TÜBİTAK

Turkish Journal of Veterinary and Animal Sciences

http://journals.tubitak.gov.tr/veterinary/

Research Article

Turk J Vet Anim Sci
(2019) 43: 500-506

doi:10.3906/vet-1812-23

Predicting the body weight of Balochi sheep using a machine learning approach

Zil E HUMA[1], **Farhat IQBAL**[2,*]

[1]Department of Zoology, Sardar Bahadur Khan Women's University, Quetta, Pakistan
[2]Department of Statistics, University of Balochistan, Quetta, Pakistan

Received: 06.12.2018 • **Accepted/Published Online:** 14.06.2019 • **Final Version:** 07.08.2019

Abstract: Various machine learning algorithms have been used to model and predict the body weight of rams of the Balochi sheep breed of Pakistan. The traditional generalized linear model along with regression trees, support vector machine, and random forests methods have been used to develop models for the prediction of the body weight of animals. The independent variables (inputs) include the body (body length, heart girth, withers height) and testicular (scrotal diameter, scrotal circumference, scrotal length, and testicular length) measurements of 131 male sheep 2–36 months of age. The performance of the models is assessed based on evaluation criteria of mean absolute error, mean absolute percentage error, correlation between observed and fitted values, coefficient of determination, and root mean squared error. A 10-fold cross-validation is done on a training dataset to check the stability of the models. A separate training dataset is used to assess the predictive performance of the developed models. The random forests model was found to provide the best results for both training and testing datasets. It was concluded that machine learning methods may provide better results than the traditional models and may help practitioners and researchers choose the best predictors for body weight of farm animals.

Key words: Body weight, ram sheep, body measurements, machine learning

1. Introduction

In the socioeconomic life of the people of Balochistan, Pakistan, sheep occupy a strategic position. The Balochi sheep is an indigenous sheep breed of Balochistan primarily reared for mutton production; it makes a significant contribution to household income in rural areas. This breed, also found in the eastern parts of Iran, is well adapted to a wide range of harsh climate conditions. Balochi sheep generally have a white medium-sized body with a fat tail and black, brown, or spotted muzzle and legs.

Body weight, an important measure of animal performance, not only provides an informative measure for feeding, health care, and breeding (selection) of animals, but has also been found to be very effective in evaluating reproductive efficacy in sheep. Reproductive performance of sheep is one of the key factors in profitability [1]. For fertility in sheep, testicular length and scrotal circumference and length, among other testicular characteristics, are considered important variables [2]. The growth and development of testicular characteristics have been reported to be closely related to the body size of animals [3].

Predicting the body weight of farm animals from various body traits observed at different growth periods for sheep [4,5], goat [6,7], and cattle [8,9] has been studied in detail in the literature. Most past studies have employed multiple linear regression analysis for modelling the body weight (dependent variable) of animals based on various body and testicular traits (independent variables). However, it has been reported that the strong correlation among independent variables, also known as multicollinearity, generally exists; as a consequence, large standard errors of the parameters have been obtained, resulting in inaccurate estimates [10]. As a remedy, few studies have used alternative methods such as ridge regression and factor analysis scores in multiple regression [5,11]. These statistical tools have also been employed for predicting the body weight of Balochi sheep using various biometrical traits [10]. However, these traditional methods are inadequate for explaining complex relationships.

Recently, a few researchers have successfully applied various data mining and machine algorithms for the prediction of live body weight of animals using morphological traits. These methods aim to map body weight from a collection to morphological measures of animals. Applied chi-square automatic interaction detector (CHAID), exhaustive CHAID (ECHAID), classification and regression tree (CART), and artificial

* Correspondence: farhatiqb@gmail.com

neural networks (ANN) data mining algorithms were used for body weight prediction for the Harnai sheep breed of Balochistan [12]. The CHAID, ECHAID, and CART algorithms were used for predicting the body weights of three dog varieties of Turkey [13]. Multivariate adaptive regression splines (MARS) algorithms along with CART were employed to estimate important variables for predicting the body weights of Turkish Tazi dogs [14]. The CART, CHAID, radial basis function (RBF), and multilayer perceptron (MLP) methods were used to find the best predictive model for body weight by means of various body measurements in the indigenous Beetal goat of Pakistan [15], whereas Aytekin et al. [16] applied the MARS algorithm to the prediction of fattening final weight of bulls from some body measurements. These studies have reported the potential of data mining algorithms in accurately predicting the nonlinear relation between body weight and morphological and biometrical traits of animals. The application of various machine learning methods for developing a body weight prediction model for animals appears to be a promising alternative, and has been further investigated in the present study.

This study aimed to determine the best soft computing methods to predict the body weight of sheep using various morphological and testicular characteristics. Another aim was to provide a robust method for modelling and predicting, in a machine learning framework, by randomly partitioning the data into training and testing parts. A cross-validation approach is applied to the training dataset to correctly model the relationship between the dependent and independent variables and to avoid overfitting of models. The testing dataset is then used to assess the predictive performance of competing models. No studies in the literature, to our knowledge, have reported on the prediction of body weight of small ruminants by exploiting the combination of machine learning methods. Therefore, this is the first study in which the traditional generalized line model and different machine learning models, namely, regression trees, support vector machine, and random forests have been employed for modelling and predicting body weight from several biometrical (body length, heart girth, and withers height) and testicular (scrotal circumference, scrotal diameter, scrotal length, and testicular length) traits taken as input variables for small ruminants.

2. Materials and methods

2.1. Dataset and variables

This study utilizes data from 131 Balochi male sheep kept in private sheep flocks and government livestock farms in the Quetta, Mastung, and Usta Mohammad districts of Balochistan, Pakistan. The dependent variable body weight (BW) and independent variables such as body length (BL), heart girth (HG), withers height (WH), scrotal length (SL), scrotal circumference (SC), scrotal diameter (SD), and testicular length (TL) were measured for sheep aged 2–36 months using tailor tape and weigh balance. Some basic descriptive statistics of variables used in the study are given in Table 1.

2.2. Machine learning models

In the present study, the following four different machine learning methods have been used.

2.2.1. Linear models

The first model, though not a pure machine learning method, is the generalized linear model (GLM), which includes linear regression as a simple and basic form [17]. The multiple linear regression model is a commonly used method for modelling the relationship between a dependent and set of independent variables. This method requires some strict assumptions, such as normality of data and no multicollinearity in independent variables, among others.

2.2.2. Regression trees

The classification and regression trees method used by Breiman et al. [18] is a recursive partitioning method that can predict both the categorical dependent variable (classification) and continuous dependent variable (regression) by building trees. The regression trees (RT) method is a variant of decision trees designed to approximate real-valued functions. This RT procedure splits the data at several points for each independent variable. At each split point, the sum of squared errors is calculated and compared across the variables. The variable yielding the lowest sum of squared errors is chosen as the root node/split point. This process is recursively continued until a stopping criterion is reached.

2.2.3. Random forests

Ensembles of regression trees known as random decision forests or simply random forests (RF) are a flexible and easy

Table 1. Mean, standard deviation (S.D.), and coefficient of variation (CV) of each variable.

Variables	Mean	S.D.	CV (%)
Body weight (kg)	39.74	19.85	49.96
Body length (cm)	24.47	12.70	51.91
Heart girth (cm)	73.54	19.14	26.02
Withers height (cm)	62.38	13.28	21.29
Scrotal length (cm)	13.17	3.56	27.01
Scrotal circumference (cm)	20.45	7.45	36.44
Scrotal diameter (cm)	10.22	3.73	36.44
Testicular length (cm)	10.89	3.49	32.02

to use machine learning algorithm. One of the problems encountered while using RT was the overfitting of data. The RF method used by Breiman [19] avoids this problem by forming multiple shallow trees instead of a single deep tree. This method identifies complex patterns in the data by randomly selecting records and variables. Accurate predictions can be achieved as the output is accumulated and the errors are cancelled out.

2.2.4. Support vector machine regression

Support vector machine is another important machine learning algorithm that can be used for both classification and regression problems in high dimensional spaces. As an alternative to a regression method, the support vector machine (SVM) regression is a popular machine learning tool that can be used to estimate a nonlinear function. The SVM regression of Vapnik et al. [20] relies on kernel functions and is thus considered a nonparametric technique. It can generally be thought of as an alternative training technique for popular neural networks models such as multilayer perceptron and radial basis function classifiers. In SVM, the problem is transformed into a quadratic optimization problem which can obtain the globally optimal solution. SVM can take care of practical problems such as nonlinearity, small sample size, local minimum, and high dimensionality of the data [21].

2.3. Model evaluation

Different evaluation criteria have been employed to assess the performance of the models developed in this study for modelling and predicting the body weight of sheep.

2.3.1. Traditional valuation measures

We consider a variety of commonly used evaluation measures in this study. These include the Pearson's coefficient of correlation (r) between the observed and predicted body weights, the coefficient of determination (R^2), the mean absolute error (MAE), the root mean squared error (RMSE), and the mean absolute percentage error (MAPE).

2.3.2. *k*-fold cross-validation

Cross-validation is a commonly used statistical method for assessing the effectiveness of a machine learning model. It is based on resampling procedure and is ideally suited for limited datasets. Cross-validation divides the data into *k* numbers of folds also known as subsamples. These subsamples are used to train and validate the model. This method uses all the data for training and validation and also for estimating the prediction error. This procedure not only helps mitigate overfitting but is also useful in determining the hyperparameters of the model. Generally, a 10-fold cross-validation is used for this purpose. Cross-validation is a popular choice among practitioners due to its simplicity and easy to implement procedure.

A common approach used by researchers is to fit competing models to the whole dataset and then evaluate the performance of the models using various evaluation measures. This approach mostly leads to optimistic results based on overfitting of the model, because one cannot just fit a model to a training dataset and hope it would accurately work for the real unseen dataset. Hence, we adopted a different approach in this study. The dataset was initially partitioned randomly into two parts, the training (75%) and testing (25%) datasets. The training dataset was used for tuning the parameters of the four machine learning methods using 10-fold cross-validation. Once the best model is developed, the testing dataset was used for the prediction of outcome variables and validation of the fitted models. Figure 1 shows the layout of the methodology used in this study. Use of an independent testing dataset for validation purposes may help to better evaluate the predictive ability of fitted models. The R program [22] was used for all statistical analysis.

3. Results

Table 2 shows the results of various evaluation measures used to evaluate a model's performance on both training and testing datasets. It can be noticed from the results of the table that although all models can be used to model the body weight of sheep, the RF methods gave the best result on all evaluation measures for both training and testing datasets. For the training dataset, the correlation coefficient between the observed and fitted values of all models considered in this research were in the range of 0.947–0.994, with RF providing the highest value. The same observation was true for the coefficient of determination, whose values ranged from 0.896 (RT) to 0.988 (RF). The mean absolute error varied from 1.242 (RF) to 4.583 (RT), whereas the mean absolute prediction error was in the range 2.810 to 14.703. The root means squared error of the RF model (2.129) was found to be the minimum among all models. The MAPE value of 2.810 for RF model was found to be the lowest compared to the MAPE of other models (6.429 for SVM to 14.073 for RT).

As mentioned earlier, a model may overfit the training data yet fail to predict the test data accurately. Hence, we evaluated the predictive performance of all models on a separate test dataset. The results of evaluation measures on the testing dataset are also presented in Table 2. The RF method was a clear winner in predicting the body weight. The values of r (0.957) and R^2 (0.916) were both found to be the highest while the values of MAE (3.275), RMSE (5.390), and MAPE (7.946) were the lowest for this machine learning method.

Table 3 presents the observed body weight (BW in kg) of Balochi sheep and the predicted values of BW obtained from all four models for a sample testing dataset. The

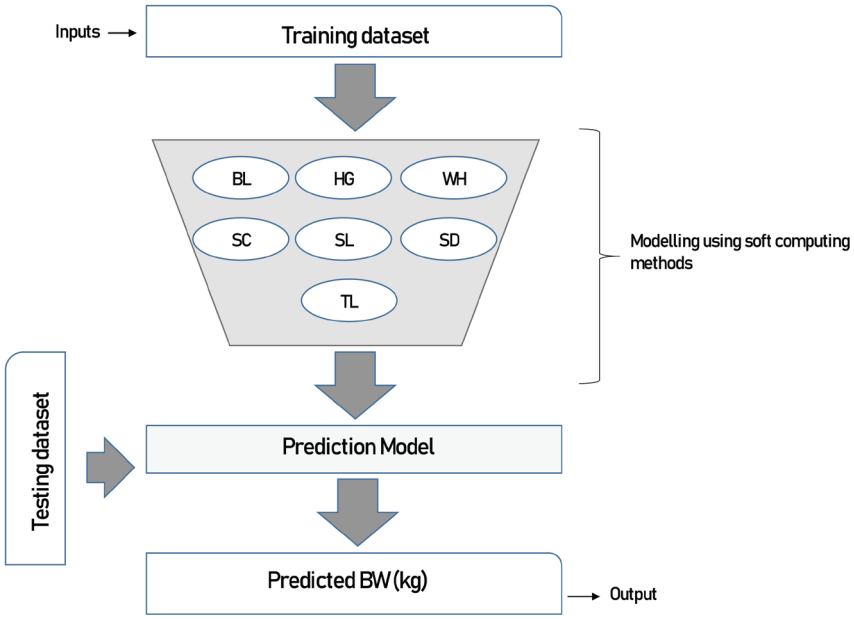

Figure 1. Prediction method for body weight (BW) of sheep using machine learning approach.

Table 2. Evaluating models based on different performance measures.

Model	Training dataset (95 samples)					Testing dataset (36 samples)				
	r	R^2	MAE	RMSE	MAPE	r	R^2	MAE	RMSE	MAPE
Linear model	0.964	0.929	3.519	5.101	9.052	0.928	0.861	5.064	6.587	12.149
Regression trees	0.947	0.896	4.583	6.197	14.703	0.924	0.854	5.871	7.023	17.419
Random forests	0.994	0.988	1.242	2.129	2.810	0.957	0.916	3.275	5.390	7.946
Support vector machine	0.988	0.976	2.169	3.097	6.429	0.947	0.897	3.934	5.938	11.086

Table 3. A sample dataset of observed vs. predicted values of body weight.

Observed BW (kg)	Linear model		Regression trees		Random forests		Support vector machine	
	Predicted BW (kg)	Error (kg)	Predicted BW (kg)	Error (kg)	Predicted BW (kg)	Error (kg)	Predicted BW (kg)	Error (kg)
18.00	16.524	–1.476	22.474	4.474	18.258	0.258	18.274	0.274
25.50	24.696	–0.836	22.474	–3.026	25.335	–0.165	27.654	2.154
30.90	35.078	4.178	22.474	–8.427	30.846	–0.054	31.710	0.810
55.00	49.717	–5.284	54.429	–0.571	53.242	–1.758	48.424	–6.576
60.00	60.468	0.468	67.556	7.556	60.644	0.643	62.358	2.358

corresponding prediction error values (in kg) are also reported. The errors of all models varied from very small to quite large values except for the random forests model. The random forests method produced the least values of residuals (prediction errors), confirming its better predictive ability than the competing methods.

Figure 2 shows the importance of predictors identified by the random forests method for describing the body

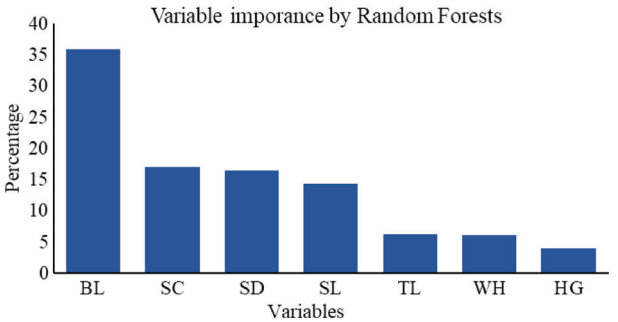

Figure 2. Variable importance by the random forests method.

weight of Balochi sheep. The most important variable was found to be the body length (BL) of sheep, accounting for around 35% of the variation in the weight of the animals. Scrotal circumference, scrotal diameter, and scrotal length were also found to be important predictors, each with approximately 15% weights. Other variables such as testicular length (TL), withers height (WH), and heart girth (HG) contributed little in predicting the body weight.

The results of 10-fold cross-validation for the best performing (random forests) method for various evaluation measures are shown in Figure 3. For all 10 iterations, the values of four evaluation measures remain almost the same, indicating the stability of the random forests method for fitting the predictive body weight model of sheep. Thus, we can say that the RF method performed better than all other models used in this study for modelling the body weight of Balochi sheep.

To further check the significant difference between the observed body weight and those predicted by the random forests method, a two-sample t-test was performed for the testing dataset and the results are presented in Table 4. The high P-value (0.762) of the test provided evidence that the difference between the observed and predicted body weights of RF method are not statistically different at 5% level of significance.

4. Discussion

A model showing good performance on training data need not be the best model for prediction. We emphasize again that our approach of modelling based on 10-fold

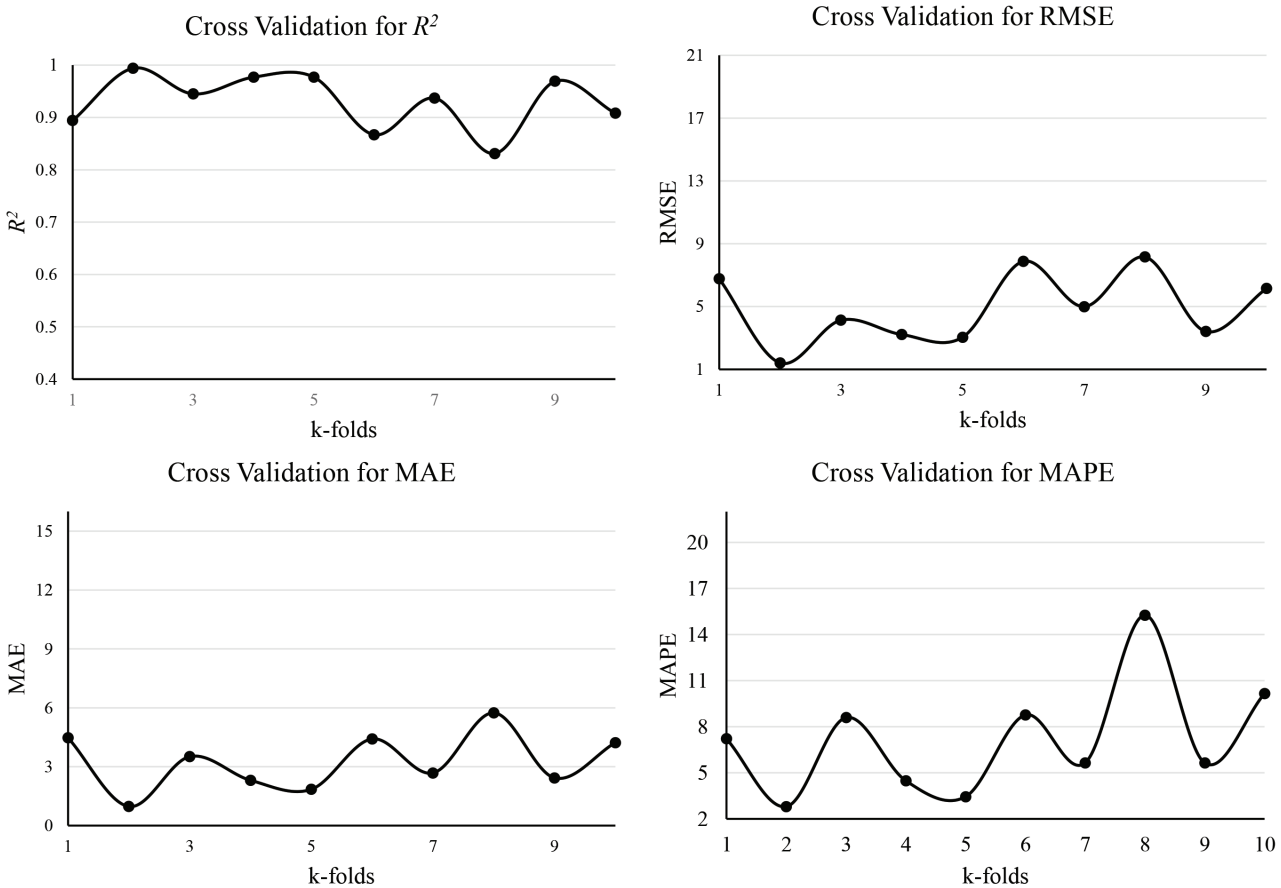

Figure 3. 10-fold cross-validation for R^2, RMSE, MAE, and MAPE by the random forests method.

Table 4. Results of t-test for difference between the observed and predicted weights for random forests.

Test variable	Result of t-test	
	t-stat	P-value
Body weight (kg)	0.304	0.762
Number of observations	36	

cross-validation for a training dataset to obtain the best model, and then exposing this model to a separate dataset for evaluation, is different from those of other studies in the literature. In this respect, an exact comparison of the results obtained from the present research on machine learning methods for body weight prediction with earlier results from classical regression and data mining methods published in the literature could not be made here. However, we may compare the predictive accuracy of our approach based on the R^2 (coefficient of determination) values. The coefficient of determination R^2 values of 0.988 and 0.916 for training and testing datasets, respectively, of the random forests method used in this study were higher than that of Jahan et al. (10), who reported R^2=0.911 for the same dataset when factor scores with multiple regression were used to model the body weight of Balochi sheep. Tariq et al. [23] predicted the body weight of an indigenous sheep breed of Balochistan using the RT method and reported a coefficient of determination value of 0.72. Their reported value of R^2 is also smaller than the R^2 values of all methods in the present research. The values of RMSE and MAPE obtained from RF and SVM in this study are smaller than those obtained from CART, CHAID, RBF, and MLP methods reported by Eyduran et al. [15]. However, the R^2 value of 0.9717 obtained from the MARS algorithm for prediction of the fattening final weight of bulls by Aytekin et al. [16] is close to the R^2 values of the RF method of this study. We observed that the RF method not only achieved much higher predictive performance than other competing methods used in this study, but also then other machine learning methods used in similar studies.

The LM method, although it performs better than the other two machine learning models for prediction, cannot be relied upon without properly checking all of its assumptions. Based on the results of both training and testing datasets, we may conclude that the random forests method clearly outperforms all other methods on different evaluation measures and can be used to develop body weight prediction models with high accuracy.

The SVM for regression can be considered the second-best model based on these evaluation measures. Surprisingly, the RT method could not provide a more accurate fit than the LM. However, the LM model may not be preferred over RT, as the former requires very strong assumptions about the data such as no multicollinearity among independent variables, which may lead to serious consequences if not addressed properly.

The RF method can be an attractive option for modelling complex relationships between variables as compared to other models for researchers based on its features. The RF method takes less time to model than other machine learning methods, especially for large datasets with a large number of parameters. It is an ensemble method more appealing for real time predictions which can handle missing values. Therefore, it can be used by researchers, academics, practitioners, and biostatisticians in modelling and predicting when the relationship between variables is complex or unknown. Our results showed that the RF provided an accurate fit to the body weight data of Balochi sheep. We also observed that the performance of all models decreased when exposed to an independent testing dataset. Hence, trusting a model based solely on its accuracy on a training dataset is not advised. A researcher needs to test the model's predictive accuracy before drawing any conclusions.

This study employed the generalized linear model, regression trees, support vector machine, and random forests methods to predict the body weight of the Balochi breed of sheep of Balochistan using various body measures. Using various evaluation measures, we found strong evidence of better performance for machine learning methods. Random forests followed by support vector machine regression and regression trees were found to provide more accurate predictions of body weight, outperforming the traditional linear model. Based on the results of the present study, we conclude that the random forests method can be used to model and predict body weight via various biometric and testis characteristics in small ruminants. The findings of this study may help researchers and practitioners to adopt the latest machine learning methods for accurate prediction of body weight using various biometrical and testicular traits in farm animals. Moreover, the k-fold cross-validation may be used each time a new model is fitted to a dataset to avoid overfitting of the model.

Acknowledgments
We are thankful to Dr. Masood Tariq for providing the research data.

References

1. Bilgin OC, Emsen E, Davis MH. Comparison of non-linear models for describing the growth of scrotal circumference in Awassi male lambs. Small Rumin Res 2004; 52: 155-160. doi: 10.1016/S0921-4488(03)00251-7
2. Koyuncu M, Uzun SK, Ozis S, Duru S. Development of testicular dimensions and size, and their relationship to age and body weight in growing. Kivircik (Western Thrace) ram lambs. Czech J Anim Sci 2005; 50: 243-248. doi: 10.17221/4164-CJAS
3. Land RB, Gauld FK, Lee GS, Webb R. Further possibilities for manipulating the reproductive process. In: Barker SF, Hammond K, McClintock AE, editors. Further Development in the Genetic Improvement of Animals. Sydney, Australia: Academic Press; 1982. pp. 59-87.
4. Cam, MA, Olfaz M, Soydan E. Possibilities of using morphometrics characteristics as a tool for body weight production in Turkish hair goats (Kilkeci). Asian J Anim Vet Adv 2010; 5: 52-59. doi: 10.3923/ajava.2010.52.59
5. Tariq MM, Eyduran E, Bajwa MA, Waheed A, Iqbal F, Javed Y. Prediction of body weight from testicular and morphological characteristics in indigenous Mengali sheep of Pakistan: using factor analysis scores in multiple linear regression analysis. Int J Agric Biol 2012; 14: 590-594.
6. Rahman F. Prediction of carcass weight from the body characteristics of black goats. Int J Agric Biol 2007; 9: 431-434.
7. Cam MA, Olfaz M, Soydan E. Body measurement reflect body weights and carcass yields in Karakaya sheep. Asian J Anim Vet Adv 2010; 5: 120-127. doi: 10.3923/ajava.2010.120.127
8. Siddiqui MU, Lateef M, Bashir MK, Bilal MQ, Muhammad G, Mustafa MI, Rehman S. Estimation of live weight using different body measurements in Sahiwal Cattle. Pak J Life Soc Sci 2015; 13: 12-15.
9. Karadas K, Tariq M. Tariq MM, Eyduran E. Measuring predictive performance of data mining and artificial neural network algorithms for predicting lactation milk yield in indigenous Akkaraman sheep. Pak J Zool 2017; 49: 1-7. doi: 10.17582/journal.pjz/2017.49.1.1.7
10. Jahan M, Tariq MM, Kakar MA, Eudyran E, Waheed A. Predicting body weight from body and testicular characteristics of Balochi male sheep in Pakistan using different statistical analyses. J Anim Plant Sci 2013; 23: 14-19.
11. Khan MA, Tariq MM, Eyduran E, Tatliyer A, Rafeeq M, Abbas F, Rashid N, Awan MA, Javed, K. Estimating body weight from several body measurements in Harnai sheep without multicollinearity problem. J Anim Plant Sci 2014; 24: 120-126.
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. Pak J Zool 2015; 47: 1579-1585. doi: 10.17582/journal.pjz/2015.47.6.1579.1585
13. 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 2017; 27: 57-64.
14. Celik S, Yilmaz O. Prediction of body weight of Turkish Tazi dogs using data mining techniques: Classification and Regression Tree (CART) and Multivariate Adaptive Regression Splines (MARS). Pak J Zool 2018; 50: 575-583. doi: 10.17582/journal.pjz/2018.50.2.55.583
15. Eyduran E, Zaborski D, Waheed A, Celik S, 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. Pak J Zool 2017; 49: 273-282. doi: 10.17582/journal.pjz/2017.49.1.273.282
16. Aytekin I, Eyduran E, Karadas K, Aksahan R, Keskin I. Prediction of fattening final live weight from some body measurements and fattening period in young bulls of crossbred and exotic breeds using MARS data mining algorithm. Pak J Zool 2018; 50: 189-195. doi: 10.17582/journal.pjz/2018.50.1.189.195
17. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A 1972; 135: 370-384. doi: 10.2307/2344614
18. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Boca Raton, FL, USA: Chapman & Hall/CRC; 1984.
19. Breiman L. Random forests. Mach Learn 2001; 45: 5-32. doi: 10.1023/A:101093340
20. Vapnik V, Golowich S, Samola A. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer M, Jordan M, and PetscheT, editors. Neural Information Processing Systems, Vol. 9. Cambridge, MA, USA: MIT Press; 1977.
21. Vapnik V. Statistical Learning Theory. DBLP; 2010.
22. Team RC. R: A Language and Environment for Statistical Computing. (Version 3.4.2). [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing; 2017.
23. Mohammad MT, Rafeeq M, Bajwa MA, Awan A, Abbas F, Waheed A, Bukhari A, Akhtar P. Prediction of body weights from body measurements using Regression Tree (RT) method for indigenous sheep breed in Balochistan, Pakistan. J Anim Plant Sci 2012; 22: 20-24.