg. The sub-group of female lambs with only 10 < LBE $\leq$ 12 was depictured with Node 6 (23.155 (S=0.854)). FL meaningfully effected BW of the female lambs with 10 < LBE $\leq$ 12 (Adj-P=0.000, F=13.582, df1=2 and df2=55).

With respect to FL characteristics, Node 6 was re-branched into Nodes 8, 9 and 10, respectively. Expressed as the sub-group of the female lambs with FL $\leq$ 13 and 10 < LBE $\leq$ 12 cm, Node 8 had the BW average of 24.100 (S=0.738) kg. The average BW of Node 9 assigned as the sub-group of the female lambs with 13 < FL $\leq$ 15 and 10 < LBE $\leq$ 12 cm was predicted as 22.774 (S=0.617) kg. The sub-group of the female lambs with FL > 15 and 10 < LBE $\leq$ 12 cm, Node 10, provided 23.294 (S=0.849) kg.

A detailed review on the importance of the body weight prediction from several linear body measurements and the statistical procedures (multiple linear regression, using factor and principle component scores in multiple linear regression and regression tree methods via CHAID or exhaustive CHAID algorithms) employed scarcely in the prediction was reported by Mahmud *et al*. (2014). In literature, rather than multiple linear regression applications, there was a more limited utilization of CHAID, exhaustive CHAID, CART and ANN algorithms in order to predict body weight from biometrical characteristics, which are strongly correlated with meat yield and reproduction characteristics, as well as to remove multicollinearity condition. For example, Yakubu (2012) applied only CART algorithm to investigate the relationship between body weight and nine morphometric traits (withers height, rump height, body length, face length, rump length, chest circumference, head width, shoulder width and rump width) of Uda sheep and statistically found chest circumference and face length in predicting the BW of the Uda sheep with approximately 62% R^2 for the BW. In other investigation, Mohammad *et al.* (2012) reported that 72 (%) of the variability in BW at yearling age was accounted for by WH, body length (BL) and chest girth (CG) via regression tree method, respectively and that the sheep with chest girth greater than 89 cm within all the sheep could produced the heaviest average BW with 36.486 kg. These two previous results were importantly lower compared to those of CART and other data mining algorithms in the current study (Table I). With the exhaustive CHAID algorithm, Khan *et al.* (2014) emphasized that 84.4 % R^2 of the variability of BW in Harnai sheep was explained by FL, WH, CG and BL, respectively. The results of Khan *et al.* (2014) were in almost agreement with the R^2 (%) estimates of exhaustive CHAID and other constructed algorithms in the study, with very limited part of same data set.

A practice way to establish the relationship between BW and morphological linear body measurements is to calculate Pearson correlation coefficients, which may cause insufficient comments, biologically (Mohammad *et al.*, 2012). Rather than using multiple linear regression, ridge regression and polynomial regression analyses; using factor and principle component scores for multiple regression analyses, robust regression methods, have been preferred for the description of the relationship, in recent years. However, more

inclusive report on the utilization of the data mining algorithms has not been yet found for the description, in contradistinction to the study where the seven quality criteria were estimated for four data mining algorithms, in an effort to prove breed characteristics of Harnai lambs.

Compared with the results of the earlier studies, the study was likely to be very different, which may be ascribed to wide variation in animal ages, managerial conditions, body measurements and environmental factors and more especially, breeds and the statistical procedures marked in all the studies.

To better assess performance of CHAID, exhaustive CHAID, CART and ANN algorithms on the subject of the more accurate description of Harnai breed standards and removing multicollinearity problem, it is recommended for further investigators to study much larger populations, a great number of efficient factors and to appraise a large number of sheep breeds in very large populations in generalization of the results in the current investigation.

REFERENCES

CANKAYA, S., 2009. A comparative study of some estimation methods for parameters and effects of outliers in simple regression model for research on small ruminants. *Trop. Anim. Hlth. Prod.*, **41**:35-41.

CANKAYA, S., ALTOP, A., KUL, E. AND ERENER, G., 2009. Body weight estimation in karayaka lambs by using factor analysis scores. *Anadolu J. agric. Sci.*, **24**: 98-102.

EYDURAN, E., KARAKUS, K., KARAKUS, S. AND CENGIZ, F., 2009. Usage of factor scores for determining relationships among body weight and body measurements. *Bulgarian J. agric. Sci.*, **15**: 374–378.

GORGULU, O., 2012. Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks. *South African J. Anim. Sci*, **42:**280-287.

GRZESIAK W., BLASZCZYK, P. AND LACROIX R., 2006. Methods of predicting milk yield in dairy cows-Predictive capabilities of Wood's lactation curve and artificial neural networks (ANNs). *Comput. Electron. Agric.* **54**:69-83.

GRZESIAK, W., RZEWUCKA-WOJCIK, E., ZABORSKI, D., SZATKOWSKA, I., KOTARSKA, K. AND DYBUS, A., 2014. Classification of daily body weight gains in beef cattle via neural networks and decision trees. *Appl. Engin. Agric.*, **30**: 307-313.

GRZESIAK, W. AND ZABORSKI, D., 2012. *Examples of the use of data mining methods in animal breeding*. (Book) ISBN 978-953-51-0720-0

JAHAN, M., TARIQ, M. M., KAKAR, M. A., EYDURAN, E. AND WAHEED, A., 2013. Predicting body weight from body and testicular characteristics of Balochi male sheep in Pakistan using different statistical Analyses. *J. Anim. Pl. Sci.*, **23**:14-19.

KHAN, M. A., TARIQ, M.M., EYDURAN, E., TATLIYER, A., RAFEEQ, M., ABBAS, F., RASHID, N., AWAN, M. A. AND JAVED, K., 2014. Estimating body weight from several body measurements in Harnai sheep without multicollinearity problem. *J. Anim. Pl. Sci*., **24**: 120-126.

MAHMUD, M.A., SHABA, P. AND ZUBAIRU, U.Y., 2014. Live body weight in small ruminants-a review. *Global J. Anim. Sci. Res.*, **2**:102-108.

MENDES, M. AND AKKARTAL, E., 2009. Regression tree analysis for predicting slaughter weight in broilers. *Italian J. Anim. Sci*., **8**: 615-624

MOHAMMAD, M. T., RAFEEQ, M., BAJWA, M. A., AWAN, M. A., ABBAS, F., WAHEED, A., BUKHARI, F. A. AND AKHTAR, P., 2012. Prediction of body weight from body measurements using regression tree (RT) method for indigenous sheep breeds in Balochistan, Pakistan. *J. Anim. Pl. Sci.*, **22**: 20-24.

MOHAMMAD M. T, FARHAT, I., ECEVIT, E., MASROOR A. B., ZIL E H., AND ABDUL W., 2013. Comparison of non-linear functions to describe the growth in mengali sheep breed of Balochistan. *Pakistan J. Zool.*, **45:** 661-665.

SHAHINFAR, S., MEHRABANI-YEGANEH, H., LUCAS, C., KALHOR, A., KAZEMIAN, M. AND WEIGEL, K. A., 2012. *Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems*. Computational and Mathematical Methods in Medicine, Article ID 127130, 9 pages. doi:10.1155/2012/127130.

TAKMA, C., ATIL, H. AND AKSAKAL, V., 2012. Comparison of multiple linear regression and artificial neural network models goodness of fit to lactation milk yields. *Kafkas Univ. Vet. Fac. J.*, **18**:941-944.

YAKUBU, A., 2009. Fixing collinearity instability in the estimation of body weight from morpho-biometrical traits of West African dwarf goats. *Trakia J. Sci.*, 7: 61-66.

YAKUBU, A., 2012. Application of regression tree methodology in predicting the body weight of Uda sheep. *Anim. Sci. Biotech.*, **45**: 484-490.

(Received 2 April 2015, revised 29 April 2015)