

## Article

# Prediction of Body Weight by Using PCA-Supported Gradient Boosting and Random Forest Algorithms in Water Buffaloes (*Bubalus bubalis*) Reared in South-Eastern Mexico

Armando Gomez-Vazquez <sup>1</sup>, Cem Tırınk <sup>2</sup> , Alvar Alonzo Cruz-Tamayo <sup>3</sup> , Aldenamar Cruz-Hernandez <sup>1</sup>, Enrique Camacho-Pérez <sup>4</sup> , İbrahim Cihangir Okuyucu <sup>5</sup>, Hasan Alp Şahin <sup>6</sup> , Dany Alejandro Dzib-Cauch <sup>7</sup>, Ömer Gülboy <sup>5</sup>, Ricardo Alfonso Garcia-Herrera <sup>1,\*</sup> and Alfonso J. Chay-Canul <sup>1</sup>

<sup>1</sup> División Académica de Ciencias Agropecuarias, Universidad Juárez Autónoma de Tabasco, Villahermosa C.P. 86280, Tabasco, Mexico; armando.gomez@ujat.mx (A.G.-V.); aldenamar.cruz@ujat.mx (A.C.-H.); alfonso.chay@ujat.mx (A.J.C.-C.)

<sup>2</sup> Department of Animal Science, Faculty of Agriculture, Iğdir University, TR76000 Iğdir, Turkey; cem.tirink@igdir.edu.tr

<sup>3</sup> Facultad de Ciencias Agropecuarias, Universidad Autónoma de Campeche, Escárcega C.P. 24350, Campeche, Mexico; alalcruz@uacam.mx

<sup>4</sup> Facultad de Ingeniería, Universidad Autónoma de Yucatán, Av. Industrias No Contaminantes s/n, Mérida C.P. 97302, Yucatán, Mexico; enrique.camacho@correo.uady.mx

<sup>5</sup> Department of Animal Science, Faculty of Agriculture, Ondokuz Mayıs University, TR55139 Samsun, Turkey; cihangir.okuyucu@omu.edu.tr (İ.C.O.); omergulboy@gmail.com (Ö.G.)

<sup>6</sup> Research Institute of Hemp, Ondokuz Mayıs University, TR55139 Samsun, Turkey; h.alpsahin@gmail.com

<sup>7</sup> Tecnológico Nacional de México, Instituto Tecnológico Superior de Calkiní, Av. Ah-Canul, Calkiní C.P. 24900, Campeche, Mexico; dadzib@itescam.edu.mx

\* Correspondence: ricardo.garcia@ujat.mx



**Citation:** Gomez-Vazquez, A.; Tırınk, C.; Cruz-Tamayo, A.A.; Cruz-Hernandez, A.; Camacho-Pérez, E.; Okuyucu, İ.C.; Şahin, H.A.; Dzib-Cauch, D.A.; Gülboy, Ö.; Garcia-Herrera, R.A.; et al. Prediction of Body Weight by Using PCA-Supported Gradient Boosting and Random Forest Algorithms in Water Buffaloes (*Bubalus bubalis*) Reared in South-Eastern Mexico. *Animals* **2024**, *14*, 293. <https://doi.org/10.3390/ani14020293>

Academic Editor: Markku Saastamoinen

Received: 4 December 2023

Revised: 4 January 2024

Accepted: 12 January 2024

Published: 17 January 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Simple Summary:** Accurately estimating body weight is crucial for managing water buffalo health and optimizing feeding strategies. This study explored the effectiveness of machine learning models in predicting body weight based on body measurements. Principal component analysis was employed to reduce the dimensionality of the data and identify the most relevant features. Subsequently, Gradient Boosting and Random Forest algorithms were utilized to predict body weight using the reduced data set. The Gradient Boosting algorithm demonstrated superior performance compared to the Random Forest algorithm. These findings suggest that the combination of principal component analysis and Gradient Boosting offers a reliable and effective method for estimating body weight in water buffaloes. This approach holds promise for improving animal production and health management practices. Future research could focus on enhancing the applicability and generalizability of these models to diverse water buffalo populations across various geographical regions.

**Abstract:** This study aims to use advanced machine learning techniques supported by Principal Component Analysis (PCA) to estimate body weight (BW) in buffaloes raised in southeastern Mexico and compare their performance. The first stage of the current study consists of body measurements and the process of determining the most informative variables using PCA, a dimension reduction method. This process reduces the data size by eliminating the complex structure of the model and provides a faster and more effective learning process. As a second stage, two separate prediction models were developed with Gradient Boosting and Random Forest algorithms, using the principal components obtained from the data set reduced by PCA. The performances of both models were compared using  $R^2$ , RMSE and MAE metrics, and showed that the Gradient Boosting model achieved a better prediction performance with a higher  $R^2$  value and lower error rates than the Random Forest model. In conclusion, PCA-supported modeling applications can provide more reliable results, and the Gradient Boosting algorithm is superior to Random Forest in this context. The current study demonstrates the potential use of machine learning approaches in estimating body weight in water buffaloes, and will support sustainable animal husbandry by contributing to decision making processes in the field of animal science.

**Keywords:** principal component analysis; gradient boosting; random forest; buffalo; body weight

## 1. Introduction

In recent years, buffalo (*Bubalus bubalis*) breeding has gained a place as an important breeding activity in the livestock sector in Mexico, as it is a source of milk, dairy products, and meat [1]. Buffaloes offer many important advantages compared to cattle, such as having better adaptation abilities and greater resistance to tropical animal diseases, as well as better utilization of low-quality feed in terms of nutritional quality [2]. In Mexico, buffaloes live in states such as Veracruz, Tabasco, Chiapas and Campeche, which have a hot and humid climate with large swamps [3]. Although producers perceive buffalo farming as profitable, much research is necessary regarding animal production parameters [4]. Growth rate is a characteristic of livestock production's adaptability and economic suitability [5], making it an essential parameter in animal production.

For this reason, body weight (BW) appears as the most critical information in production systems, as it will vary depending on many financial characteristics [6,7]. Accurate BW prediction is a basis in animal science studies, such as animal healthcare management, animal husbandry, and determining drug doses and feeding optimization [8]. BW estimation poses a complex challenge in identifying and modeling many processes in animal breeding due to many factors that include computationally demanding situations, from determining herd management strategies to genetic selection. In this context, it is evident that more research is needed to estimate BW accurately and reliably.

Advances that will further the ability to benefit from these complex data sets have occurred in machine learning and many statistical approaches [9]. Principal Component Analysis (PCA) helps to separate high-dimensional data into their components in the most informative way [10]. In this form, PCA is emerging as a leading technique to simplify analytical processes that can be applied later to complex and high-dimensional data sets. PCA alleviates the high-dimension problem and increases the interpretability of the model without sacrificing critical information [11].

However, transforming the explanatory variables for BW prediction through PCA is only a precursor to the predictive modeling journey. The trick is that providing valid and reliable predictions depends on choosing robust algorithms to exploit the reduced feature space [12]. In this context, algorithms such as Gradient Boosting and Random Forest are powerful prediction methods known for their high prediction abilities.

Combining PCA with Gradient Boosting and Random Forest algorithms is a sequential application of these methods and a strategic approach to improving the performance of Gradient Boosting and Random Forest algorithms, which are predictive algorithms [13]. This combination aims to leverage the strengths of PCA, such as feature extraction and noise reduction capabilities, Gradient Boosting's ability to optimize loss functions, and Random Forest's ensemble strategy that increases accuracy and controls overfitting.

The current study aims to provide empirical evidence on the collective impact of these methods on estimating BW. With our approach, BW underlines the importance of methodical feature engineering followed by the application of complex algorithms, paving the way for a robust prediction framework that has the potential to revolutionize prediction applications.

## 2. Materials and Methods

The buffalo were cared for according to the ethical guidelines and animal experimentation regulations of the Department of Agricultural Sciences of the Universidad Juárez Autónoma de Tabasco (approval code: UJAT-2012-IA-18) on a commercial farm located in Isla, Veracruz State, Mexico. The climatic conditions of the region are hot and humid, with summer rains, and the average annual temperature and precipitation are 25 °C and 2750 mm, respectively.

The experiment was carried out at the commercial farm “Polcay” in the municipality of Sabancuy (18°99' N 91°14' W), located northeast of the municipality of Carmen in the southwest of the state of Campeche, Mexico. The climatic condition of the region is warm and sub-humid, with summer rains, and an average annual temperature of 26.7 °C and rainfall of 1412 mm. The animals grazed on native grasses such as *Cenchrus echinatus* (Mul), *Dactyloctenium aegyptium* (chimes su'uk), *Sporobolus virginicus* (ch'ilibil su'uk), and *Spartina spartinae* (k'oxolaak), and grasses such as *Brachiaria brizantha* and *Panicum maximum* ex *Poaceae*, plus water ad libitum.

BW and body measurements were taken in 130 Murrah buffaloes aged 6 to 10 months (78 females and 52 males). The body measurements recorded were: (1) hearth girth (HG), (2) thorax width (TW), (3) hip width (HW), (4) body length (BL) and (5) diagonal body length (BDL), (6) withers height (WH), (7) rump height (RH) and (8) rib depth (RD), respectively. BW was recorded by weighing the animals on a fixed platform scale with a capacity of 2000 kg and an accuracy of 0.5 kg (Revuelta, Torreon, Coahuila, Mexico), while body measurements were recorded using a flexible fibreglass tape measure (Truper®, Truper, S. A. de C. V., San Lorenzo, Mexico) and a 65 cm forcipule, as previously described by [6].

### Statistical Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique proposed by Karl Pearson in 1901, and is used in almost all fields of science [10,14]. Originating in the early 20th century, PCA has been proposed as a basic linear method for reducing dimensions in a variety of applications, such as compressing existing data sets [12,15]. The main purpose of PCA is as a statistical tool that expresses the variability occurring in the data set by creating a new, compact subset of variables known as principal components [16]. The technique reduces the size of data in a high-dimensional structure by projecting the initial data onto a new axis defined by these principal components [12]. PCA is also performed by constructing a linear subspace of reduced dimensions that captures the critical variations present in the data set. In other words, it enables the determination of orthogonal directions that effectively explain the variance of the data. In addition, building sub-dimensions allows data to be reflected in these orthogonal directions [10,17]. Furthermore, the process of PCA involves determining a linear transformation that maximizes the data variance by calculating the eigenvalues and eigenvectors of the data's covariance matrix [10,12]. Here, eigenvectors define the essential directions that maximize the variance, while eigenvalues show the variance explained by each principal component [12,18]. In this way, the principal components with the highest eigenvalues are prioritized, effectively achieving dimensional reduction [19]. The reliability of PCA is limited to linear features, as it often struggles with data showing non-linear features.

After dimensionality reduction through PCA, a new perspective is gained in estimating BW using Gradient Boosting and Random Forest algorithms to take advantage of the dimensionally reduced and important feature set. The logic in choosing these algorithms is twofold: First, the Gradient Boosting algorithm is known for its predictive accuracy, especially in data sets where the relationship between explanatory variables and the outcome is complex and non-linear. Secondly, Random Forest emerges as a highly effective algorithm for feature selection after PCA by leveraging the power of multiple decision trees to improve prediction accuracy and control overfitting. Both methods are well suited to handling reduced-dimensional datasets generated by PCA. This makes them ideal for building a predictive model that is both effective and performs well.

Ensemble learning completes the process by combining the predictive power of various models, such as Random Forest, Boosting and Bagging, to increase the overall accuracy of the prediction to the response variable. The Random Forest (RF) algorithm, which is one of the ensemble learning algorithms and aims to create many decision trees, prevents the overfitting problem by eliminating the high correlation between trees, and provides a balanced model [20]. The Random Forest algorithm is an algorithm that adds a layer of

randomness to the Bagging algorithm [21]. The Random Forest algorithm consists of three processes [22]. The first process of the algorithm is to determine the individual trees. The second process develops a regression tree for each sample with un-pruned aspects. The last process is to predict the latest data from the constructed tree [8].

Boosting algorithms are algorithms that iteratively combine learners that are slightly better than random learners into stronger learners [23]. One of the Boosting algorithms, the Gradient Boosting algorithm, works based on decision trees, similar to the Random Forest algorithm. In addition, Gradient Boosting can also be considered an ensemble method [24,25]. Furthermore, it differentiates itself from other algorithms with its unique community-building approach. This algorithm combines different explanatory variables sequentially with a partial shrinkage on them, and thus can be used in variable selection [25,26]. The strategy of the Gradient Boosting algorithm, unlike the Random Forest algorithm, consists of a process that involves sequentially adding trees to the ensemble, each of which is adjusted according to the cumulative error of the ensemble's predictions. The Gradient Boosting algorithm can be shown as below:

$$y = \mu + \sum_{n=1}^N v h_n(y; X) + e, \quad (1)$$

where  $y$  is defined as the actual response variable vector,  $\mu$  is the mean for the sample of the study,  $v$  is defined as the shrinkage parameter,  $h_n$  is defined as the predictor model, and  $e$  emphasizes the vector of error term for the obtained model. The building of Gradient Boosting requires the cautious tuning of hyper-parameters.

The obtained models of the current study were compared using the goodness of fit criteria, as given below [27]:

1. Coefficient of determination ( $R^2$ ):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (2)$$

2. Root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

3. Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

All statistical evaluations were made using R and Python software [28,29]. Descriptive statistics were used to provide the necessary information about the data. Descriptive statistics for explanatory and response variables were performed using the "psych" package available in R software [30]. Pearson correlation analysis was used with the "corrplot" package in R software to visualize the relationship between explanatory and response variables [31]. Principal component analysis was carried out using the "stats" package in R software [28]. To visualize the scree plot from the PCA, the "factoextra" package was used [32]. For the partitioning of the data set into train and test sets, the "caret" package was used [33]. "gbm" and "randomForest" packages were used to apply the Gradient Boosting and Random Forest algorithms used to estimate BW from the loadings obtained as a result of PCA analysis [22,34]. Python software was used to visualize the 3D plots.

### 3. Results

Table 1 presents descriptive statistics of different physical traits separated by the buffaloes' sex. While the number of observations (n) for female buffaloes is 78, this number is 52 for males. According to Table 1, the average live weight (BW) of female buffaloes is  $223.14 \pm 20.10$  kg, while this value for males is  $230.48 \pm 24.23$  kg, indicating that males are slightly heavier. The average height (HG) in both sexes is close— $149.33 \pm 6.18$  cm in females and  $148.31 \pm 6.67$  cm in males. Other measures such as TW, HW, BL, BDL, WH, RH and RD also show similar variances for both sexes, but overall indicate slightly higher means and a wider range of distribution in males. These findings highlight differences and variations in physical characteristics between sexes, which should be considered when developing body weight prediction models. These measurements can be considered important parameters for understanding and managing biodiversity among buffalo populations.

**Table 1.** Descriptive statistics of the response and explanatory variables.

Sex	Variables	n	Mean $\pm$ Std. Deviation	Min	Max
Female	BW (kg)	78	$223.14 \pm 20.10$	184	294
	HG (cm)		$149.33 \pm 6.18$	138	168
	TW (cm)		$30.18 \pm 3.81$	24	53
	HW (cm)		$38.76 \pm 2.51$	31	44
	BL (cm)		$67.27 \pm 6.75$	54	92
	BDL (cm)		$88.15 \pm 5.17$	66	100
	WH (cm)		$107.91 \pm 5.55$	95	118
	RH (cm)		$110.37 \pm 4.03$	100	122
Male	RD (cm)	52	$58.88 \pm 4.83$	50	70
	BW (kg)		$230.48 \pm 24.23$	176	285
	HG (cm)		$148.31 \pm 6.67$	130	162
	TW (cm)		$29.25 \pm 2.37$	23	37
	HW (cm)		$37.17 \pm 2.51$	26	42
	BL (cm)		$65.33 \pm 6.69$	55	79
	BDL (cm)		$88.94 \pm 4.43$	70	101
	WH (cm)		$108.75 \pm 5.01$	95	118
RH (cm)	$111.6 \pm 5.74$	98	128		
RD (cm)	$56.44 \pm 3.25$	49	69		

In Figure 1, the correlation coefficients between live weight (BW) and various body measurements in buffaloes are expressed in three groups: female, male and the whole population. This graphically illustrates how relationships between these measures may vary across sex and the general population. In this context, there appear to be moderate correlation coefficients between live weight and other measurements for female buffaloes. This indicates that body measurements in females show a relationship with live weight, but are not high enough to conclude that this relationship is strong. This suggests that body measurements of female buffaloes may have more complex relationships with live weight, and that these relationships may be less linear. In this context, correlation coefficients are generally higher for male buffaloes, indicating that body measurements have a stronger and perhaps more linear relationship with body weight in males. This indicates that certain body measurements may be a good indicator of live weight, as well as growth and body composition in males. When the general population was examined, moderate correlation coefficients could be observed when both male and female measurements were averaged. This indicates that differences between sexes keep the correlation values of the general population in balance. General population analysis shows that combining data from both sexes makes the relationships between body measurements and body weight more homogenized.

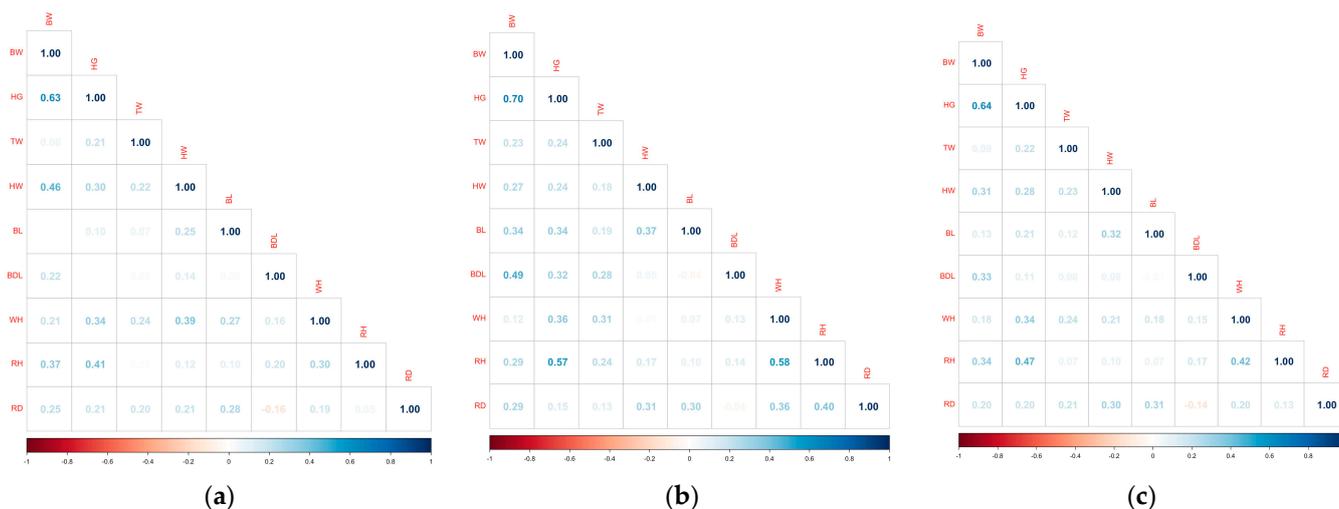


Figure 1. Correlation matrix of the dataset. (a) female; (b) male; (c) all.

The results of the correlation analysis emphasize that sex is an important factor in developing strategies for managing and feeding buffaloes according to sex, and show that individualized approaches may be required. Due to the relatively low correlation coefficients, especially in female buffaloes, it is believed that using Gradient Boosting and Random Forest algorithms, as well as PCA analysis, will provide more reliable results in model estimation.

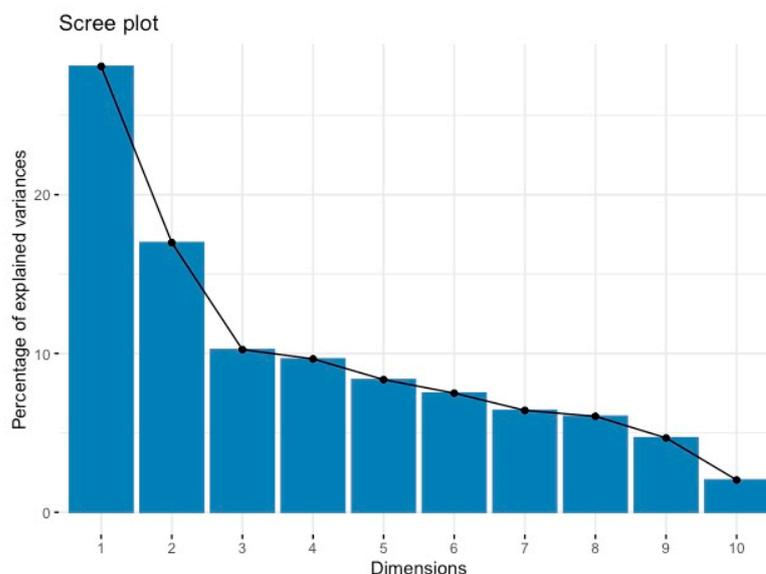
The loadings obtained as a result of the PCA analysis and the information about the variances explained in each principal component are presented in Table 2 and Figure 2.

Table 2. Loadings of principal components for sex and physical features.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Sex	0.077	−0.540	0.267	0.070	−0.314	0.578	−0.095	−0.260	0.165	−0.300
BW	−0.408	−0.281	−0.311	0.337	0.151	0.300	−0.131	−0.050	−0.152	0.623
HG	−0.463	−0.138	−0.014	0.210	0.353	0.044	0.351	0.206	−0.310	−0.578
TW	−0.246	0.189	−0.002	−0.711	0.211	0.510	0.115	0.186	0.189	0.096
HW	−0.349	0.304	−0.321	0.030	−0.116	−0.053	0.176	−0.718	0.304	−0.161
BL	−0.265	0.302	0.057	0.233	−0.713	0.132	0.266	0.418	0.056	0.078
BDL	−0.164	−0.352	−0.592	−0.333	−0.314	−0.248	−0.347	0.223	0.010	−0.239
WH	−0.353	−0.117	0.440	−0.370	−0.254	−0.257	−0.014	−0.296	−0.534	0.164
RH	−0.351	−0.316	0.357	0.029	0.129	−0.395	0.039	0.147	0.661	0.129
RD	−0.291	0.392	0.233	0.167	0.084	0.102	−0.783	0.054	0.020	−0.213
Variance	0.281	0.170	0.102	0.097	0.084	0.075	0.064	0.060	0.047	0.020

Table 2 shows the loadings obtained as a result of PCA analysis and the variance values explained by each principal component. In this context, it provides important findings related to examining the physical characteristics of buffaloes and the effects of gender on basic components. PC1 presents the largest explained variance in the data set. The first four principal components explain more than 65% of the total explained variance, and the first five principal components explain 73%. These ratios show that the first five principal components represent the greatest variation between body weight and other measurements of buffalos. The gender variable has a very large effect on the second principal component (PC2). This shows that gender explains a significant part of the variance explained by this component. This shows that the effect of gender on physical characteristics is important, and that this variable defines a significant part of the variance in the body structure of buffalos. Body weight (BW) has an extremely high positive loading on the tenth principal component (PC10) while presenting negative loadings on

the other principal components. This indicates that body weight has a complex structure of variability among different fundamental components, and that this characteristic is associated with a variety of physical measurements in different dimensions. These results may require the development of gender-specific strategies in the rearing and management of buffalos. In practices aimed at monitoring the health status of animals and in feeding and breeding programs, the relationships of variables such as gender and live weight with other physical measurements should be taken into account. The role of PCA in identifying these components is critical to the development of other models that predict such features and allow for more accurate and effective predictions.



**Figure 2.** Scree plot of variance explained by principal components.

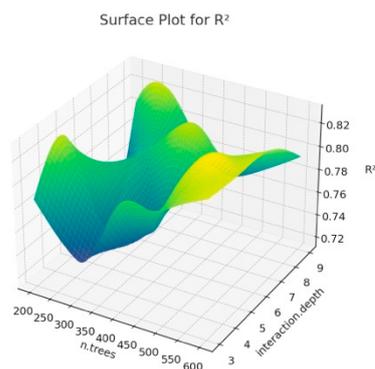
The percentage contribution of each principal component obtained from the PCA analysis to the total variance is also expressed in Figure 2.

According to Figure 2, a typical sloping line shows that the first component has the highest percentage of explained variance, and then the contribution of each additional component decreases. In addition, focusing on the first five components may be sufficient, especially since the first four components explain 65% of the total variance, and the first five components explain 73%. This is consistent with the preservation of the most important information from the data by significantly reducing the data set's size while preserving the defined amount of variance. Additionally, this scree plot and PCA results are critical for managing the complexity and size of the dataset, supporting the application of powerful machine learning algorithms such as Gradient Boosting and Random Forest.

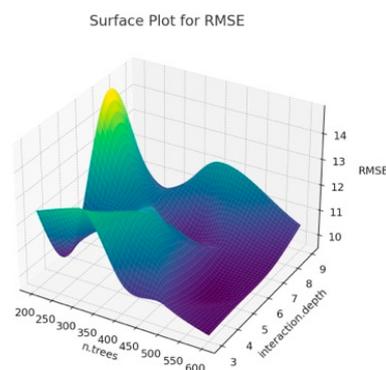
The surface plot obtained when we estimated BW from the first six principal components that explain 81% of the variance as a result of PCA analysis using different hyperparameter values of the Gradient Boosting algorithm is given in Figures 3–5. It is important to interpret the 3D surface plots obtained in Figures 3–5. In this way, the optimum hyperparameters (n.trees and interaction.depth) are determined, and the hyperparameter values that affect the model's performance are seen.

According to Figure 3, it can be observed that the shrinkage and interaction.depth values have a significant impact on  $R^2$ . In addition, the fluctuations in  $R^2$  seen in the graph show that the model better captures the overall structure of the data set. The reason for these fluctuations seen in  $R^2$  is the overfitting problem and the increase in the number of n.trees, which may cause the training time of the model to increase, and thus, the performance to decrease. In addition, the effects of hyperparameters may vary due to the unique structure of the dataset. When Figure 4 is examined, it is observed how RMSE changes at certain n.trees and interaction.depth values. It shows at what point the

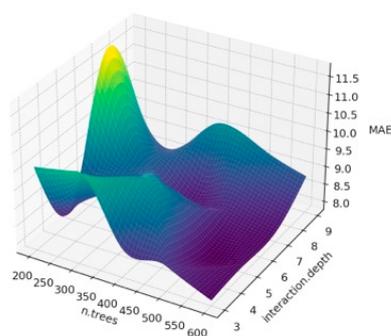
model obtained for different hyperparameter combinations can give the optimal RMSE value. Lower MAE values indicate better model performance. It is possible to see trends similar to RMSE in MAE charts. It is also seen that MAE generally decreases with lower interaction depth values and increasing n.tree values. As a result, the 3D surface plots in Figures 3–5 show the sensitivity of the GBM model to hyperparameters, and how the model obtained by determining these parameter values according to the graph affects the overall performance. In addition, in the current study, lower interaction depth values and increasing n.tree values generally increase the model's accuracy and reliability. Tuning the model's hyperparameters based on these observations also increases the accuracy of BW predictions.



**Figure 3.** Surface plot for Gradient Boosting algorithm results according to n.trees and interaction.depth for  $R^2$ .



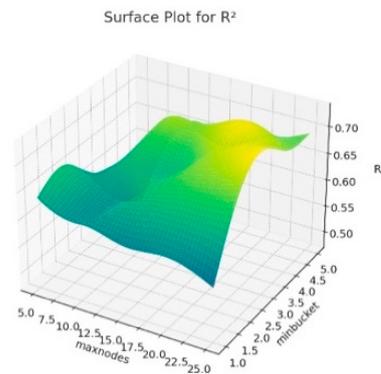
**Figure 4.** Surface plot for Gradient Boosting algorithm results according to n.trees and interaction.depth for RMSE.



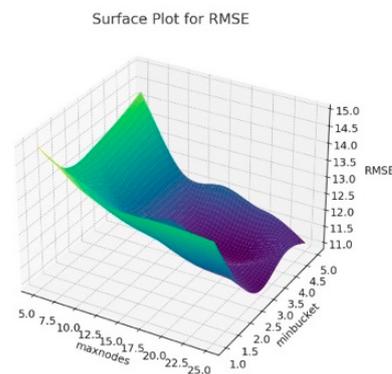
**Figure 5.** Surface plot for Gradient Boosting algorithm results according to n.trees and interaction.depth for MAE.

The 3D surface graphics created for RMSE,  $R^2$  and MAE corresponding to the hyperparameters of the resulting Random Forest model are presented in Figures 6–8. In

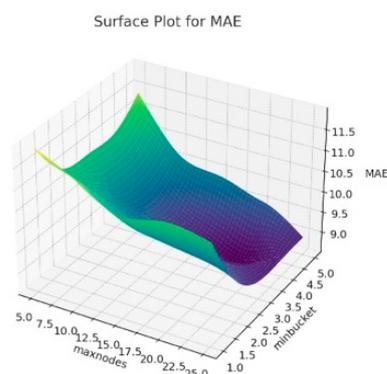
this context, this visualizes how the maxnodes and minbucket hyperparameters affect the model's performance. The term "minbucket" in Figure 4 indicates the minimum number of observations that should be present in the terminal nodes (leaves) in a decision tree. As the "minbucket" value increases, the resulting model becomes less detailed and generalized, which can reduce the risk of overfitting. The term "maxnodes" refers to a decision tree's maximum number of nodes. As the number of nodes in the resulting model increases, the model has a better fit, which may also increase the risk of overfitting.



**Figure 6.** Surface plot for Random Forest algorithm results according to maxnodes and minbucket for R<sup>2</sup>.



**Figure 7.** Surface plot for Random Forest algorithm results according to maxnodes and minbucket for RMSE.



**Figure 8.** Surface plot for Random Forest algorithm results according to maxnodes and minbucket for MAE.

According to Figure 6, it can be seen that R<sup>2</sup> generally increases as the "minbucket" increases, indicating that less detailed models fit the model better. Additionally, it can be seen that R<sup>2</sup> varies again as "maxnodes" increases. In Figure 7, for RMSE values, the

increases in the “minbucket” term can decrease the RMSE values towards the good fit of the model. The same comment can be made for the MAE values in Figure 8. As a result, these graphs show that the “minbucket” and “maxnodes” hyperparameters significantly impact the obtained model performance. Generally, larger “minbucket” values increase the model’s generalization ability, while the effect of “maxnodes” is more complex and based on the dataset. Adjusting the model in line with this information can optimize its performance.

In evaluating the model performances of the Gradient Boosting and Random Forest algorithm,  $R^2$ , RMSE and MAE values are examined. In this context, the model performances are presented in Table 3 with the optimal hyperparameter values for each model. For both models, optimum values of hyperparameters such as the number of trees (n.trees or ntree), tree depth (interaction.depth or maxnodes), shrinkage, and the minimum number of observations in the node (n.minobsinnode or Minbucket) are specified.

**Table 3.** The goodness of fit criteria of the Gradient Boosting and Random Forest algorithms for optimal hyperparameter values.

Hyperparameters of the Models					
Gradient Boosting algorithm			Random Forest algorithm		
n.trees	600		ntree	200	
interaction.depth	3		maxnodes	20	
shrinkage	0.01		node_size	5	
n.minobsinnode	5		minbucket	5	
Goodness of Fit Criteria					
Gradient Boosting algorithm	Train	Test	Random Forest algorithm	Train	Test
$R^2$	0.823	0.818	$R^2$	0.704	0.684
RMSE	4.998	6.418	RMSE	6.870	9.425
MAE	3.971	5.287	MAE	5.306	8.939

According to Table 3, for the Gradient Boosting algorithm,  $R^2$  for the training set is 0.823 and for the test set is 0.818; these high values indicate that the model predicts the data well. It is seen that the RMSE and MAE values are low in the training set and slightly high in the test set. However, this shows that the model fits the training data well, but makes slightly more errors in the test data. However, looking at RMSE and MAE, it can be said that the errors are still at an acceptable level. For the Random Forest algorithm,  $R^2$  values are lower than Gradient Boosting, indicating that the model is less capable of predicting BW. Additionally, the RMSE and MAE values are higher than Gradient Boosting for both the training and testing sets, indicating that the Random Forest model has higher error rates.

#### 4. Discussion

In the current study, principal component analysis (PCA) and two machine learning algorithms, Gradient Boosting and Random Forest, were applied to estimate body weight (BW) in buffaloes. PCA analysis was used to reduce the dimensionality in the dataset and extract the most significant features. This method aims to improve the calculation time and the model’s generalizability by ensuring that our model is trained on fewer, more practical features. Then, using Gradient Boosting and Random Forest algorithms, BW was estimated from the data, the sizes of which were reduced.

The Gradient Boosting algorithm predicted the BW quite well, with the model showing high  $R^2$  values in the training and test sets. In addition, the RMSE and MAE values show that the model’s error is acceptable. These results indicate that the algorithm achieves a strong performance in estimating body weight in buffaloes.

On the other hand, the Random Forest algorithm showed relatively poorer performance than the Gradient Boosting algorithm, with lower  $R^2$  values and higher error rates

(RMSE and MAE). This suggests either that Random Forest is not the optimal model for this dataset, or that the algorithm's hyperparameters should be tuned to provide the best fit.

Various statistical methods have been used to estimate the BW in several species of animal. One of them is PCA, which has been employed to work body conformation features and advance some unobservable components to describe the body conformation of water buffaloes [35]. In addition, PCA has been performed in the morphological description of native goats, describing a significant proportion of the difference in BW [36]. Furthermore, another usage of PCA has been applied to obtain an unbiased explanation of different pre-aging body forms for Uda sheep [37]. These studies indicate that using PCA is quite an effective method of predicting body weight in different livestock species.

Besides PCA, several machine learning algorithms have been used in livestock science. One of these studies emphasized using artificial neural networks in estimating the milk yield in dairy cows, showcasing machine learning algorithms in livestock sciences [38]. In addition, [39] established a prediction model on calving using recurrent neural networks, determining the potential use of machine learning in predicting animal-related measures. Additionally, it points to the use of multi-trait genetic principal components to predict reproductive traits in buffaloes [40].

However, it is important to note that the use of specific algorithms such as PCA-based Gradient Boosting and Random Forest has not been described in the literature. Although the use of machine learning algorithms has been seen in livestock sciences for various purposes, including predicting milk yield and reproductive characteristics, few studies specifically focus on estimating body weight in buffaloes with the use of these algorithms. Although live weight estimation has been achieved in buffaloes using different algorithms, the lack of a PCA-supported algorithm shows a need for more studies, especially those using with multi-dimensional data sets.

In estimating body weight from biometrical features, the Multivariate Adaptive Regression Splines (MARS) algorithm was evaluated within the scope of several goodness of fit criteria [5]. In this context, the aforementioned study was designed to predict body weight for several train and test set proportions. The researchers determined a 70%-30% split between the train and test sets as the most reliable model. Although the methods used were different, they showed similar performance in terms of prediction. Even though Gradient Boosting lagged behind in the train set, the test set gave more reliable results than the aforementioned study.

Ref. [41] proposed a new approach, which is based on Principal Component Analysis (PCA) and light gradient boosting machine (LightGBM) algorithms, for predicting stellar atmospheric parameters from photometric data. To this end, the researchers used several algorithms such as Random Forest, LightGBM, XGBoost, Gradient Boosting decision tree, ANN, support vector regression and linear regression with PCA. In this context, the PCA + LightGBM algorithm was the most reliable method for this study within the scope of the calculation time and RMSE value range. Although it appeared as the best method in this study, it does not provide much information related to the discussion because it does not create a similar data structure.

Ref. [42] used several algorithms, such as the MARS algorithm, Bayesian ridge regression, Ridge regression, support vector machines, Gradient Boosting, Random Forests, XGBoost algorithm, artificial neural networks, classification and regression trees, polynomial regression, K-nearest neighbours and Genetic Algorithms for predicting weight in sheep. According to the results of this study, the five most reliable methods were MARS, Bayesian ridge regression, Ridge regression, support vector machines and Gradient Boosting algorithms. When the results are compared with the current study, we see that the evaluation criteria used are the same. This is an important criterion for comparing studies. Both studies show similar results.

As a result, it has been determined that the Gradient Boosting algorithm provides superior results over Random Forest in terms of prediction performance and minimizing model error. Other articles in the Section 4 also concluded that the PCA-based Gradient

Boosting algorithm is more reliable. However, to increase the generalization capacity of both models and reduce possible overfitting problems, it is recommended to study additional data analysis methods and different hyperparameter tuning techniques in many areas. The accurate estimation of water buffaloes' live weight is critical to animal health and herd management practices. In this context, it is believed that the results of the present study will make a significant contribution to studies carried out in the field. It is also noteworthy that the results of this study only concern the Murrah breed reared in Mexico, and so the model should be tested on other breeds such as Bufalypso, Mediterranean and Swamp.

## 5. Conclusions

This study examined how the body weight of water buffaloes can be estimated using machine learning models based on body measurements. In the study, PCA analysis was used to reduce the size of the features and select the most significant predictors. With this method, the principal components obtained from the data set were used for training Gradient Boosting and Random Forest algorithms.

Our comparative results have shown that the Gradient Boosting algorithm provides better results than the Random Forest algorithm in performance metrics such as  $R^2$ , RMSE and MAE. These results reveal that the Gradient Boosting algorithm is more effective than the Random Forest algorithm in estimating the body weight of water buffaloes.

In conclusion, the use of dimensionality reduction with PCA and the Gradient Boosting algorithm produces effective and reliable results in estimating the body weight of water bison. These findings may provide significant benefits in animal production and health management, particularly in optimizing feeding strategies and developmental monitoring. Future studies may contribute to the development of machine learning-based body weight prediction models by further increasing the applicability and generalizability of these models for water buffalo populations in different geographies.

**Author Contributions:** Conceptualization, A.G.-V., A.A.C.-T., A.C.-H. and A.J.C.-C.; methodology, A.G.-V., A.J.C.-C., A.A.C.-T., Í.C.O. and H.A.Ş.; validation, C.T., D.A.D.-C. and A.J.C.-C.; formal analysis, A.G.-V., C.T., Ö.G. and R.A.G.-H.; investigation, A.G.-V., A.C.-H. and A.A.C.-T.; resources, A.G.-V., R.A.G.-H. and A.J.C.-C.; data collection, A.G.-V.; data analysis, A.A.C.-T., C.T. and D.A.D.-C.; writing—original draft preparation, A.G.-V., C.T., D.A.D.-C. and A.J.C.-C.; writing—review and editing, A.G.-V., C.T., E.C.-P. and R.A.G.-H.; visualization, A.G.-V., C.T., Ö.G. and R.A.G.-H.; supervision, E.C.-P., H.A.Ş. and Ö.G.; project administration A.G.-V. and A.J.C.-C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to. There was no need to seek ethical approval because there was no clinical application applied to the animals.

**Informed Consent Statement:** Written informed consent has been obtained from the patient(s) to publish this paper.

**Data Availability Statement:** To acquire the data please contact the author R.A.G.-H.

**Conflicts of Interest:** The authors declare no conflicts of interest and none of the authors of this paper have a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

## References

1. Mota-Rojas, D.; Bragaglio, A.; Braghieri, A.; Napolitano, F.; Domínguez-Oliva, A.; Mora-Medina, P.; Álvarez-Macías, A.; De Rosa, G.; Pacelli, C.; José, N. Dairy Buffalo Behavior: Calving, Imprinting and Allosuckling. *Animals* **2022**, *12*, 2899. [[CrossRef](#)] [[PubMed](#)]
2. Torres-Chable, O.M.; Ojeda-Robertos, N.F.; Chay-Canul, A.J.; Peralta-Torres, J.A.; Luna-Palomera, C.; Brindis-Vazquez, N.; Blitvich, B.J.; Machain-Williams, C.; Garcia-Rejon, J.E.; Baak-Baak, C.M. Hematologic RIs for Healthy Water Buffaloes (*Bubalus bubalis*) in Southern Mexico. *Vet. Clin. Pathol.* **2017**, *46*, 436–441. [[CrossRef](#)]

3. Peralta-Torres, J.A.; Torres-Chablé, O.M.; Segura-Correa, J.C.; Ojeda-Robertos, N.F.; Chay-Canul, A.J.; Luna-Palomera, C.; Severino-Lendechy, V.H.; Aké-Villanueva, J.R. Ovarian Dynamics of Buffalo (*Bubalus bubalis*) Synchronized with Different Hormonal Protocols. *Trop. Anim. Health Prod.* **2020**, *52*, 3475–3480. [[CrossRef](#)] [[PubMed](#)]
4. Hernández-Herrera, G.; Lara-Rodríguez, D.A.; Vázquez-Luna, D.; Ácar-Martínez, N.; Fernández-Figueroa, J.A.; Velásquez-Silvestre, M.G. Water Buffalo (*Bubalus bubalis*): An Approach to Sustainable Management in Southern Veracruz, Mexico. *Agroproductividad* **2018**, *11*, 27–32.
5. Ağyar, O.; Tırınk, C.; Önder, H.; Şen, U.; Piwczynski, D.; Yavuz, E. Use of Multivariate Adaptive Regression Splines Algorithm to Predict Body Weight from Body Measurements of Anatolian Buffaloes in Türkiye. *Animals* **2022**, *12*, 2923. [[CrossRef](#)] [[PubMed](#)]
6. Ramos-Zapata, R.; Dominguez-Madriral, C.; García-Herrera, R.-A.; Camacho-Perez, E.; Lugo-Quintal, J.M.; Tyasi, T.L.; Gurgel, A.L.C.; Itavo, L.C.V.; Chay-Canul, A.J. Predicting Live Weight Using Body Volume Formula in Lactating Water Buffalo. *J. Dairy Res.* **2023**, *90*, 138–141. [[CrossRef](#)] [[PubMed](#)]
7. Ruiz-Ramos, J.; Torres-Chable, O.M.; Peralta-Torres, J.A.; Ojeda-Robertos, N.F.; Luna-Palomera, C.; Portillo-Salgado, R.; Tyasi, T.L.; Gurgel, A.L.C.; Itavo, L.C.V.; Chay-Canul, A.J.; et al. Estimation of Body Weight Using Body Measurements in Female Water Buffaloes Reared in Southeastern Mexico. *Trop. Anim. Health Prod.* **2023**, *55*, 137. [[CrossRef](#)]
8. Tırınk, C. 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 Univ. Vet. Fak. Derg.* **2022**, *28*, 411–419.
9. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep Learning and Process Understanding for Data-Driven Earth System Science. *Nature* **2019**, *566*, 195–204. [[CrossRef](#)]
10. Hasan, B.M.S.; Abdulazeez, A.M. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *J. Soft Comput. Data Min.* **2021**, *2*, 20–30.
11. Vinutha, M.R.; Chandrika, J.; Krishnan, B.; Kokatnoor, S.A. EPCA—Enhanced Principal Component Analysis for Medical Data Dimensionality Reduction. *SN Comput. Sci.* **2023**, *4*, 243. [[CrossRef](#)]
12. Kocuvan, P.; Hrastič, A.; Kareska, A.; Gams, M. Predicting a Fall Based on Gait Anomaly Detection: A Comparative Study of Wrist-Worn Three-Axis and Mobile Phone-Based Accelerometer Sensors. *Sensors* **2023**, *23*, 8294. [[CrossRef](#)]
13. Abba, S.I.; Pham, Q.B.; Usman, A.G.; Linh, N.T.T.; Aliyu, D.S.; Nguyen, Q.; Bach, Q.-V. Emerging Evolutionary Algorithm Integrated with Kernel Principal Component Analysis for Modeling the Performance of a Water Treatment Plant. *J. Water Process Eng.* **2020**, *33*, 101081. [[CrossRef](#)]
14. Kurita, T. Principal Component Analysis (PCA). In *Computer Vision: A Reference Guide*; Springer: Cham, Switzerland, 2019; pp. 1–4.
15. Sorzano, C.O.S.; Vargas, J.; Montano, A.P. A Survey of Dimensionality Reduction Techniques. *arXiv* **2014**, arXiv:1403.2877.
16. Anowar, F.; Sadaoui, S.; Selim, B. Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (Pca, Kpca, Lda, Mds, Svd, Lle, Isomap, Le, Ica, t-Sne). *Comput. Sci. Rev.* **2021**, *40*, 100378. [[CrossRef](#)]
17. Reddy, G.T.; Reddy, M.P.K.; Lakshmana, K.; Kaluri, R.; Rajput, D.S.; Srivastava, G.; Baker, T. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* **2020**, *8*, 54776–54788. [[CrossRef](#)]
18. Ait-Sahalia, Y.; Xiu, D. Principal Component Analysis of High-Frequency Data. *J. Am. Stat. Assoc.* **2019**, *114*, 287–303. [[CrossRef](#)]
19. Van Der Maaten, L.; Postma, E.O.; van den Herik, H.J. Dimensionality Reduction: A Comparative Review. *J. Mach. Learn. Res.* **2009**, *10*, 13.
20. Lakshmanaprabu, S.K.; Shankar, K.; Ilayaraja, M.; Nasir, A.W.; Vijayakumar, V.; Chilamkurti, N. Random Forest for Big Data Classification in the Internet of Things Using Optimal Features. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2609–2618. [[CrossRef](#)]
21. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
22. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R. News* **2002**, *2*, 18–22.
23. Freund, Y.; Schapire, R.; Abe, N. A Short Introduction to Boosting. *J. Jpn. Soc. Artif. Intell.* **1999**, *14*, 1612.
24. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
25. González-Recio, O.; Jiménez-Montero, J.A.; Alenda, R. The Gradient Boosting Algorithm and Random Boosting for Genome-Assisted Evaluation in Large Data Sets. *J. Dairy Sci.* **2013**, *96*, 614–624. [[CrossRef](#)] [[PubMed](#)]
26. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
27. Zaborski, D.; Ali, M.; Eyduran, E.; Grzesiak, W.; Tariq, M.M.; Abbas, F.; Waheed, A.; Tırınk, C. Prediction of Selected Reproductive Traits of Indigenous Harnai Sheep under the Farm Management System via Various Data Mining Algorithms. *Pak. J. Zool.* **2019**, *51*, 421. [[CrossRef](#)]
28. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
29. VanRossum, G.; Drake, F.L. *The Python Language Reference*; Python Software Foundation: Amsterdam, The Netherlands, 2010; Volume 561.
30. Revelle, W. *Procedures for Personality and Psychological Research*; Northwestern University: Evanston, IL, USA, 2015.
31. Wei, T.; Simko, V.; Levy, M.; Xie, Y.; Jin, Y.; Zemla, J. Package ‘Corrplot’. *Statistician* **2017**, *56*, e24.
32. Kassambara, A.; Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7.999. Available online: <http://www.sthda.com/english/rpkgs/factoextra> (accessed on 20 November 2023).

33. Kuhn, M.; Caret: Classification and Regression Training. R Package Version 6.0-93. 2022. Available online: <https://CRAN.R-project.org/package=caret> (accessed on 20 November 2023).
34. Greenwell, B.M.; Boehmke, B.C.; McCarthy, A.J. A Simple and Effective Model-Based Variable Importance Measure. *arXiv* **2018**, arXiv:1805.04755.
35. Vohra, V.; Niranjana, S.K.; Mishra, A.K.; Jamuna, V.; Chopra, A.; Sharma, N.; Jeong, D.K. Phenotypic Characterization and Multivariate Analysis to Explain Body Conformation in Lesser Known Buffalo (*Bubalus bubalis*) from North India. *Asian-Australas. J. Anim. Sci.* **2015**, *28*, 311. [[CrossRef](#)]
36. Okpeku, M.; Yakubu, A.; Peters, S.; Ozoje, M.; Ikeobi, C.; Adebambo, O.; Imumorin, I. Application of Multivariate Principal Component Analysis to Morphological Characterization of Indigenous Goats in Southern Nigeria. *Acta Agric. Slov.* **2011**, *98*, 101. [[CrossRef](#)]
37. Salako, A.E. Principal Component Factor Analysis of the Morphostructure of Immature Uda Sheep. *Int. J. Morphol.* **2006**, *24*, 571–774. [[CrossRef](#)]
38. Shahinfar, S.; Page, D.; Guenther, J.; Cabrera, V.; Fricke, P.; Weigel, K. Prediction of Insemination Outcomes in Holstein Dairy Cattle Using Alternative Machine Learning Algorithms. *J. Dairy Sci.* **2014**, *97*, 731–742. [[CrossRef](#)]
39. Keceli, A.S.; Catal, C.; Kaya, A.; Tekinerdogan, B. Development of a Recurrent Neural Networks-Based Calving Prediction Model Using Activity and Behavioral Data. *Comput. Electron. Agric.* **2020**, *170*, 105285. [[CrossRef](#)]
40. Agudelo-Gómez, D.A.; Pelicioni Savegnago, R.; Buzanskas, M.E.; Ferraudo, A.S.; Prado Munari, D.; Cerón-Muñoz, M.F. Genetic Principal Components for Reproductive and Productive Traits in Dual-Purpose Buffaloes in Colombia. *J. Anim. Sci.* **2015**, *93*, 3801–3809. [[CrossRef](#)]
41. Liang, J.; Bu, Y.; Tan, K.; Pan, J.; Yi, Z.; Kong, X.; Fan, Z. Estimation of Stellar Atmospheric Parameters with Light Gradient Boosting Machine Algorithm and Principal Component Analysis. *Astron. J.* **2022**, *163*, 153. [[CrossRef](#)]
42. Hamadani, A.; Ganai, N.A. Artificial Intelligence Algorithm Comparison and Ranking for Weight Prediction in Sheep. *Sci. Rep.* **2023**, *13*, 13242. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.