PREDICTING BODY WEIGHT OF THREE CHICKEN GENOTYPES FROM LINEAR BODY MEASUREMENTS USING MARS AND CART DATA MINING ALGORITHMS

ASSAN, N.^{1,2} – Mpofu, M.³ – Musasira, M.³ – Mokoena, K.^{4*} – Tyasi, T. L.⁴ – MWAREYA, N.⁵

¹Zimbabwe Open University, Faculty of Agriculture, Department of Agriculture Management, Bulawayo Regional Campus, Bulawayo, Zimbabwe

²University of South Africa, College of Agriculture and Environmental Sciences, Department of Agriculture and Animal Health, South Africa

³Matopos Research Station, Ministry of Lands and Agriculture, Department of Research and Extension, Private Bag 5137, Bulawayo, Zimbabwe

4 School of Agricultural and Environmental Sciences, Department of Agricultural Economics and Animal Production, University of Limpopo, Private Bag X1106, Sovenga 0727, Limpopo, South Africa

⁵Zimbabwe Open University, Faculty of Agriculture, Department of Mathematics, Mutare Regional Campus, Bulawayo, Zimbabwe

> **Corresponding author e-mail: koenamokoena41@gmail.com; phone/fax: +27-15-268-2803*

> > (Received 22nd Nov 2023; accepted 9th Feb 2024)

Abstract. The aim of the current study was to predict the body weight from linear body measurements of Astrolope, Boschveld and indigenous Sacco genotype using Classification and regression tree (CART) and Multivariate Adaptive Regression Spline (MARS) algorithm. A total of 389 body weight (BW) records, including five continuous predictors such as Neck length (NL), body circumference (BC), shank length (SL), body length (BL) and shank circumference (SC) were used. The best model was selected based on goodness of fit, such as, standard deviation ratio (SDR), root mean square error (RMSE), coefficient of variation (CV), adjusted coefficient of determination (ARsq), coefficient of determination (Rsq) and Pearson's correlation coefficients (PC). The Rsq (%) values ranged from 59 (MARS) to 69 (CART). The lowest SDR was recorded by CART (0.56) and the highest by MARS (0.70). The CART was selected to be the best algorithm with sex, genotype, SC, SL, BL, NL, and BC as influential predictor of BW. The heaviest body weight on females of genotype (Boschveld, Sacco) was recorded when BL was less than 43 cm and BL higher than 47 cm. The goodness of fit criteria suggest that CART model outperformed the MARS model on predicting the body weight of the three genotypes. The findings will assist farmers in the prediction of body wight and selection of heavier chickens.

Keywords: *Indigenous chickens, Astrolope, Boschveld, regression tree, Sacco*

Introduction

Farmers use body weight in their daily management at the farm such as monitoring the feed intake, response to feed and growth rate (Haq et al., 2020). The selection for higher body weight might improve next generations body weight, which is helpful to farmers to improve their herd (Hlokoe et al., 2022). The use of linear body measurements is important in local farmers where weighing scales are not readily available (Nweke-Okorocha and Gunn, 2022). The prediction of body weight plays a major role in influence the selling price, good maintenance, production level, knowing the performance of the offspring, and used in seed selection (Ajayi et al., 2008; Akramullah et al., 2021). Tyasi et al. (2020) employed the tree-based regression tree methods to estimate the body weight of indigenous Potchefstroom Koekoek breed, native to South Africa and Hy-line silver brown commercial layer using morphometric traits. The decision trees (CART, exhaustive Chi-square Automatic Interaction Detector (Ex-CHAID), and Chi-square Automatic Interaction Detector (CHAID)) and MARS algorithms have been used for predicting body weight in different animal species; namely, dogs (Celik and Yilmaz, 2018), cattle (Hlokoe et al., 2022), sheep (Abbas et al., 2021) and goats (Mokoena et al., 2022; Rashijane et al., 2023). However, there is limited literature on the prediction of body weight from linear body measurements of chicken genotypes (Astrolope, Boschveld and Sacco) using data mining algorithm. Hence, CART and MARS algorithm were employed to predict the body weight from linear body measurements of commercial and indigenous Sacco genotype, native to Zimbabwe. The results will help chicken farmers on the best traits to select in predicting body weight in chickens, selection for breeding and production improvement.

Materials and Methods

Study area

The research was conducted at the Matopos Research Station (20 0 23' S, 310 30' E), which is found approximately thirty kilometres south-western of Bulawayo, Zimbabwe. The setting is 800 meters above sea level and receives irregular rainfall of only 450 mm annually (Homann et al., 2007). The temperatures during the summer are quite high, with the average maxima and the lowest temperature of the warmest months being 21.6 and 11.4°C, respectively. According to Hagreveas et al. (2004), there is a chance of severe droughts in the area.

Animal management

A deep litter (semi intensive poultry) house was used to rear three chicken breeds namely Sacco (SC), Astrolope (AC), and Boschveld (BC). The Feeding, medication and watering of the birds were done in accordance with that of Matopos Research Institute Animal Nutrition Section, Zimbabwe. The poultry house was 5m x 6m in dimension with proper ventilation. Each poultry house was having 25 chickens of which five were cocks. In rain and cold season plastic sheeting was used to control room temperature. The chickens were fed grower ration, routine vaccination and other management practices were done. The birds were fed the same feed throughout the experimental period and clean water was supplied ad-libitium. Periodically chickens were supplemented with pearl millet and were allowed to scavenge during the day.

Data sampling

The BW and linear body measurements of 389 chickens ($SC = 82$, $AC = 154$, BC =153) were estimated on 6 months old chickens. The body weight was collected using electronic weighing scale, whereas, linear body measurements $(BC = body)$ Circumference, $BL = Body Length$, $SL = Shank Length$, $SC = Shank Circumference$, NL= Neck Length) were recorded using tape measure and Vernier callipers. The study birds were caught into an empty box, and their individual body weight was recorded. An empty box was used to weigh the chicken, the box was first placed on an electronic scale adjusted the scale to zero rendering the box void and place the birds in the box to record the weights.

Classification and regression tree (CART) algorithm

CART was first proposed by Breiman et al. (1984) as a duplication algorithm tree that is created by continually splitting a node into pairs of child nodes, beginning with the root node, which houses the entire learning sample. CART and CHAID algorithms were reported in detail by Akin et al. (2017).

Multivariate Adaptive Regression Spline (MARS) algorithm

MARS was developed by Friedman (1991), defining it as a non-parametric regression method. The MARS algorithm was conducted as explained by Sengül et al. (2020) in the current study, and its prediction equation can be written as follows:

$$
f(x) = \beta_0 + \sum_{m=1}^{m} \beta_m \lambda_m(x) \tag{Eq.1}
$$

where $f(x)$ is the expected response, β_0 and β_m are parameters that are calculated to give the best data fit, and m is the number of BFs in the model. In the MARS model, the basis function composed of be a single univariable spline function or a combination of more than one spline function for diverse predictor inputs. The spline BF, $\lambda_m(x)$, is defined as:

$$
\lambda_m(x) = \prod_{k=1}^{k_m} [s_{km}(X_{\nu(k,m)} - t_{k,m})]
$$
 (Eq.2)

where $t_{k,m}$ denotes the knot location; s_{km} denotes the right/left regions of the corresponding step function, taking either 1 or -1; v(k, m) denotes the predictor variable's label; and km is the number of knots. Following the procedure of Sengül et al. (2020), the pruning process was used to remove the basic functions that had a low contribution to the model fitting performance following the generalised cross-validation error (GCV):

$$
GCV(\lambda) = \frac{\sum_{i=1}^{n} (y_i - y_{ip})^2}{(1 - \frac{M(\lambda)}{n})^2}
$$
 (Eq.3)

where n represents the number of training cases, yⁱ shows the observed value of the responsible variable, y_{ip} as the estimated value of the response variable and $M(\lambda)$ represents the penalty function for the complex of the model with λ terms.

The following goodness of fit test criteria were computed for training and test datasets: Coefficient of Determination (Rsq):

$$
\text{Rsq} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
$$
(Eq.4)

Pearson's correlation coefficient (r):

$$
r = \frac{cov(y_i y_{ip})}{s_{yi} s_{Yip}}
$$
 (Eq.5)

Adjusted Coefficient of Determination (Adj. R^2):

$$
Adj. R^2 = 1 - \frac{\frac{1}{n-k-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}
$$
(Eq.6)

Standard deviation ratio (SD_{ratio}):

$$
SD_{ratio} = \sqrt{\frac{\frac{1}{n-1} \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{Eq.7}
$$

Mean absolute percentage error (MAPE):

$$
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}}{y_i} \right| .100
$$
 (Eq.8)

Akaike Information Criteria (AIC):

$$
AIC = nln\left[\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}\right] + 2k
$$
 (Eq.9)

Relative approximation error (RAE):

$$
RAE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)}{\sum_{i=1}^{n} Y_i^2}}
$$
(Eq.10)

Root-mean-square error (RMSE):

$$
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
$$
 (Eq.11)

Coefficient of variation (CV):

$$
CV(\%) = \sqrt{\frac{\frac{1}{n-1} \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2}{\overline{Y}}} \times 100
$$
 (Eq.12)

Performance index (PI):

$$
PI = \frac{rRMSE}{1+r} \tag{Eq.13}
$$

Mean absolute deviation (MAD):

$$
MAD = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100
$$
 (Eq.14)

Statistical analysis

The data was analysed using Statistical Package for the Social Sciences (IBM SPSS, 2022) v.29.0 (IBM Corp., NY, USA), with a probability of 5% for significance. The data set was split into two data sets, training (54) and test sets at (46) proportions. A ten-fold cross-validation resampling method was used to select the best MARS models with degree $= 1:9$ and n prune $= 2:38$ as a number of selected terms within the scope of live body weight estimation in the training set. The statistical evaluation of the MARS data mining algorithm for the prediction of body weight was performed using Package in R. EhoGof package (version 0.1.1, Igdir, Turkiye) developed by Eyduran, 2022 in R was implemented to reveal the predictive performance of the optimal MARS model.

Results

The correlation between the linear body measurements and BW are represented by the heat map in *Fig. 1*. The BW had a statistically significant correlation ($p < 0.05$) with BL, SL, NL and SC, however, BW showed no significant correlation with BC. NL and SC showed the highest significant correlation at $p < 0.01$. BC had no statistical correlation with BL, BW, NL, and SC, however, it showed to be statistically correlated with SL ($p <$ 0.05).

*Figure 1. Heat map of Pearson's correlation between linear body measurements and body weight. Pearson correlation colour illustration, a high correlation is red, mid correlation is white and low correlation is blue. BW: body weight; NL: neck length; BC: body circumference; BL: body length; SL: shank length; SC: shank circumference, * significant at P < 0.05, ns not significant and ** significant at P < 0.01*

The MARS model performance

Table 1 gives the results of performance of the MARS model of training and test dataset based on goodness of fit. The results revealed that the training data set for the proportion 54% (training) and 46% (test) achieved the best predictive model. The training set had the lowest RMSE, SDR, CV, RAE, PI, MAD and MAPE. The Pearson's correlation coefficients (PC) and Coefficient of determination (Rsq) values of the training set were higher than that of the test set.

Table 2 shows the MARS data mining algorithm model which included GENSC, SEXM, and NL. The MARS model yielded 3 basic functions of single order term variation with an intercept of 1318.45. The effect of linear body measurements with the positive and negative coefficient on BW were described by MARS. The cock had a constructive influence on BW with a coefficient of 360.68. The influence on BW was 49.66 in the positive direction when $NL > 12$ cm, however, the GENSC had a negative influence on BW with a coefficient of -152.76.

The optimal MARS predictive model is presented below:

 $y = 1318.45$

- 152.76 * GENSC

- + 360.68 * SEXM
- $+$ 49.66 $*$ max (0, NL 12)

where 1318.45 is the first term known as intercept on MARS prediction equation for body weight, GENSC is the second term with coefficient of - 152.76, the third term is SEXM whose coefficient is 360.68 . max $(0, NL - 12)$ is the fourth term of the model whose cut point is -0.12 units in NL with coefficient of 49.66.

Decision tree diagram to estimate body weight by CART

The regression tree diagram created by CART algorithm in predicting BW is presented in *Fig. 2*. Sex, Gen (BC, SC), SC, SL, BL, NL, and BC were the influential predictor of BW. The 1608 g was recorded at the top regression diagram. At first depth, SEX-F had an influence on BW with hens weighing 219 g lighter than the average BW. At second tree depth, the hens recorded 18% of the BW with $NL < 14$ cm and 35% of BW with $NL > 14$ cm. the third tree depth recorded GEN = SC and $SL > -4.3$ cm recorded 3% of the BW and SL < -4.3 recorded 23% of the BW. In Cock, at first tree depth, SEX-F had an influence on BW with cock weighing 248 g heavier than the average BW. The GEN $= BC$, SC further divided second tree depth into third and fourth tree depth. GEN=BC explain 28% of BW, which divided into third tree depth with SC < 3.7 cm with 9% BW, $SC > 3.7$ cm with 19% of BW. and GEN= SC explained 19% of BW. The GEN = SC was further divided into $BL > -47$ cm which recorded 15% of BW.

Figure 2. Regression tree diagram constructed by CART algorithm

The CART model performance

The performance of the CART model of training and test dataset based on goodness of fit results are given in *Table 3*. The outcome revealed that the raining data set for the proportion 54% (training) and 46% (test) achieved the best predictive model. The training set had the lowest RMSE, SDR, CV, RAE, PI, MAD and MAPE. The Pearson's correlation coefficients (PC) and Coefficient of determination (Rsq) values of the training set out performed that of the test set.

Criterion	Train	Test
Relative approximation error (RAE)	0.02	0.04
Mean error (ME)	0.000	19.33
Relative root mean square error (RRMSE)	12.42	20.84
Akaike's information criterion (AIC)	1875.17	2503.62
Mean absolute percentage error (MAPE)	10.08	110.00
Coefficient of variation (CV)	12.46	20.85
Pearson's correlation coefficients (PC)	0.83	0.55
Standard deviation ratio (SDR)	0.56	0.89
Corrected Akaike's information criterion (CAIC)	1875.17	2503.62
Mean relative approximation error (MRAE)	0.01	0.01
Adjusted coefficient of determination (ARsq)	0.69	0.20
Performance index (PI)	6.78	13.44
Mean absolute deviation (MAD)	152.17	239.98
Coefficient of determination (Rsq)	0.69	0.20
Root mean square error (RMSE)	199.75	337.78

Table 3. Predictive performance of CART model for training and test data set

Discussion

Body weight provides adequate information on the biometric structure of the animal as well as its physiological conditions (Abbas et al., 2021). The current study showed a significant correlation between linear body measurements and body weight in three chicken genotypes. The findings are similar to the results of Sadick et al. (2020) who concluded that linear body measurements are correlated to body weight in Cobb broiler chickens. Ukwu et al. (2014) conducted a study on Nigerian indigenous chickens and came to conclusion that the body weight is significantly correlated to body weight with shank length. The results of the study suggest that body weight may be improved using body length, neck length, shank length and shank circumference in the investigated chicken genotypes. Although correlation gives the association between body weight linear and body measurements, it does not give the effect of linear body measurements on body weight (Rashijane et al., 2023). Hence, Multivariate Adaptive regression splines (MARS) and Classification and Regression Tree (CART) were used to determine the effect of linear body measurements on body weight in the present study. The results indicated that the CART model was the superior model to predict body weight with the highest coefficient of correlation, coefficient of determination, lowest RMSE and SDR. The CART model described 69% of variation in body weight, with sex, genotype, shank circumference, shank length, body length, neck length and body circumference being the influential predictor of body weight. However, the MARS model recognized males and genotype SC as influencers on the body weight. Due to lack of literature on the comparison on MARS and CART in chickens, the discussion used different animal species. Faraz et al. (2021) reported that the MARS model is best predictor of body weight in Pakistan Thali sheep with goodness of fit of $R2 = 0.90$, Adj. $R2 = 0.89$, SD ratio = 0.312 and $r = 0.95$. The study conducted in Turkish Tazi Dogs using MARS and CART to predict body weight indicated that MARS outperformed CART with the goodness of fit of R2 = 0.92, Adj. R2 = 0.89, $r = 0.96$, SD ratio = 0.28 (Celik and Yilmaz, 2018). The findings of Hlokoe et al. (2022) showed the highest $R2 = 0.993$, Adj. $R2 = 0.991$ with the SD ratio $= 0.081$ and lowest RMSE $= 5.97$ indicating that the MARS model is the best fit model as compared with the developed CART model. Celick (2019) reported the MARS

model as the best predictor of body weight in Pakistan goats with the goodness of fit of RMSE = 3.32, R2 = 0.91, Adj. R2 = 0.86, SD ratio = 0.30 and $r = 0.95$.

Conclusion

The finding of the study revealed a positive association between body weight and linear body measurements. The results imply that body length, shank circumference, neck length and shank length can be used to estimate the body weight. According to the results of MARS algorithm, neck length had a positive influence on body weight of males, whereas genotype (Sacco) had a negative influence on the body weight. The CART algorithm recorded the heaviest body weight on females of genotype (Boschveld, Sacco) with $43 \leq BL > 47$. The goodness of fit criteria suggest that CART model was the best predictor of body weight. The findings will assist farmers in prediction of body weight and selection of heavier chickens. There is a need more studies to be performed on the comparison of other data mining algorithms in chickens to predict body weight using linear body measurements.

Acknowledgements. Authors would like to extend their deepest appreciation to the Matopos Research Station for allowing us to use their chickens for data collection.

REFERENCES

- [1] Abbas, A., Ullah, M. A., Waheed, A. (2021): Body weight prediction using different data mining algorithms in Thalli sheep: A comparative study. – Veterinary World 14(9): 2332- 2338.
- [2] Ajayi, F. O., Ejiofor, O., Ironkwe, M. O. (2008): Estimation of body weight from linear body measurements in two commercial meat-type chicken. – Global journal of agricultural sciences 7(1): 57-59.
- [3] 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: 303-316.
- [4] Akramullah, M., Badaruddin, R., Adrianto, A., Tiya, N. A. D., Dewi, Y. L. (2021): Phenotypic Correlation between Body Weight and Body Part Size of Native Chickens in Moramo District, South Konawe Regency, Indonesia. – Advances in Biological Sciences Research, volume 20 Proceedings of the International Conference on Improving Tropical Animal Production for Food Security (ITAPS 2021).
- [5] Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984): Classification and regression trees. – Monterey, CA: Wadsworth and Brooks.
- [6] Celik, S. (2019): Comparing predictive performances of tree-based data mining algorithms and MARS algorithm in the prediction of live body weight from body traits in Pakistan goats. – Pakistan Journal of Zoology 51(4): 1447-1456.
- [7] Celik, S., Yilmaz, O. (2018): Prediction of Body Weight of Turkish Tazi Dogs using Data Mining Techniques: Classification and Regression Tree (CART) and Multivariate Adaptive Regression Splines (MARS). – Pakistan Journal of Zoology 50(2): 575-583.
- [8] Eyduran, E. (2020): EhaGoF: Calculates Goodness of Fit Statistics. R package Version 0.1.1. – Available online: https://CRAN.R-project.org/package=ehaGoF (accessed on 13 October 2022).
- [9] Faraz, A., Tirink, C., Eyduran, E., Waheed, A., Tauqir, N. A., Nabeel, M. S., Tariq, M. M. (2021): Prediction of live body weight based on body measurements in Thalli sheep under

tropical conditions of Pakistan using cart and mars. – Tropical Animal Health and Production 53(2): 301.

- [10] Friedman, J. H. (1991): Multivariate Adaptive Regression Splines. Annals of Statistic 19: 1-141.
- [11] Hagreveas, S. K., Bruce, D., Beffa, L. M. (2004): Disaster mitigation options for livestock production in communal farming systems in Zimbabwe. – 1. Background information and literature review. P.O. Box 776, Bulawayo, Zimbabwe: ICRISAT and FAO Rome, Italy, 56.
- [12] Haq, M. S., Budisatria, I. G. S., Panjono, P., Maharani, D. (2020): Prediction of live body weight using body measurements for jawa brebes (jabres) cattle. – Journal of Animal and Plant Sciences 30(3): 552-559.
- [13] Hlokoe, V. R., Mokoena, K., Tyasi, T. L. (2022): Using multivariate adaptive regression splines and classification and regression tree data mining algorithms to predict body weight of Nguni cows. – Journal of Applied Animal Research 50(1): 534-539.
- [14] Homann, S., van Rooyen, A. F., Moyo, T., Nengomahsa, Z. (2007): Goat production and marketing: Baseline information for semi-arid Zimbabwe. – P.O. Box 776. Bulawayo, Zimbabwe: International Crops Research Institute for the Semi-Arid Tropics, 84.
- [15] IBM SPSS (2022): Statistical packages for social sciences for windows: base system user's guide. – IBM statistics, 29. Chicago: SPSS Inc. doi:10.2527/ jas.2013-6967.
- [16] Mokoena, K., Molabe, K. M., Sekgota, M. C., Tyasi, T. L. (2022): Predicting body weight of Kalahari Red goats from linear body measurements using data mining algorithms. – Veterinary World 15(7): 1719-1726.
- [17] Nweke-Okorocha, G. O., Gunn, H. H. (2022): Estimation of Body Weight Using Linear Body Measurements of Fulani Ecotype Chickens in The Southwestern Nigeria. – Journal of Research in Agriculture and Animal Science 9(3): 15-22.
- [18] Rashijane, L. T., Mokoena, K., Tyasi, T. L. (2023): Using Multivariate Adaptive Regression Splines to Estimate the Body Weight of Savanna Goats. – Animals 13: 1146.
- [19] Sadick, A. M., Aryee, G., Poku Jr, P. A., Kyere, C. G. (2020): Relationship between body weight and linear body measurements in the Cobb broiler chicken. – World Journal of Biology Pharmacy and Health Sciences 4(2): 001-006.
- [20] Sengül, T., Çelik, S., Sengül, Ö. (2020): Use of multivariate adaptive regression splines (MARS) for predicting parameters of breast meat in quails. – Journal of Animal and Plant Sciences 30: 786-793.
- [21] Tyasi, T. L., Eyduran, E., Celik, S. (2021): Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa. – Tropical Animal Health and Production 53: 7.
- [22] Ukwu, H. O., Okoro, V. M. O., Nosike, R. J. (2013): Statistical Modelling of Body Weight and Linear Body Measurements in Nigerian Indigenous Chicken. – Journal of Agriculture and Veterinary Science 7(1): 27-30.