RESEARCH ARTICLE Indian Journal of Animal Research, Volume 58 Issue 5: 884-889 (May 2024)

Artificial Intelligence Techniques for the Prediction of Body Weights in Sheep

Ambreen Hamadani[1], Nazir Ahmad Ganai[1], Safeer Alam[1], Syed Mudasir[1], Tariq A Raja[1], Ishraq Hussain[1], Haider Ali Ahmad[1]

10.18805/IJAR.B-4831

ABSTRACT

Background: Artificial intelligence (AI) is transforming all spheres of life and it has the potential to revolutionize animal husbandry as well. In this regard, an attempt was made to compare two AI techniques for predicting 12-month body weights of animals; *viz*. Principal Component regression (PCR) and Ordinary Least Squares (OLS) for datasets of Corriedale sheep spanning 11 years.

Methods: PCR models were trained by varying proportions of training and testing datasets. The dataset was subject to PCR before analysis and tested (PCA dataset). A separate dataset was also created by feature selection of the PCA (PCA+FS dataset) variables.

Result: The highest correlation coefficients between test and predicted variables for two datasets (PCA dataset and PCA+FS dataset) created among the multiple models trained using PCR were 0.982 and 0.9741. In terms of error, R2 or correlation coefficient, the PCA dataset performed better than the PCA+FS dataset. The second principal component had the highest explained variance in OLS (86.16%) and the highest coefficient of determination (R2) using OLS was obtained for the PCA dataset *viz*. 0.980. It is concluded that both the algorithms tested in this study were satisfactorily trained in their prediction of the body weights with OLS performing better than PCA in terms of R2 value.

Key words: Dimensionality reduction, Feature selection, Machine learning, PCA, Principal component regression.

INTRODUCTION

Machine learning is changing the world in every known dimension. It is seeping into all spheres including the biological sphere and transforming it for the better. This is reason enough for applying the various techniques of machine learning in animal sciences. Such techniques have recently found application in all areas of Animal Husbandry including production, management, breeding, welfare, farm sustainability, health surveillance and disease prevention, as well as in human welfare through animals and environmental protection.

As the global technological revolution is becoming a reality, data resources are becoming abundant (Zhang *et al*., 2021). These data resources can therefore be mined to yield information that was hitherto unknown. The prediction and analysis of animal production using big data is gaining more and more importance. This focus area of ML, which in itself is a sub area of Artificial Intelligence (Cihan *et al*., 2017), will enable breeders and farm managers to build models by learning from this data to predict the performance of superior animals and use that knowledge to eliminate unproductive animals from the farm early.

The prediction models once built do not require training again (Moncaster, 2020) and therefore can be reused for years even by people with no knowledge of data analytics. This is a major advantage of building robust Machine Learning Models in Animal Sciences over other techniques.

Sheep is one of the most useful domestic animals in the Jammu and Kashmir. It contributes greatly to the economy of farmers especially the poor and marginalized ones. This is largely owing to the fact that the population of J & K is a mutton consumer and the agro-climatic conditions in J & K are favourable. Early prediction of body weights in sheep would therefore improve animal management, improve the utilization of resources including medication, feed, *etc*. It would also help in effective replacement of males and females. This weight information can also be useful for the early determination of the worth of the animal and therefore lead to early and efficient economic planning of the farm (Shirzeyli *et al*., 2013).

While body measurements have been used for the prediction of body weights by various authors (Afolayan *et al*., 2006; Cam *et al*., 2010a; Gül *et al*., 2005; Khan *et al*., 2006; Moaeen-ud-Din *et al*., 2006; Rahman, 2007; Pesmen and Yardimci, 2008; Shirzeyli *et al*., 2013), there may also be other factors which have often not been taken into

[1]Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar-190 025, Jammu and Kashmir, India.

Corresponding Author: Ambreen Hamadani, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar-190 025, Jammu and Kashmir, India.
Email: escritor005@gmail.com

How to cite this article: Hamadani, A., Ganai, N.A., Alam, S., Mudasir, S., Raja, T.A.,Hussain, I. and Ahmad, H.A. (2024). Artificial Intelligence Techniques for the Prediction of Body Weights in Sheep. Indian Journal of Animal Research. 58(5): 884-889. doi: 10.18805/IJAR.B-4831.

Submitted: 17-11-2021 **Accepted:** 07-09-2022 **Online:** 14-09-2022

consideration for the prediction of body weight at a later age (Cho *et al*., 2020). ML models can take all these factors into consideration for more accurate predictions.

Principal components regression is an important ML technique for analysing multiple regression data as it takes care of the multicollinearity within the data (Jolliffe, 1982) during analysis unlike other popular techniques like ordinary least squares. By adding a bias degree to the estimates, principal components regression helps in reducing the standard errors and therefore the estimates so procured are more reliable. We, therefore, attempted to predict body weight of animals from data available earlier at the farm using a lesser known but very useful technique in Machine Learning. In this study, three different types of data sets were compared based on the type of data reduction technique used.

MATERIALS AND METHODS

Data collection and digitization

Data for Corriedale breed for the last 11 years (2011-2021) was collected from Mountain Research Station for Sheep and Goat, J & K, India. Economically important traits like birth weight, weaning weight, 6-month weight, 9-month weight, 12-month weight as well as morphometric measurements at various ages were taken into consideration for the research. Pedigree data was also collected. For the prediction of body weight, features such as morphometric measurements and body weights at earlier ages and other relevant factors like breed and sex were acquired. These constituted a total of 64 features.

Data cleaning

The raw data was manually cleaned and duplicate values were removed. Unreliable data points like those having the same animal as dam and sire were removed. Removed noisy rows. Identified and removed rows that contain duplicate observations for the animal identification numbers. Data cleaning was done in both Python and R.

Iterative imputation

Missing data cause bias to creep into the analysis thus making it arduous to analyse (Barnard and Meng, 1999). Rows with too many missing variables were removed completely. For the rows within the dataset where the number of missing values were less, imputation was used for filling up the missing variables in the dataset. The missing values were treated as MAR (missing at random) values (Wu *et al*., 2004). Data imputation for the current dataset was done iteratively in Python using the Scikit-learn open-source machine learning library (Pedregosa *et al*., 2011) using Bayesian ridge regression (Tipping, 2001; MacKay, 1992).

Winsorization

Outliers were detected using boxplots and histograms using matplotlib library in Python (Hunter, 2007). To handle the outliers in the dataset, winsonization technique was used. Winsorization was done in Python using the SciPy library (Gerard-Marchant, 2007).

Data types and feature encoding

For the machine learning models two types of encoding were performed.

Label encoding

This type of encoding was done for variables that had too many categories or values. One Hot Encoding: One hot encoding technique was applied to nominal data present in the data set.

Data normalization/standardization

Normalization was done in Python (Pedregosa *et al*., 2011). For multivariate data, this was done feature-wise *i.e*., independently for each feature of the data.

Dimensionality reduction

In order to reduce the number of input variables in the dataset, dimensionality reduction was performed in Python. Two methods were used for this purpose.

1. **Principal Component Analysis (Pearson, 1901):** The PCA was fit on the training set and mapping, or transformation was applied to both the training and test set.
2. **Feature selection:** Through feature selection, an optimal feature subset was selected based on the one that optimized the scoring function which was done using sklearn (Pedregosa *et al*., 2011). Feature selection in Python was done based on an F-test estimate of the degree of linear dependency between two numerical variables: the input and the output. This was treated as a regression predictive modelling problem (Pedregosa *et al*., 2011).

Feature selection was done both for the original datasets as well as the extracted features from PCA. As a result, three separate datasets were created for the prediction of body weight from morphometric measurements: Principal Component analysis (PCA), feature selection (FS), PCA and feature selection (PCA+FS).

Multicollinearity was checked using pair plots in Python (Waskom, 2021). The variance inflation factor was also checked for all variables before analysis.

Principal component regression

Principal component regression was performed on both the datasets *viz*. PCA + FS and PCA. PCR was utilized by finding M-linear combinations (also known as principal components) of the p-predictors and least squares was employed to fit the linear regression model. In this model, principal components were used as predictors (Sutter *et al*., 1992).

The scoring criteria for evaluating the models were mean squared error, mean absolute error, Coefficient of determination value and correlation coefficient. The following percentages of data were used for constructing the model:

1. Testing data (10% of the dataset), training data (90% of the dataset), validation data (10% of training data).

2. Testing data (20% of the dataset), training data (80% of the dataset), validation data (10% of training data).
3. Testing data (20% of the dataset), training data (80% of the dataset), validation data (20% of training data).

Ordinary least squares

A prediction equation for arriving at the 12-month body weight for both the datasets was derived using ordinary least squares in python. The models used had the following structure:

$$Yik = a + biXi + \ldots + bk\ Xk + eik$$

Where:

Yik= 12 month body weight.

a= The intercept.

bk= Regression coefficient estimated.

Xk= Different principal components.

eik= Random error ~ NID (0, σ2e).

The accuracy of fitting the regression models was calculated employing coefficient of determination (R2).

RESULTS AND DISCUSSION

Datasets

Considering all variables having variance above 95% in the principal component analysis (PCA), a total of 23 features were retrieved in the PCA dataset. The FS dataset was created by using features in the original dataset having F scores greater than 10. By this way the number of features within our dataset were 28. The body weights taken at various ages from weaning had the greatest feature scores. This is expected as is also evident from the growth curves of various animals in which body weight is also the most important parameter (Swatland, 1994). For the dataset containing features selected after PCA, 6 features were selected for the final dataset (PCA and Feature Selection). The analysis was carried out at SKUAST-K, J & K, India during the year 2021.

The selected features for PCA and PCA+FS datasets were devoid of multicollinearity. A pairplot of the dataset derived from the PCA results did not contain any multicollinearity as expected (Fig 1). However, a number of features in the dataset containing features derived only from feature selection had high multicollinearity. Eg the weights taken at various months were highly collinear with each other (Fig 2). The variance inflation factors for all PCA variables was 1 from which it may be inferred that the variables were not correlated. The variance inflation factor for most of the features for the FS dataset was greater than 5 and therefore this dataset was not used for further evaluation.

The PCA technique eliminated the multicollinearity within the dataset. PCA is one of the most common ways to reduce multicollinearity in the dataset which has been reported by multiple authors (Rahayu *et al*., 2017; Pinto *et al*., 2006). Our results suggest that PCA allowed for a better understanding of the complex correlations among the traits while ensuring that the number of traits was reduced which was also suggested by Pinto *et al*. (2006). This is especially important as the pressure on farms increases to produce more food. Thus, solutions to problems (Hamadani *et al*., 2021) like quick farm predictions, reducing inbreeding (Nabi *et al*., 2022), must be sought.

Principal component regression

The scoring statistics for models 1, 2 and 3 for PCA and PCA+FS datasets are given in Table 1, 2 and 3 respectively. Similar plots were obtained for models 2 and 3 as well. Our results indicate that PCR is an efficient technique for the prediction of body weights in sheep. For all models the mean squared error, mean absolute error and the root mean squared errors were lower for the PCA testing dataset. The correlation coefficients between predicted and the actual values were also higher for the PCA. The effect of increasing the proportion of validation dataset did not have any effect on the final testing errors or correlations. From this we may infer that FS may have resulted in the loss of some explained

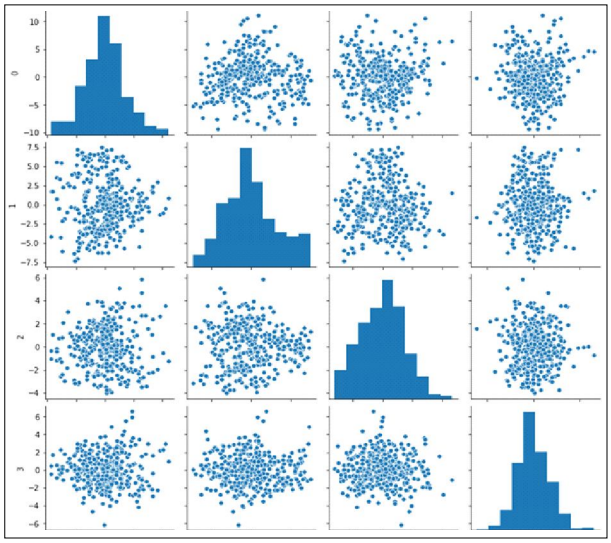

Fig 1: Pair plot of the first 4 variables of the PCA dataset containing 22 variables.

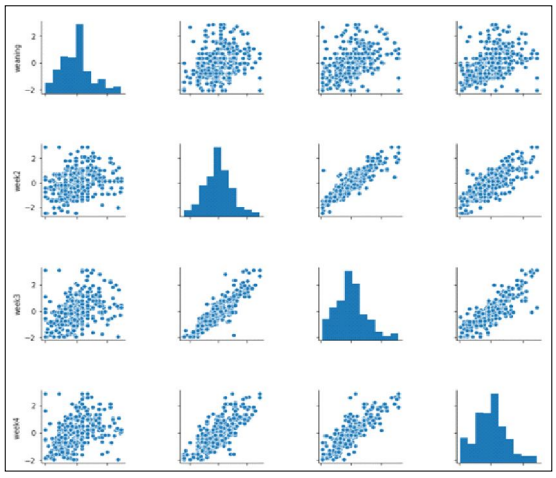

Fig 2: Pair plot of the first 4 variables (body weights at week 2, 3, 4 and weaning) of the dataset without PCA.

variance and hence the increased number of features in the PCA dataset caused the model to converge better with lower error.

The coefficients of determination in our study were high for all models (Table 1, 2, 3). High R2 values (0.988, 0.929, 0.976) by using various machine learning approaches was also reported by Huma and Iqbal, (2019). Valsalan *et al*., (2020) also used principal component analysis in Malabari goats to arrive at the growth performances and found the PCA model to have a coefficient of determination (R2) value equal to 74%. A tenfold cross validation approach was reported to train the best model by Huma and Iqbal, (2019) which also correlated with our results. Hamadani *et al*. (2022) also used principal component regression for the prediction of genetic merit in sheep and reported a correlation coefficient to 0.74 between true and predicted labels.

Table 1: Validation and testing results for model 1 for PCR.

Measure	PCA		PCA + FS	
	Validation	Test	Validation	Test
Mean absolute error	0.344	0.759	0.429	0.865
Mean squared error	0.245	1.169	0.327	1.387
Root mean squared error	0.973	1.081	0.965141	1.178
Coefficient of determination		0.921		0.907
Correlation coefficient		0.979		0.974
Data split	10%	10%	10%	10%

Table 2: Validation and testing results for model 2 for PCR.

Measure	PCA		PCA + FS	
	Validation	Test	Validation	Test
Mean absolute error	0.349	0.559	0.436	0.669
Mean squared error	0.246	0.670	0.356	0.947
Root mean squared error	0.974	0.819	0.962	0.973
Coefficient of determination		0.949		0.928
Correlation coefficient		0.982		0.970
Data split	10%	20%	10%	20%

Table 3: Validation and testing results for model 3 for PCR.

Measure	PCA		PCA + FS	
	Validation	Test	Validation	Test
Mean absolute error	0.353	0.559	0.437	0.669
Mean squared error	0.251	0.670	0.358	0.947
Root mean squared error	0.974	0.819	0.963	0.973
Coefficient of determination		0.949		0.928
Correlation coefficient		0.982		0.970
Data split	20%	20%	20%	20%

The models where a greater proportion of the dataset is allocated to testing had lower mean squared error as well as mean absolute error. From this we can infer that the model trained sufficiently well even for a lesser proportion of data (80%) and could perform well for a greater proportion of testing data (20%) as well. The correlation coefficient was also greater for the model having 20% data allocated for training.

Ordinary least squares

Multiple regression equations for the PCA dataset are given in Table 4. All features constituting the 2nd prediction equation in Table 4 were significant and from that we may infer that in OLS, these variables contribute to the overall prediction of bodyweight. Multiple regression equations for the PCA+FS dataset are given in Table 5. Both models had a high coefficient of determination and were highly significant. However, the R2 value of the PCA dataset was greater than the PCA dataset which correlates with our earlier results obtained by the PCR method. A high coefficient of determination (0.92) was also reported by Kumar *et al*. (2018) for the prediction of adult body weight from linear body measurements while Topai and Macit (2003) and Agamy *et al*. (2015) reported a coefficient of determination of 0.784 and 0.70 for the prediction of body weights in Morkaraman Sheep and three Egyptian Sheep breeds respectively.

In the regression analysis, the second principal component explained most of the variance. Our result is in consonance with Valsalan *et al*. (2020) who reported that the inclusion of the first 2 components accounted for a significant improvement in the amount of variance (R2 = 74%). Pinto *et al*. (2006) also reported that the first five principal components explained nearly 93.3% of the variation and

Table 4: Prediction equations for PCA using OLS.

Prediction equations for PCA	R2
22.4622* - 0.1532×1* -0.8614×2* -0.4704×3* +0.3378×4* +0.4764×5*	-
0.0960×6* +0.0531×7* +0.0484×8* +0.0050×9 -0.0593×10*-0.0209×11	-0.98
0.1310×12* -0.1325×13* -0.0054×14 -0.0044×15 -0.4537×16* -0.2177×17*	+0
0.0886×18* +0.2359×19* +0.0124×20 -0.0718×21 -0.0516×22 +0.0057×23	
22.4622* - 0.1532×1* -0.8614×2* -0.4704×3* +0.3378×4* +0.4764×5*	-
0.0960×6* +0.0531×7* +0.0484×8* -0.0593×10* -0.1310×12* -0.1325×13*	-0.98
0.4537×16* -0.2177×17* + 0.0886×18* +0.2359×19*	0

Table 5: Prediction equations for PCA + FS using OLS (*indicates significance).

Prediction equation	R2
22.4622* - 0.1532×1* -0.8614×2* -0.4704×3* +0.3378×4* +0.4764×5* -0.4537×6*	0.970*
22.4622* - 0.1532×1* -0.8614×2* -0.4704×3* +0.3378×4* +0.4764×5*	0.959*

the first one alone explained about 66%. Khan *et al*. (2013) also reported the first two principal components to show maximum variance (61.86% and 26.14%). The components explaining a majority of the variance can be used for selection and breeding especially for the construction of selection indices (Ibe, 1989).

CONCLUSION

It is concluded that Machine Learning approaches give good results for the prediction of body weights in animals and these approaches can be extrapolated to a number of other situations in animal sciences. Both the datasets trained well, however the dataset without feature selection performed better in both OLS and PCR. Both the tested algorithms are similar in their prediction of the body weights with OLS performing slightly better than PCR in terms of R2 value.

Conflict of interest

All authors declare that they have no conflict of interest.

REFERENCES

Afolayan, R.A., Adeyinka, I.A. and Lakpini, C.A.M. (2006). The estimation of live weight from body measurements in Yankasa sheep. Czech Journal Animal Science. 51: 343-348.

Agamy, R., Abdel-Moneim, A.Y., Abd-Alla, M.S., Abdel-Mageed, I.I. and Ashmawi, G.M. (2015). Using Linear Body measurements to predict body weight and carcass characteristics of three egyptian fat-tailed sheep breeds. Asian Journal of Animal and Veterinary Advances. 10(7): 335-344. 10.3923/ajava.2015.335.344.

Barnard, J. and Meng, X.L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. Statistical Methods in Medical Research. 8(1): 17-36.

Cam, M.A., Olfaz, M. and Soydan, E. (2010a). Possibilities of using morphometrics characteristics as a tool for body weight production in Turkish hair goats (Kilkeci). Asian Journal Animal and Veterinary Advances. 5(1): 52-59.

Cho, H., Seoyoung, J., Mingyung, L., Kyewon, K., Hamin, K., Eunkyu, P., Minkook, K., Seokman, H. and Seongwon, S. (2020). Analysis of the factors influencing body weight variation in hanwoo steers using an automated weighing system. Animals. 10(8): 1270. https://doi.org/10.3390/ani10081270.

Cihan, P., Gökçe, E., Kalıpsız, O. (2017). A review of machine learning applications in veterinary field. Kafkas Univ Vet Fak Derg. 23(4): 673-680.10.9775/kvfd.2016.17281.

Gerard-Marchant, P.G. (2007). Scipy.stats.mstats.winsorize. https://docs.scipy.org/doc/scipy/reference/generated/scipy. stats.mstats.winsorize.html.

Gül, S., Görgülü, Ö., Keskin, M., Bicer, O. and Sarı, A. (2005). Some production equations of live weight from different body measurements in Shami (Damascus) goats. Asian Journal Animal and Veterinary Advances. 4(5): 532-534.

Hamadani, A., Ganai, N.A., Mudasir, S, Shanaz, S., Alam, S. and Hussain, I. (2022). Evaluation of artificial intelligence algorithms for the prediction of genetic merit, 1 April 2022, Preprint (Version 1) available at Research Square. 10. 21203 /rs.3.rs-1488946/v1.

Hamadani, A., Ganai, N.A. and Rather, M.A. (2021). Genetic, phenotypic and heritability trends for body weights in Kashmir Merino Sheep. Small Ruminant Research. 205: 106542. 10. 1016/j.smallrumres. 106542.

Huma, Z. and Iqbal, F. (2019). Predicting the body weight of Balochi sheep using a machine learning approach. Turkish Journal of Veterinary and Animal Sciences. 43: 500-506. 10.3906/vet-1812-23.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering. 9(3): 90-95. https://doi.org/10.1109/MCSE.2007.55.

Ibe, S.N. (1989). Measures of size and conformation in commercial broilers. J. Anim. Breed. Genet. 106(1989): 461-469.

Jolliffe, I.T. (1982). A note on the use of principal components in regression. Journal of the Royal Statistical Society, Series. C. 31(3): 300-303. 10.2307/2348005. JSTOR 2348005.

Khan, H., Muhammad, F., Ahmad, R., Nawaz, G., Rahimullah and Zubair, M. (2006). Relationship of body weight with linear body measurements in goat. Journal of Agricultural and Biological Science. 1(3): 51-54.

Khan, T.A., Tomar, A.K.S., Dutt, T. and Bhushan, B. (2013). Principal component regression analysis in lifetime milk yield prediction of crossbred cattle strain Vrindavani of North India. A Indian Journal of Animal Sciences. 83: 1288-1291.

Kumar, S., Dahiya, S.P., Malik, Z.S. and Patil, C.S. (2018). Prediction of body weight from linear body measurements in sheep. Indian Journal of Animal Research. (52): 1263-1266. doi: 10.18805/ijar.B-3360.

MacKay, D.J.C. (1992). Bayesian interpolation. Computation and Neural Systems. 4(3): 1992.

Moaeen-ud-Din, M., Ahmad, N., Iqbal, A. and Abdullah, M. (2006). Evaluation of different formulas for weight estimation in Beetal, Teddi and crossbred (Beetal × Teddi) goats. Journal of Animal and Plant Sciences. 16: 3-4.

Moncaster, T. (2020). Machine learning as a service-serving reusable ML models. https://www.codemotion.com/magazine/dev-hub/machine-learning-dev/machine-learning-as-a-service-serving-reusable-ml-models/.

Nabi, N., Ganai, N.A., Shanaz, S., Aalam, S., Shabir, M., Majid, R., Bukhari, S., Mir, S.A., Hamadani A. and Rather, M.A. (2022). Effect of inbreeding coefficient on growth and fitness traits in a closed flock of corriedale sheep. Indian Journal of Animal Research. 56(5): 525-530. doi: 10.18805/IJAR.B-4254.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 2(11): 559-572. doi: 10.1080/14786440109462720.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. *et al*. (2011). Scikit-learn: Machine learning in {P}ython. Journal of Machine Learning Research. 12: 2825-2830.

Pesmen, G. and Yardimci, M. (2008). Estimating the live weight using some body measurements in Saanen goat. Archivos de Zootecnia. 11: 30-40.

Pinto, L.F.B., Packer, I.U., De Melo, C.M.R, Ledur, M.C. and Countiho, L.L. (2006). Principal components analysis applied to performance and carcass traits in the chicken. Anim. Res. 55: 419-425.

Rahayu, S., Sugiarto, T., Madu, L., Holiawati and Subagyo, A. (2017). Application of principal component analysis (PCA) to reduce multicollinearity exchange rate currency of some countries in asia period 2004-2014. International Journal of Educational Methodology. 3(2): 75-83.

Rahman, F. (2007). Prediction of carcass weight from the body characteristics of black goats. International Journal of Agriculture and Biology. 9(3): 431-434.

Shirzeyli, F.H., Lavvaf, A. and Songklanakarin, A.A. (2013). Estimation of body weight from body measurements in four breeds of Iranian sheep. J. Sci. Technol. 35(5): 507-511.

Sutter, J.M., Kalivas, J.H. and Lang, P.M. (1992). Which principal components to utilize for principal component regression. Journal of Chemometrics. 6(4): 217-225. doi: 10.1002/cem.1180060406.

Swatland, H.J. (1994). Structure and Development of Meat Animals and Poultry. Switzerland: Taylor and Francis.

Topai, M. and Macit, M. (2003). Prediction of body weight from body measurements in morkaraman sheep. Journal of Applied Animal Research. 25(2): 97-100.

Tipping, E. (2001). Sparse bayesian learning and the relevance vector machine. Journal of Machine Learning Research. 1(2001): 211-244.

Valsalan, J., Sadan, T. and Venketachalapathy, T. (2020). Multivariate principal component analysis to evaluate growth performances in Malabari goats of India. Trop Anim Health Prod. 52: 2451-2460. https://doi.org/10.1007/s11250-020-02268-9.

Waskom, M.L. (2021). Seaborn: Statistical data visualization. Journal of Open-Source Software. 6(60): 3021.10.21105/joss.03021.

Wu, J., Tillett, R., McFarlane, N., Ju, X., Siebert, J.P. and Schofield, P. (2004). Extracting the three-dimensional shape of live pigs using stereophotogrammetry. Computers and Electronics in Agriculture. 44(3): 203-222.

Yilmaz, O. Cemal, I. and Karaca, O. (2013). Estimation of mature live weight using some body measurements in Karya sheep. Tropical Animal Health and Production. 45: 397-403.

Zhang, S., Su, Q. and Chen, Q. (2021). Application of machine learning in animal disease analysis and prediction. Current Bioinformatics. 16(7): 972-982. https://doi.org/10.2174/1574893615999200728195613.