Veterinary Medicine and Science

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Ruminants

Body Weight Estimation in Holstein × Zebu Crossbred Heifers: Comparative Analysis of XGBoost and LightGBM Algorithms

Jose Herrera-Camacho[1] | Cem Tırınk[2] | Rosa Inés Parra-Cortés[3] | Lütfi Bayyurt[4] | Rashit Uskenov[5] | Karlygash Omarova[6] | Aizhan Makhanbetova[6] | Kadyrbai Chekirov[7] | Alfonso Juventino Chay-Canul[8]

[1]Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, Mexico | [2]Department of Animal Science, Igdir University, Faculty of Agriculture, Iğdır, Türkiye | [3]Universidad de Ciencias Aplicadas y Ambientales U.D.C.A, Área de Ciencias Agropecuarias, Grupo de Investigación en Ciencia Animal, Bogotá, Colombia | [4]Faculty of Agriculture, Department of Animal Science, Tokat Gaziosmanpaşa University, Tokat, Türkiye | [5]Agronomic Faculty, Saken Seifullin Kazakh Agrotechnical University, Astana, Kazakhstan | [6]Faculty of Veterinary and Livestock Technology, Saken Seifullin Kazakh Agrotechnical University, Astana, Kazakhstan | [7]Kyrgyz-Turkish Manas University, Bishkek, Kyrgyz Republic | [8]División Académica de Ciencias Agropecuarias, Universidad Juárez Autónoma de Tabasco, Villahermosa, Tabasco, México

Correspondence: Cem Tırınk (cem.tirink@gmail.com)

Received: 17 February 2025 | **Revised:** 16 April 2025 | **Accepted:** 16 May 2025

Funding: The authors received no specific funding for this work.

Keywords: body weight prediction | crossbred heifer | LightGBM | machine learning | XGBoost

ABSTRACT

This study evaluates the effectiveness of XGBoost and LightGBM algorithms for estimating the live weight of Holstein×Zebu crossbred heifers. The study compares the performance of both algorithms using a wide range of biometric measurements and tests various hyperparameter settings. The research results show that the XGBoost algorithm provides almost perfect agreement with an R^2 value of 0.999 on the training set and high performance with an R^2 value of 0.986 on the test set. The LightGBM algorithm also achieved effective results with R^2 values of 0.986 and 0.981 on both training and test sets. The machine learning algorithms used in the current study stand out as having the potential to provide a practical and economical solution for live weight estimation in livestock enterprises and especially for herd management applications in rural areas through input variables such as body measurements, milk yield, etc. However, the obtained results in the current study reveal the potential of machine learning algorithms for live weight estimation in the livestock sector and indicate that advanced research is needed for the optimisation of these algorithms.

1 | Introduction

Cattle are one of the animal species suitable for domestication due to their herbivorous and fast-growing characteristics, and they have existed as an integral part of human society in various environments since the beginning of civilisation (Felius et al. 2014). The domestication of cattle began after the domestication of sheep and goats, which were smaller and easier to manage. In this context, some studies are needed in cattle breeding to increase the benefits obtained from animal products obtained from these domesticated animals. Livestock-based products include milk, meat and leather; thanks to these products, cattle have become one of the most important domestic animal species that serve humanity and provide economic income. Among these products, meat production is one of the most critical aspects of cattle breeding. Determining the live weight, which directly affects meat production, is of great importance because live weight varies depending on many factors such as the animal's health status, nutritional quality and genetic potential (Godfray et al. 2018; Tırınk 2022). Therefore, accurate and effective determination of live weight in the meat production process is a factor that directly affects both economic efficiency and sustainable livestock practices (Warner et al. 2010).

Meat production is a critical component of cattle breeding for both human health and rural development. As a source of protein, meat is one of the basic elements of a balanced diet and thus plays an important role in protecting public health (Li 2017). On the other hand, beef production stimulates economic activities in rural areas, provides a source of income for agricultural enterprises and thus contributes to the revitalisation of rural economies. In this context, sustainable meat production methods have the potential to increase economic efficiency while minimising environmental impact and should be at the core of rural development strategies (Jónsdóttir and Gísladóttir 2023).

In the tropical regions of Mexico, sustainable meat production of cattle breeds such as Zebu and Brahman, which have high environmental adaptability, supports the efficient use of natural resources in these regions and makes significant contributions to the rural economy (Parra-Bracamonte et al. 2015; Domínguez-Viveros 2023).

In addition, the most widely used system in animal breeding for milk and meat production in tropical regions and Mexico is the dual-purpose production system, which provides both milk and meat yield. In these production systems, crossbreds of Bos taurus and Bos indicus species (especially Holstein × Gyr) are generally preferred. Milk production is considered an important food source for both calf nutrition and human consumption (Román-Ponce et al. 2013; Vázquez-Martínez et al. 2024). Due to the high-temperature stress and disease pressure seen in tropical and subtropical climate conditions, it is reported that crossbred cows (Bos taurus × Bos indicus) perform better, especially in large-area and pasture animal breeding systems (Peralta-Torres et al. 2021). Therefore, the F1 Holstein × Zebu crossbreds used in the current study are among the genotypes widely preferred in tropical regions, in line with both adaptation to environmental conditions and dual-purpose production aims.

Meat production plays a major role in meeting nutritional security and protein needs, especially in developing countries (Smith et al. 2013). In this context, the development of effective and sustainable meat production methods is of strategic importance for both regional development and global food security. The development of cattle farming with meat-oriented approaches increases Mexico's agricultural production capacity while also directly contributing to the improvement of living standards in rural areas.

It will be easier to achieve global food security and sustainable development goals with scientific studies to meet the meat demand due to the increasing world population (Pérez-Escamilla 2017). In this context, it will be easier to achieve this goal with new methods to be implemented, such as developing innovative breeding methods, improving animal welfare and accelerating genetic improvement studies, increasing efficiency in meat production and ensuring environmental sustainability of production processes.

While enabling the development of innovative solutions in livestock sectors such as cattle breeding, scientific studies are needed to strengthen rural economies. In this context, knowing the live weights of animals is essential in animal husbandry activities, determining the breeding strategy of the herd and herd management. Knowing the live weights will facilitate the calculation of the optimum amount of feed per animal in the herd, the determination of drug doses, the more reliable determination of marketing prices and the determination of the optimum slaughter times of animals (Tırınk 2022).

In the literature, multivariate statistical methods based on biometric measurements have been used to provide breed characterisation in terms of live weight (Khan et al., 2014; Ali et al., 2015; Faraz et al., 2021; Yavuz and Şahin, 2022). Especially in multivariate statistical methods where the classical regression approach is used, some assumptions that need to be provided may not be provided depending on the structure of the data. In this case, the applicability of the classical statistical approach may be limited. If these assumptions are neglected, the accuracy and validity of the model may be affected and misleading results may be obtained. Therefore, it is important to check these assumptions meticulously and, when necessary, to use alternative methods, especially more flexible modelling techniques (Mendes and Akkartal, 2009; Sahin et al., 2018; Tırınk, 2020; Yavuz and Şahin, 2022). Algorithms such as data mining and machine learning, which have been proposed to eliminate such problems, can be used for this purpose (Tyasi et al., 2021; Kurnaz et al., 2021). In this context, several researches have been implemented in the scope of multivariate statistical methods to predict several traits such as birth weight, weaning weight, live weight and final weight in various species and breeds (Eyduran et al., 2017; Eyduran et al., 2019). The primary aim of predicting these traits is to define how these traits vary between the species and dissimilar breeds within the same species. These changes can provide significant convenience in decision-making, especially in herd management processes (such as selection practices and breed characterisation). This situation will contribute significantly to the development of sustainable animal husbandry.

For this aim, in the current study, the live weight was tried to be estimated from several body measurements with modern multivariate statistical approaches such as LightGBM and XGBoost algorithms.

2 | Material and Methods

2.1 | Animals

Body weight and biometric measurements were performed on 100 crossbred dairy cows (Holstein × Zebu). In this context, the animals used in the current study mainly consist of F1 generation individuals of Holstein × Zebu crossbreeds. Without supplementary feed, the cows were fed on pastures containing Cynodon nlemfuensis (star grass) and Brachiaria humidicola (humidicola grass). Data was collected on a commercial farm called 'Rancho la Esperanza', located in the municipality of

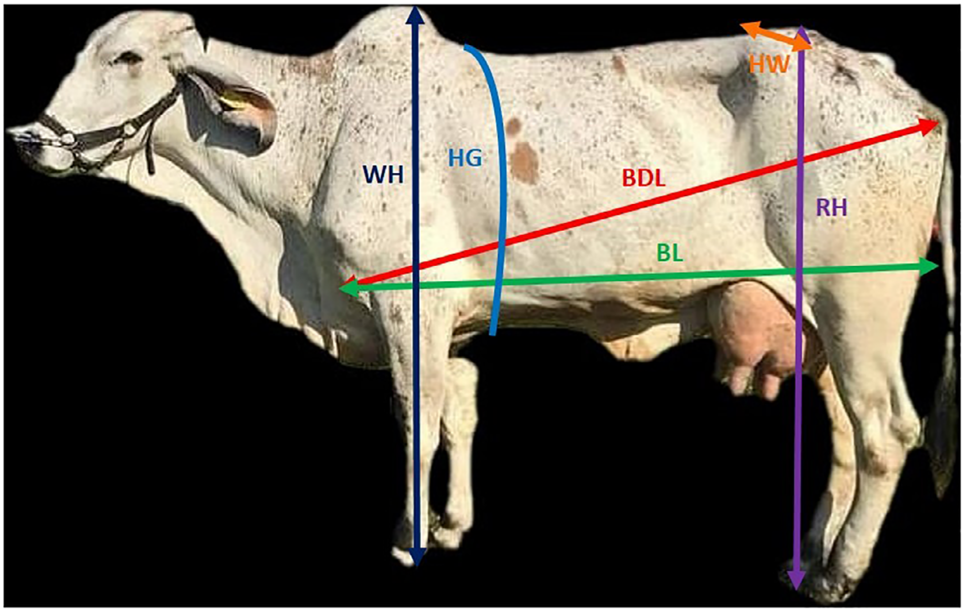

FIGURE 1 | Biometric measurements.

Juarez, Chiapas, Mexico, at an altitude of 120 m above sea level. This farm is located 10 km from the Juarez-Reforma Road and its coordinates are recorded as latitude 17°36′27″N and longitude 93°11′35″W.

The data set was divided into two different sets, such as training and testing sets. In this context, the usage ratio for the training and testing sets was 80%-20%, respectively. Biometric measurements were expressed in centimetres and heart girth (HG), withers height (WH), rump height (RH), hip width (HW), body length (BL) and body diagonal length (BDL) (Figure 1) were recorded according to the methods established by Oliveira et al. (2013) and Bretschneider et al. (2014). A flexible fibre tape measure (Truper) and a large calliper (Haglof, Sweden) of 65 cm in length were used for measurements. Animals were weighed on a fixed platform weighing scale with a capacity of 2000 kg and a sensitivity of 1 kg.

2.2 | A Highly Efficient Gradient Boosting Decision Tree (LightGBM)

The LightGBM algorithm used in this study, which was planned to estimate live weight from various body measurements, was proposed by Microsoft's Distributed Machine Learning Toolkit (DMTK) in 2016 (Ke et al. 2017; Li et al. 2024). The LightGBM algorithm uses histogram-based algorithms that separate continuous feature values into discrete bins (Alsabti et al. 1998, Li et al. 2007). The LightGBM algorithm has a fast training speed and lower memory usage. Besides, it supports GPU and parallel learning simultaneously and can process large datasets (Cai et al. 2021; Li et al. 2024). Additionally, a histogram subtraction technique that reaches the target leaf after removing neighbouring leaves using the parent of the target leaf can contribute to accelerating convergence (Cai et al. 2021). LightGBM uses the leaf-based tree growth method, which selects the leaf with the highest delta loss during growth (Cai et al. 2021). There are some tuning parameters in the implementation phase of the LightGBM algorithm (Table 1).

2.3 | eXtreme Gradient Boosting Algorithm (XGBoost)

The XGBoost algorithm was developed by Chen and Guestrin in 2016. This algorithm was designed with the aim of providing a model that is continuously improved with the contributions of various scientists (Chen and Guestrin 2016). XGBoost uses parallel tree boosting techniques, which are known for their ability to solve data science problems quickly and effectively (Punuri et al. 2023). It is designed as an optimised distributed gradient boosting library and has been effectively used in various fields such as supervised learning, natural language processing and recommender systems. XGBoost works by integrating a set of decision trees trained on different subsets of the data and combines these trees in a way that minimises the total amount of error. Unlike other boosting algorithms, XGBoost includes regularisation mechanisms, which help prevent overfitting of the model (Gertz et al. 2020). In applications such as predicting the weight of heifers, the combined results of XGBoost's decision trees provide more accurate predictions than those of a single decision tree. Moreover, thanks to its regularisation capability, XGBoost is a modelling tool that is more resistant to overfitting (Sagi and Rokach 2021). These features make XGBoost particularly powerful in heifer body weight predictions.

As goodness-of-fit criteria, root mean square error (RMSE), mean absolute error (MAE) and determination coefficient (R^2) are used to compare the prediction models. In addition, sensitivity analysis was performed by taking into account the variable importance values for the best estimation algorithm.

All statistical assessments were completed using R and Spyder software (R Core Team 2020; Raybaut 2009). Descriptive statistics

TABLE 1 | Hyperparameters of the LightGBM model.

Parameters	Description
num_leaves	It determines the number of leaves in each decision tree. This parameter directly affects the complexity and capacity of the model. Since LightGBM is a tree-based learning algorithm, it is of great importance how deep each tree can go and how many detailed features it can learn.
min_data_in_leaf	Setting this parameter to a higher value may limit the resulting tree from getting too deep.
learning_rate	Determines the step size at each iteration as we move towards the minimum of the loss function.

TABLE 2 | Descriptive statistics.

	Mean ± Std. deviation	Min	Max	CV (%)
BW	318.70±99.95	182	550	31.36
HG	161.77±17.64	133	198	10.91
WH	125.62±11.27	104	157	8.97
RH	128.87±8.86	108	155	6.87
HW	41.94±6.84	30	57	16.31
BL	75.31±9.35	60	103	12.42
BDL	93.80±9.30	76	115	9.91

were used to obtain detailed information about the data set. Descriptive statistics for explanatory and response variables were performed using the 'psych' package of the R program (Revelle 2017). For a visual representation of the relationship between explanatory and response variables, Pearson correlation analysis was performed using the 'corrplot' package in R (Wei et al. 2017). The 'LightGBM' package was used to apply the LightGBM algorithm used to estimate BW from the body measurements (Shi et al. 2023). In addition, the Python programming language was used to create 3D surface graphics and analyses were performed via the Spyder interface. In this context, the mpl_toolkits.mplot3d module of the Matplotlib library in the Python programming language was used for drawing 3D surface graphics (Hunter 2007).

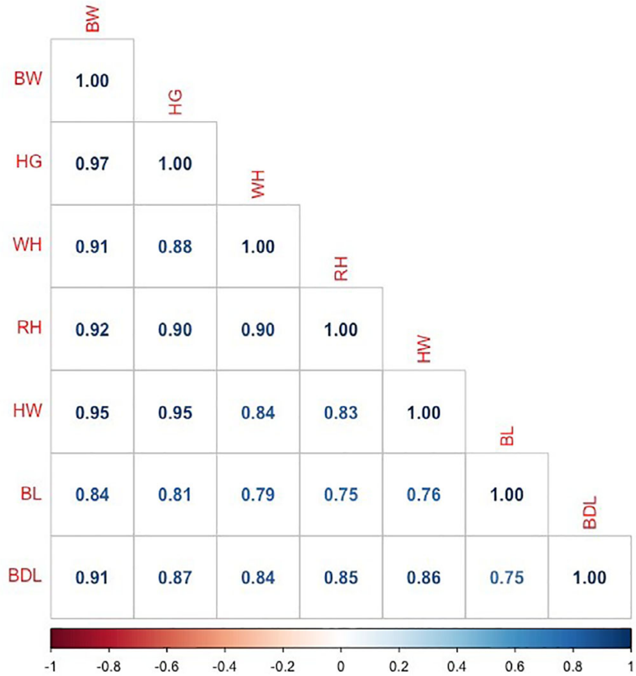

FIGURE 2 | Pearson correlation analysis results.

3 | Results

Table 2 contains statistical data on various biometric measurements, including body weight (BW), heart girth (HG), withers height (WH), rump height (RH), hip width (HW), body length (BL) and BDL. The means, standard deviations, minimum and maximum values and coefficients of variation (CV %) in the table describe the general distribution of the biometric characteristics of the measured animals and the variance of these characteristics within the population.

According to Table 2, BW stands out as the measurement with the highest variation (CV 31.36%) and exhibits a wide distribution between 182 kg and 550 kg. This large variance may indicate that the studied population shows significant heterogeneity under genetic or environmental influences. On the other hand, measurements such as RH and WH have a relatively more homogeneous distribution with very low coefficients of variation of 6.87% and 8.97%, respectively. This situation indicates that certain traits are more stable within the population and perhaps selective breeding practices have an effect on these traits. Other measurements such as HW and BL have average levels of variability and reveal that these measurements also have significant differences among individuals within the population. In particular, HW and BL measurements may be affected by factors such as selection criteria or differences in breeding conditions. In conclusion, the data in Table 2 show that biometric traits show a wide variation and that these traits reflect inter-animal and/or intragroup differences. Such data may be of critical importance in developing animal breeding and selection strategies, as different biometric measurements provide important clues regarding the health, productivity and adaptability of animals. Detailed analysis of these measurements may form the basis for breed development and genetic selection studies.

The correlation matrix in Figure 2 quantitatively presents the relationships between various biometric measurements.

According to Figure 2, in particular, the high correlations between BW and HG and HW (0.97 and 0.95, respectively) indicate that these measurements have a strong relationship with the overall body size of the animals. Similarly, a high correlation of 0.90 between WH and RH indicates that these two measurements are closely related. On the other hand, the value of 0.91 between

TABLE 3 | Hyperparameters for XGBoost.

Hyperparameters	Min	Max	Increment
eta	0.05	0.30	0.05
max_depth	3	10	1
min_child_weight	1	6	1

TABLE 4 | Hyperparameters for LightGBM.

Hyperparameters	Min	Max	Increment
learning_rate	0.01	0.10	0.001
num_leaves	10	100	5
min_data_in_leaf	10	100	5

BL and BDL indicates that these two measurements reflect the dimensions of the animal in a similar way. These findings provide important data for use in animal breeding and selective breeding programs, and reveal how these measurements are related to each other under genetic or environmental influences. Understanding such relationships is critical for developing breeding strategies.

Table 3 shows the ranges of the three basic hyperparameters selected for the XGBoost algorithm and the increments of these ranges. The parameter 'eta', which determines the learning rate, varies between 0.05 and 0.30 and is adjusted in 0.05 increments. This controls how much the model changes at each step and how fast the learning process progresses. The parameter 'max_depth' limits the maximum depth of the trees to values from 3 to 10, which indicates how deep the model can learn on the data. Finally, the parameter 'min_child_weight' determines the minimum weight threshold of the branch and is examined with values from 1 to 6. Careful tuning of these parameters helps to prevent both overfitting of the model and ensure sufficient learning performance.

Table 4 details the hyperparameter settings for LightGBM. The 'learning_rate' parameter was examined from 0.01 to 0.10 in 0.001 intervals to adjust the learning rate of the model. This fine-tuning is critical for the model to generalise better on different datasets. 'Num_leaves', the number of leaves per tree, is another parameter that determines the complexity of the model and was evaluated in increments of 5 between 10 and 100. Finally, the 'min_data_in_leaf' parameter determines the minimum number of data that should be present in a leaf node, which was also adjusted in increments of 5 between 10 and 100. These parameters were strategically chosen to optimise the performance of LightGBM and reduce overfitting, especially on large datasets.

Table 5 presents the results of the goodness of fit criteria for the XGBoost and LightGBM algorithms according to the optimum hyperparameter values. For both models, R^2, RMSE and MAE values on the training and test sets were examined.

According to Table 5, in the XGBoost model, the selected hyperparameters were determined as eta value 0.1, maximum depth (max_depth) 8 and minimum child weight (min_child_weight) 1. The model showed an almost perfect fit in terms of R^2 value

TABLE 5 | The goodness of fit criteria results according to optimum hyperparameter values for each model.

Hyperparameters of the model					
XGBoost			**LightGBM**		
eta	0.1		learning_rate	0.028	
max_depth	8		num_leaves	10	
min_child_weight	1		min_data_in_leaf	10	
Goodness of fit criteria					
XGBoost	**Train**	**Test**	**LightGBM**	**Train**	**Test**
R^2	0.999	0.986	R^2	0.986	0.981
RMSE	0.421	11.470	RMSE	13.432	14.521
MAE	0.221	8.272	MAE	9.598	10.403

in the training set (0.999) and a high R^2 value of 0.986 in the test set. While the RMSE value was measured as 0.421 in the training set, it was measured as 11.470 in the test set, which shows that the errors in the predictions increased in the test set. The MAE value was determined as 0.221 in the training set and 8.272 in the test set. For the LightGBM model, the learning rate (learning_rate) was set as 0.028, the number of leaves (num_leaves) as 10 and the minimum number of data per leaf (min_data_in_leaf) as 10. The R^2 value of this model was 0.986 in the training set and 0.981 in the test set, and these results show that the model has a high level of explanatory power in both sets. The RMSE value was determined as 13.432 in the training set and 14.521 in the test set; and the MAE was calculated as 9.598 and 10.403, respectively. When the results of both models are examined, it is seen that XGBoost performs better than LightGBM in the training set, but both models achieve similar R^2 values in the test set. In terms of RMSE and MAE values, LightGBM exhibited higher errors, especially in the test set. These results provide important information on how the hyperparameter settings and model structures of the models should be optimised according to the characteristics of the data sets.

Figures 3–6 present the goodness of fit criteria of the optimisation results performed on the training and test sets for the XGBoost and LightGBM algorithms with three-dimensional graphics. These figures visualise how the hyperparameter settings of the algorithms affect the performance of the model.

Figures 3 and 4 show the R^2, RMSE and MAE values obtained on the training and test sets for the XGBoost algorithm. In the training set (Figure 3), XGBoost exhibits a performance with an almost perfect R^2 value and low RMSE and MAE values. This shows that the model fits the training data very well and makes high-accuracy predictions. In the test set (Figure 4), the R^2 value has slightly decreased, while the RMSE and MAE values have increased. This indicates that the generalisation ability of the model is slightly lower than the training set but still shows an acceptable performance. Figures 5 and 6 contain similar goodness of fit criteria for the LightGBM algorithm. The R^2 value obtained for the training set (Figure 5) is quite high, but the RMSE and MAE values are slightly higher than XGBoost. This shows that LightGBM is not as effective as XGBoost in fitting the training set.

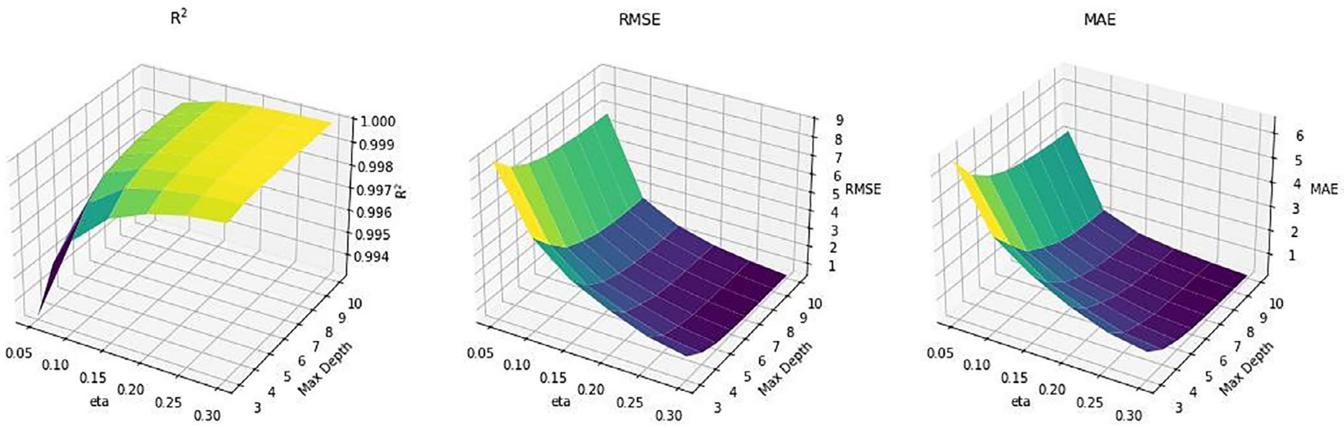

FIGURE 3 | The goodness of fit criteria according to optimized results for XGBoost (train).

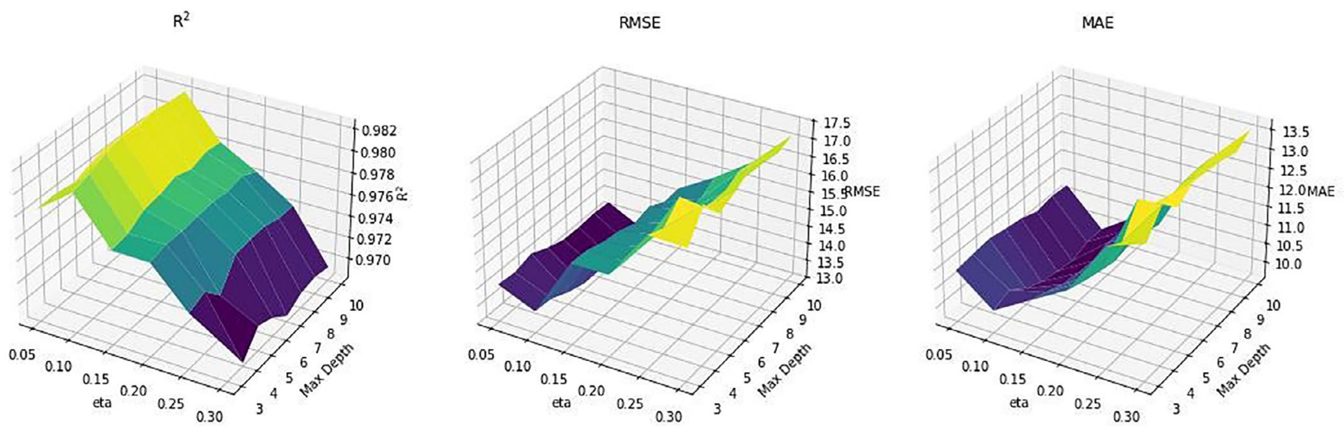

FIGURE 4 | The goodness of fit criteria according to optimized results for XGBoost (test).

The R^2 value for the test set (Figure 6) remains relatively high, but the RMSE and MAE values increase even more, indicating that the model's prediction errors on the test data increase. This situation reveals that the generalisation performance of LightGBM shows more variance compared to XGBoost. These figures allow us to comparatively evaluate the performance of both algorithms on different data sets. The XGBoost algorithm generally exhibits better goodness-of-fit criteria than LightGBM on the training and test sets.

Figure 7 visually presents the variable importances determined based on the sensitivity analysis for the XGBoost algorithm.

According to Figure 7, the effect of different variables on the model is expressed as a percentage and these importance values are represented by coloured bars. As seen in the visual, the HW variable has the highest importance in the model estimates with 0.51%, which shows that HW is the factor that affects the model outputs the most. When we look at the other variables; HG has an importance of 0.42%, RH has an importance of 0.02% and BL has an importance of 0.01%. Although these variables also play an important role in the model estimates, they are not as decisive as HW. The BDL variable, which has the lowest importance, is the feature that has the least effect on the model estimates with 0.01%. This visual is quite useful in understanding which features the model takes into account more and how these features shape the model estimates. In particular, the fact that the HW variable has a much higher importance than all other features may require a more detailed examination of the effect of this variable on the model.

When the results of both models are examined, it is seen that the XGBoost algorithm performs better than the LightGBM algorithm in the training data, but both models reach similar R^2 values in the test data. In terms of RMSE and MAE values, it was determined that the error rates were higher in the LightGBM model, especially in the test data. These results provide important information on how model hyperparameter settings and model configurations should be optimised according to the characteristics of the data set. However, it should be noted that the models developed in this study were trained only with data from this animal material consisting of Holstein × Zebu crossbreeds, and therefore the results obtained may vary from herd to herd, breed to breed and even from species to species.

4 | Discussion

In the present study, XGBoost and LightGBM algorithms were used to estimate the live weight (BW) of Holstein×Zebu crossbred

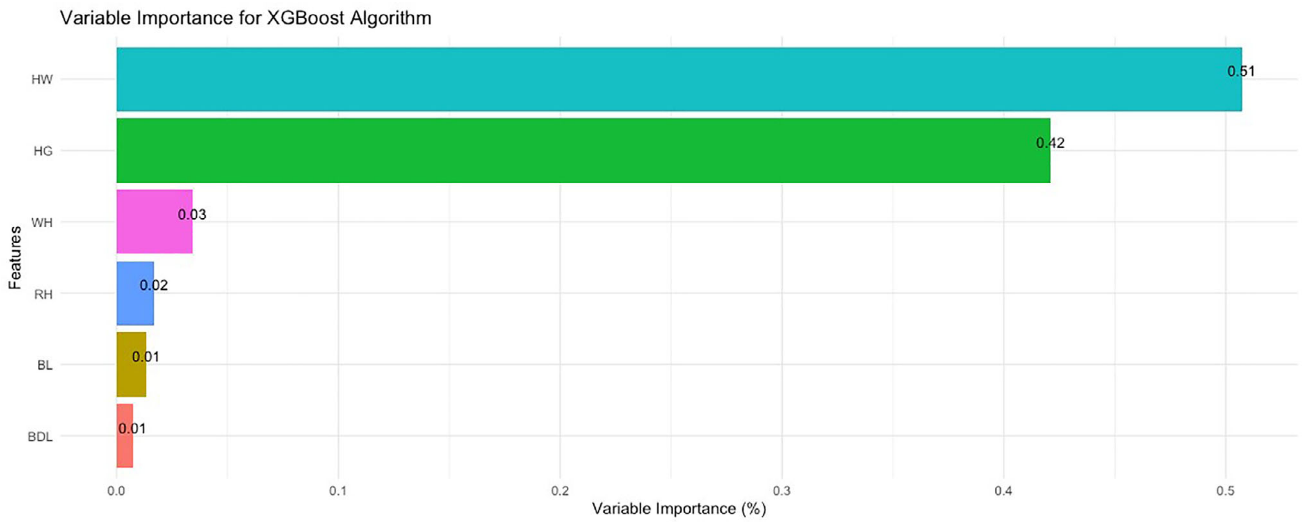

FIGURE 5 | The goodness of fit criteria according to optimized results for LightGBM (train).

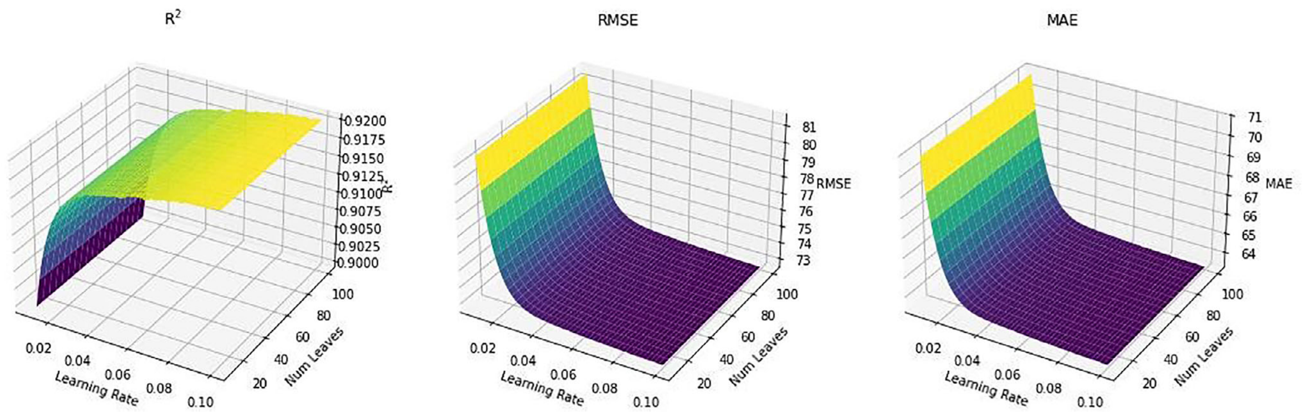

FIGURE 6 | The goodness of fit criteria according to optimized results for LightGBM (test).

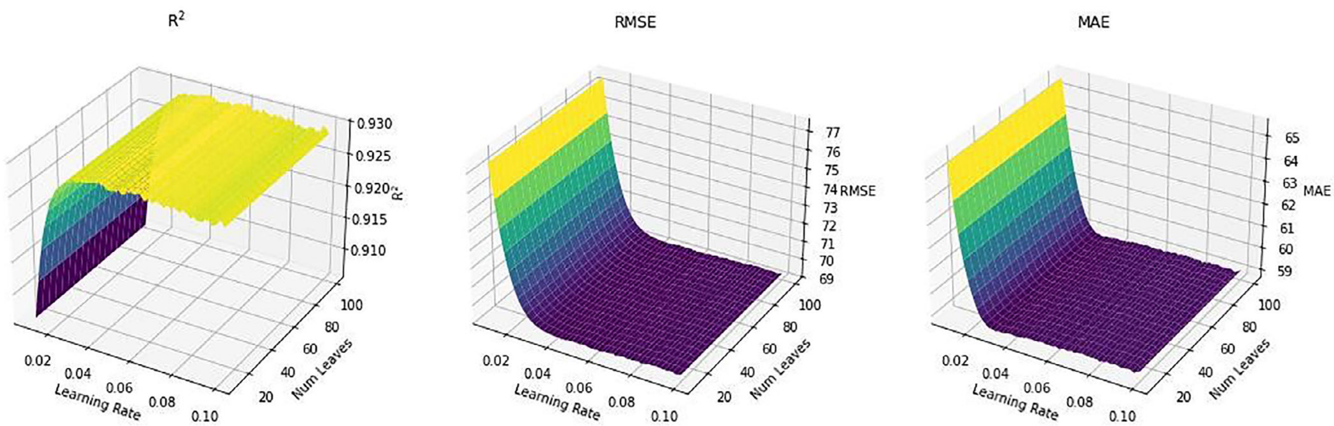

FIGURE 7 | Variable importance according to sensitivity analysis for XGBoost.

heifers. The results show that both algorithms perform quite successful estimations with high levels of R^2 values; XGBoost reached R^2 values of 0.999 for the training set and 0.986 for the test set, while LightGBM reached R^2 values of 0.986 for the training set and 0.981 for the test set. The strength of the study lies in the ability of the algorithms to achieve high accuracy rates in live weight estimations using different parameter settings and various biometric measurements. In particular, the optimum hyperparameters selected for the XGBoost algorithm and the configurations set for LightGBM enabled the models to successfully model the complexity of the data and thus played a critical role in achieving high R^2 values. In addition, another strength of the study is the use of a data set covering a wide range of measurements, which provides the opportunity to test the algorithms' ability to learn data structures of different sizes in a wide spectrum. These provide valuable information on how machine learning models can be optimised for live weight estimation and guide future studies in this area. This study provides scientific contributions that will improve data-based decision-making processes, especially in the livestock sector, enabling effective and sustainable livestock farming practices. Furthermore, it demonstrates the potential of machine learning algorithms for estimating live weights of crossbred heifers but also highlights the need for additional studies to increase the generalisability of the model.

The limited number of studies conducted in this field of study makes it difficult to fully evaluate the potential of advanced machine learning techniques for estimating the live weights of hybrid heifers. Studies focusing on crossbred breeds such as Holstein × Zebu are critical to understanding the effects of genetic diversity and environmental interactions on the growth performance of these animals. However, there are not enough examples in the existing literature regarding developing and testing such comprehensive models. This increases the need to produce scientific data to support scientists and industry professionals in the challenges of optimising livestock practices and developing genetic improvement strategies. Therefore, it is of great importance to conduct more research in this field so that modern tools such as machine learning and artificial intelligence can be used more effectively in the animal husbandry sector. Such studies will pave the way for innovations that will increase efficiency and sustainability in the sector by providing accurate, reliable and fast estimations on critical parameters such as live weight estimation.

Li et al. (2021) tried to compare several machine learning algorithms such as LightGBM, XGBoost, Random Forest, Gradient Boosting Tree, BA-LightGBM and CRBA-LightGBM for predicting the aero-engine exhaust gas temperature from several traits. CRBA-LightGBM algorithm was to determine the most reliable prediction model. However, LightGBM and XGBoost algorithms in this study showed similar results for predicting aero-engine exhaust gas temperature. According to the findings of Li et al. (2021), the LightGBM and XGBoost algorithms exhibited reliable prediction performance. These results support the conclusions of our current study. However, more information about the dataset is needed to compare the authors' research and our current study directly.

Rufo et al. (2021) evaluated various machine learning algorithms such as KNN, SVM, NB, BG, RF, XGBoost and LightGBM to diagnose diabetes mellitus. The research showed that the LightGBM algorithm provided more reliable results regarding AUC values. In addition, it was determined that this algorithm required relatively high computational times on both training and test sets. Therefore, diagnoses made with LightGBM are considered more reliable. Compared to the current study's findings, despite being a classification and prediction task, the successful results obtained by the LightGBM algorithm also don't support the results of this study.

Shehadeh et al. (2021) tested various machine learning algorithms such as Modified Decision Tree (MDT), LightGBM, and XGBoost for residual value estimation of heavy construction equipment. The research results showed that MDT provided higher accuracy compared to others in terms of R^2 value. Although LightGBM and XGBoost also achieved high R^2 values, MDT exhibited the highest performance. If a comparison is made between LightGBM and XGBoost, it will be revealed that LightGBM and XGBoost had great results. In light of these results, when the evaluation criteria of the models other than MDT are similar, it is seen that LightGBM and XGBoost algorithms give similar results in the rankings in the current study. However, since MDT stands out as the most effective model, it is difficult to make a direct comparison without more information about the dataset of the authors' study.

Alabdullah et al. (2022) made a comparison using LightGBM and XGBoost algorithms to measure the rapid chloride penetration resistance of metakaolin. In this study, the performance indices used in the training and test sets to evaluate the performance of both models are the same as the indices used in our study. According to the findings of Alabdullah et al., the LightGBM algorithm exhibited superior prediction performance compared to XGBoost. These results don't support the conclusions of our current study. However, more information about the dataset is needed to directly compare the authors' research and our current study.

Liang et al. (2022) developed a new methodology based on Principal Component Analysis (PCA) and LightGBM, which aims to estimate stellar atmospheric parameters using photometric data. In addition to PCA and linear regression, the researchers evaluated various algorithms such as Random Forest, LightGBM, XGBoost, Gradient Boosting Decision Tree, Artificial Neural Networks and Support Vector Regression. In the study, the LightGBM algorithm integrated with PCA was determined to be the most effective method for this study in terms of both computational time and RMSE performance. These findings don't support the results of our research and do not provide a comprehensive discussion, as more information on the structural details of the dataset is required to determine the most suitable method.

Yang (2022) compared the LightGBM and XGBoost algorithms to determine the prediction performance for the remaining service life of Li-ion batteries. The author optimised the model parameters to improve prediction precision and performed LightGBM and XGBoost algorithms on the lifespan prediction of lithium batteries. As goodness-of-fit criteria, mean absolute error was used to compare the prediction models. According to the results of this study, both LightGBM and XGBoost algorithms had similar and higher prediction performance. While these findings do not provide a comprehensive discussion, they support the results of our research. Still, more information is needed about the structural details of the dataset to determine the most appropriate method.

According to other studies in the literature and the findings obtained in the current study, XGBoost algorithm shows superior performance compared to LightGBM. This may be due to the XGBoost being more resistant to overfitting due to its internal regularisation mechanisms. In particular, the high noise ratio or relatively small sample size in our dataset suggests that XGBoost's regularisation features may be effective in improving the overall performance of the model. In addition, XGBoost is particularly strong in modelling interactions between features. It is likely that the interactions between variables in our dataset were better handled by XGBoost, thus increasing the model's predictive success. These results demonstrate the usability and effectiveness of machine learning models in estimating live weight and provide valuable information on how XGBoost's modelling capacity can be optimised according to the specific characteristics of datasets.

In conclusion, this study investigated the effectiveness of XGBoost and LightGBM algorithms in estimating live weights of Holstein × Zebu crossbred heifers. The results obtained showed that both algorithms can make successful predictions with high R^2 values. While the XGBoost algorithm showed excellent performance, especially on the training set, the LightGBM algorithm obtained consistent results on both training and test sets. However, the increased error rates observed in the performance of both models on the test set reveal the potential for overfitting and the limits of generalisation abilities in the algorithms, and this situation reveals the need for further studies on this subject. These findings emphasise the importance of machine learning techniques in the development of live weight estimation models and indicate the need for additional research and development studies to further improve these models. In addition, the study encourages the integration of advanced analytical tools to support decision-making processes in the livestock industry and optimise animal welfare. Future studies will provide the opportunity to further fine-tune the algorithm parameters, validate studies on different crossbred combinations, and test the robustness of the models under various environmental conditions. Thus, significant contributions will be made to the development of sustainable practices and genetic improvement strategies in the livestock sector.

Author Contributions

Conceptualisation: Cem Tırınk and Lütfi Bayyurt. Methodology: Jose Herrera-Camacho, Cem Tırınk, Rosa Inés Parra-Cortés, Lütfi Bayyurt, Rashit Uskenov, Karlygash Omarova, Aizhan Makhanbetova, Kadyrbai Chekirov and Alfonso Juventino Chay-Canul. Validation: Cem Tırınk and Lütfi Bayyurt. Formal analysis: Cem Tırınk and Lütfi Bayyurt. Investigation: Jose Herrera-Camach, Cem Tırınk, Rosa Inés Parra-Cortés, Lütfi Bayyurt, Rashit Uskenov, Karlygash Omarova, Aizhan Makhanbetova, Kadyrbai Chekirov and Alfonso Juventino Chay-Canul. Resources: Jose Herrera-Camach, Rosa Inés Parra-Cortés and Alfonso Juventino Chay-Canul. Data curation: Jose Herrera-Camach, Cem Tırınk, Rosa Inés Parra-Cortés, Lütfi Bayyurt, Rashit Uskenov, Karlygash Omarova, Aizhan Makhanbetova, Kadyrbai Chekirov and Alfonso Juventino Chay-Canul. Writing–original draft preparation: Cem Tırınk, Lütfi Bayyurt and Alfonso Juventino Chay-Canul. Writing–review and editing: Jose Herrera-Camach, Cem Tırınk, Rosa Inés Parra-Cortés, Lütfi Bayyurt, Rashit Uskenov, Karlygash Omarova, Aizhan Makhanbetova, Kadyrbai Chekirov and Alfonso Juventino Chay-Canul. Visualisation: Cem Tırınk and Lütfi Bayyurt. Supervision: Cem Tırınk, Lütfi Bayyurt and Alfonso Juventino Chay-Canul. Project administration: Jose Herrera-Camach, Rosa Inés Parra-Cortés and Alfonso Juventino Chay-Canul. All authors have read and agreed to the published version of the manuscript.

Ethics Statement

There is no need to take ethical approval because there is no clinical application on the animals.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Alabdullah, A. A., M. Iqbal, M. Zahid, K. Khan, M. N. Amin, and F. E. Jalal. 2022. "Prediction of Rapid Chloride Penetration Resistance of Metakaolin Based High Strength Concrete Using Light GBM and XGBoost Models by Incorporating SHAP Analysis." *Construction and Building Materials* 345: 128296.

Ali, M., E. Eyduran, M. M. Tariq, et al. 2015. "Comparison of Artificial Neural Network and Decision Tree Algorithms Used for Predicting Live Weight at Post Weaning Period From some Biometrical Characteristics in Harnai Sheep." *Pakistan Journal of Zoology* 47, no. 6: 1579–1585.

Alsabti, K., S. Ranka, and V. Singh. 1998. "CLOUDS: A Decision Tree Classifier for Large Datasets." In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*2–8. AAAI Press.

Bretschneider, G., A. Cuatrin, D. Arias, and D. Vottero. 2014. "Estimation of Body Weight by an Indirect Measurement Method in Developing Replacement Holstein Heifers Raised on Pasture." *Archivos De Medicina Veterinaria* 46, no. 3: 439–443.

Cai, J., X. Li, Z. Tan, and S. Peng. 2021. "An Assembly-Level Neutronic Calculation Method Based on LightGBM Algorithm." *Annals of Nuclear Energy* 150: 107871.

Chen, T., and C. Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. ACM.

Domínguez-Viveros, J. 2023. "Estimation of Growth Parameters in Brahman Cattle From Mexico." *Journal of Agriculture and Rural Development in the Tropics and Subtropics (JARTS)* 124, no. 1: 77–82.

Eyduran, E., D. Zaborski, A. Waheed, S. Celik, K. Karadas, and W. Grzesiak. 2017. "Comparison of the Predictive Capabilities of Several Data Mining Algorithms and Multiple Linear Regression in the Prediction of Body Weight by Means of Body Measurements in the Indigenous Beetal Goat of Pakistan." *Pakistan Journal of Zoology* 49, no. 1: 257–265.

Eyduran, E., M. Akin, and S. P. Eyduran. 2019. *Application of Multivariate Adaptive Regression Splines Through R Software*, Nobel Academic Publishing, Ankara.

Faraz, A., C. Tirink, E. Eyduran, et al. 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: 1–12.

Felius, M., M. L. Beerling, D. S. Buchanan, B. Theunissen, P. A. Koolmees, and J. A. Lenstra. 2014. "On the History of Cattle Genetic Resources." *Diversity* 6, no. 4: 705–750.

Gertz, M., K. Grobe-Butenuth, W. Junge, et al. 2020. "Using the XGBoost Algorithm to Classify Neck and Leg Activity Sensor Data Using On-Farm Health Recordings for Locomotor-Associated Diseases." *Computers and Electronics in Agriculture* 173: 105404.

Godfray, H. C. J., P. Aveyard, T. Garnett, et al. 2018. "Meat Consumption, Health, and the Environment." *Science* 361, no. 6399: eaam5324.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9, no. 3: 90–95.

Jónsdóttir, S., and G. Gísladóttir. 2023. "Land Use Planning, Sustainable Food Production and Rural Development: A Literature Analysis." *Geography and Sustainability* 4, no. 4: 391–403.

Ke, G., Q. Meng, T. Finley, et al. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30: 3146–3154.

Khan, M. A., M. M. Tariq, E. Eyduran, et al. 2014. "Estimating Body Weight From Several Body Measurements in Harnai Sheep Without Multicollinearity Problem." *The Journal of Animal & Plant Sciences* 24, no. 1: 120–126.

Kurnaz, B., H. Önder, D. Piwczyński, M. Kolenda, and B. Sitkowska. 2021. "Determination of the Best Model to Predict Milk Dry Matter in High Milk Yielding Dairy Cattle." *Acta Scientiarum Polonorum Zootechnica* 20, no. 3: 41–44.

Li, C. 2017. "The Role of Beef in Human Nutrition and Health." In *Ensuring Safety and Quality in the Production of Beef*, *Volume* 2, edited by M. Dikeman, 329–336. Burleigh Dodds Science Publishing.

Li, D., J. Peng, and D. He. 2021. "Aero-Engine Exhaust Gas Temperature Prediction Based on LightGBM Optimized by Improved Bat Algorithm." *Thermal Science* 25, no. 2A: 845–858.

Li, P., Q. Wu, and C. Burges. 2007. "Mcrank: Learning to Rank Using Multiple Classification and Gradient Boosting." *Advances in Neural Information Processing Systems* 20: 897–904.

Li, S., X. Dong, D. Ma, B. Dang, H. Zang, and Y. Gong. 2024. "Utilizing the Lightgbm Algorithm for Operator User Credit Assessment Research." Preprint, arXiv, March 21. https://doi.org/10.54254/2755-2721/75/20240503.

Liang, J., Y. Bu, K. Tan, et al. 2022. "Estimation of Stellar Atmospheric Parameters With Light Gradient Boosting Machine Algorithm and Principal Component Analysis." *The Astronomical Journal* 163, no. 4: 153.

Mendes, M., and E. Akkartal. 2009. "Regression Tree Analysis for Predicting Slaughter Weight in Broilers." *Italian Journal of Animal Science* 8: 615–624.

Oliveira, A. S., D. C. Abreu, M. A. Fonseca, and P. M. B. Antoniassi. 2013. "Development and Evaluation of Predictive Models of Body Weight for Crossbred Holstein-Zebu Dairy Heifers." *Journal of Dairy Science* 96, no. 10: 6697–6702.

Pérez-Escamilla, R. 2017. "Food Security and the 2015–2030 Sustainable Development Goals: From Human to Planetary Health." *Current Developments in Nutrition* 1, no. 7: e000513.

Parra-Bracamonte, G. M., J. C. Martínez-González, A. M. Sifuentes-Rincón, V. R. Moreno-Medina, and E. Ortega-Rivas. 2015. "Meat Tenderness Genetic Polymorphisms Occurrence and Distribution in Five Zebu Breeds in Mexico." *Electronic Journal of Biotechnology* 18, no. 5: 365–367.

Peralta-Torres, J. A., Y. Izquierdo-Camacho, V. H. Severino-Lendechy, and J. C. Segura-Correa. 2021. "Milk Production of Holstein X Gir Cows in a Dual Purpose System in the Tropics." *Journal MVZ Cordoba* 27, no. 1: e2359.

Punuri, S. B., S. K. Kuanar, M. Kolhar, et al. 2023. "Efficient Net-XGBoost: An Implementation for Facial Emotion Recognition Using Transfer Learning." *Mathematics* 11, no. 3: 776.

Raybaut, P. 2009. "Spyder-Documentation." Pythonhosted. Org.

Revelle, W. R. 2025. psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.5.3. (2025-03-21).

Román-Ponce, S. I., F. D. J. Ruiz-López, H. H. Montaldo, R. Rizzi, and H. Román-Ponce. 2013. "Efectos de Cruzamiento para Producción de Leche y Características de Crecimiento en Bovinos de Doble Propósito en el Trópico húmedo." *Revista Mexicana De Ciencias Pecuarias* 4: 405–416.

Rufo, D. D., T. G. Debelee, A. Ibenthal, and W. G. Negera. 2021. "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)." *Diagnostics* 11, no. 9: 1714.

Sagi, O., and L. Rokach. 2021. "Approximating XGBoost With an Interpretable Decision Tree." *Information Sciences* 572: 522–542.

Sahin, M., E. Yavuz, and F. Uckardes. 2018. "Multicollinearity Problem and Bias Estimates in Japanese Quail." *Pakistan Journal of Zoology* 50, no. 2: 757–761.

Shehadeh, A., O. Alshboul, R. E. Al Mamlook, and O. Hamedat. 2021. "Machine Learning Models for Predicting the Residual Value of Heavy Construction Equipment: An Evaluation of Modified Decision Tree, LightGBM, and XGBoost Regression." *Automation in Construction* 129: 103827.

Shi, Y., G. Ke, D. Soukhavong, et al. 2025. Lightgbm: Light Gradient Boosting Machine. R package version 4.6.0. (2025-02-13).

Smith, J., K. Sones, D. Grace, S. MacMillan, S. Tarawali, and M. Herrero. 2013. "Beyond Milk, Meat, and Eggs: Role of Livestock in Food and Nutrition Security." *Animal Frontiers* 3, no. 1: 6–13.

Team R C. 2020. *RA Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Tırınk, C. 2020. "Estimating of Birth Weight Using Placental Characteristics in the Presence of Multicollinearity." *Black Sea Journal of Engineering and Science* 3, no. 4: 138–145.

Tırınk, C. 2022. "Comparison of Bayesian Regularized Neural Network, Random Forest Regression, Support Vector Regression and Multivariate Adaptive Regression Splines Algorithms to Predict Body Weight From Biometrical Measurements in Thalli Sheep." *Kafkas Üniversitesi Veteriner Fakültesi Dergisi* 28, no. 3: 411–419.

Tyasi, T. L., E. Eyduran, and S. Celik. 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: 1–8.

Vázquez-Martínez, I., C. Tirink, F. Casanova-Lugo, et al. 2024. "Predicting the Body Weight of Crossbred Holstein × Zebu Dairy Cows Using Multivariate Adaptive Regression Splines Algorithm." *Journal of Dairy Research* 91: 267–272.

Warner, R. D., P. L. Greenwood, D. W. Pethick, and D. M. Ferguson. 2010. "Genetic and Environmental Effects on Meat Quality." *Meat Science* 86, no. 1: 171–183.

Wei, T., V. Simko, M. Levy, Y. Xie, Y. Jin, and J. Zemla. 2017. "Package 'Corrplot'." *Statistician* 56, no. 316: e24.

Yang, S. 2022. "Prediction Method of Remaining Service Life of Li-Ion Batteries Based on XGBoost and LightGBM." In *2022 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence (AHPCAI)*, 324–327. IEEE.

Yavuz, E., and M. Şahin. 2022. "Semiparametric Regression Models and Applicability in Agriculture." *Black Sea Journal of Agriculture* 5, no. 2: 160–166.