ISSN 1807-1929
Revista Brasileira de Engenharia Agrícola e Ambiental
Brazilian Journal of Agricultural and Environmental Engineering
v.29, n.1, e283168, 2025
Campina Grande, PB – http://www.agriambi.com.br – http://www.scielo.br/rbeaa

DOI: http://dx.doi.org/10.1590/1807-1929/agriambi.v29n1e283168 ORIGINAL ARTICLE

Machine learning for ranking multivariate variables in cattle breeds raised in Paraguayan wetlands[1]

Aprendizado de máquina para classificação de variáveis multivariadas em raças bovinas criadas em pântanos do Paraguai

Walter E. Pereira[2*], Liz M. Centurión[3,4], Carolina Valdez[4] & Roberto Martínez-López[4,5]

[1] Research developed at Universidad Nacional de Asunción, Centro Multidisciplinario de Investigaciones Tecnológicas, San Lorenzo, Central, Paraguay
[2] Universidade Federal da Paraíba/Centro de Ciências Agrárias, Areia, PB, Brazil
[3] Universidad Nacional de Asunción/Facultad de Ciencias Exactas y Naturales, San Lorenzo, Central, Paraguay
[4] Programa Universitario de Becas para la Investigación Andrés Borgognon Montero, Asunción, Paraguay
[5] Universidad Nacional de Asunción/Centro Multidisciplinario de Investigaciones Tecnológicas, San Lorenzo, Central, Paraguay

HIGHLIGHTS:

Applying machine learning models reveals critical variables for breeding and selecting cattle in Paraguayan wetlands.
Shapley additive explanations detail the importance of phenotypic and blood variables in cattle.
The machine learning approach can be used for genetic selection strategies adapted to wetland environmental conditions.

ABSTRACT: This study focuses on the performance of cows for meat production raised in the wetlands of Paraguay, examining five cattle genotypes: Brahman, Brangus, and Nelore, as well as two local breeds at risk of extinction. The main objective is to identify and rank phenotypic variables, including blood, clinical, hair, and health variables, demonstrating causal linkage with the live weight of the cows analyzed. Initially, high correlations were identified between different variables included in this study; then, using advanced Machine learning (ML) techniques and the application of Shapley additive explanations (SHAP), a deeper understanding was provided of the factors strongly associated with adaptability in these environments, and, therefore, the respective zootechnical performance. The association between cattle genotypic components linked with the season of the year proved to be the most influential factor on cattle live weight. Variables such as hair length, hematocrit, phosphatase, phosphorus, creatine phosphokinase, creatinine, protein, cortisol, calcium, and the presence of endoparasites were highlighted, demonstrating their hierarchical importance for animal selection. ML models are effective tools for establishing hierarchies of relevance in complex phenotypic multivariable, which is crucial in breeding programs for different zootechnical species and in special and specific environments like wetlands.

Key words: cattle adaptability in wetlands, SHAP, phenotypic variables, blood variables

RESUMO: Este estudo foca no desempenho de vacas para produção de carne, criadas nos pântanos do Paraguai, examinando cinco genótipos bovinos; Brahman, Brangus, Nelore, bem como duas raças locais em risco de extinção. O principal objetivo é identificar e classificar variáveis fenotípicas que incluem variáveis sanguíneas, clínicas, de pelagem e saúde, demonstrando ligação causal com o peso vivo das vacas analisadas. Inicialmente, foram identificadas correlações elevadas entre diferentes variáveis incluídas neste estudo, e, então, utilizando técnicas avançadas de aprendizado de máquina e a aplicação de explicações aditivas de Shapley (SHAP), foi proporcionado um entendimento mais profundo dos fatores fortemente associados à adaptabilidade nestes ambientes, e, portanto, o respectivo desempenho zootécnico. A associação entre o componente genotípico bovino ligado à estação do ano mostrou ser o fator influente mais predominante sobre o peso vivo bovino. Variáveis como comprimento do pelo, hematócrito, fosfatase, fósforo, creatina phosphokinase, creatinina, proteína, cortisol, cálcio e a presença de endoparasitas foram destacadas, demonstrando sua importância hierárquica para a seleção animal. Os modelos de ML são ferramentas eficazes para estabelecer hierarquias de relevância em multivariáveis fenotípicas complexas, o que é crucial em programas de melhoramento genético em diferentes espécies zootécnicas, bem como em ambientes especiais e específicos, como os pântanos.

Palavras-chave: adaptabilidade bovina em pântanos, SHAP, variáveis fenotípicas, variáveis sanguíneas

• Ref. 283168 – Received 09 Feb, 2024
* Corresponding author - E-mail: wep@cca.ufpb.br
• Accepted 22 Jul, 2024 • Published 30 Jul, 2024
Editors: Ítalo Herbet Lucena Cavalcante & Carlos Alberto Vieira de Azevedo

Introduction

Paraguay is a country with strong cattle farming, producing in different edaphoclimatic and agroecological systems. Two of these environments are special and specific, such as the semi-arid Chaco and the freshwater wetlands, known as Ramsar sites (RAMSAR, 2023). They are special environments due to their vulnerable sensitivity level and specific for their unique edaphoclimatic characteristics.

This study proposes an approach of multivariate phenotypic variables, including blood, clinical, parasitic, hair metrics, and traditional live weights variables. It is postulated that this comprehensive approach can greatly approximate the scientific community to more efficient conceptualizations and hierarchizations than those currently used in the zootechnical/environmental field, with the support of different predictive models of machine learning (ML).

Artificial intelligence (IA) is a field that combines computer science (statistical models and algorithms) and datasets to create intelligent systems capable of performing tasks that normally require human intelligence (Sarker, 2023). ML is a subfield of IA that uses complex algorithms to solve challenging problems that are difficult to address with traditional approaches. For example, machine learning algorithms, including deep learning, can be used to assess some behaviors of confined animals, such as feeding, posture, and aggression, allowing breeders to take measures to ensure the welfare of cows and maximize productivity (Chen et al., 2021).

The main uses of ML identified in animal production are the detection of diseases (especially mastitis), the prediction of milk production, and the prediction of milk quality using classification or regression models (Slobe et al., 2021). Similarly, other studies in the area were recorded (Ruchay et al., 2022; Xu et al., 2024).

The main objective is to identify and rank phenotypic variables that include blood, clinical, hair, and health variables, demonstrating causal linkage with the live weight of the cows analyzed.

Material and Methods

In this study, a dataset derived from the study published on adaptive variables of different cattle genotypes raised in the wetlands of Ñeembucú, Paraguay, and its areas of influence (Martínez-López, 2020; Martínez-López et al., 2021) was used, in livestock farms located at the coordinates reported in Table 1.

Since this study did not involve any invasive procedure in data collection, no ethics committee approval was required. Evaluations were conducted on 80 cows between four and five years of age in four seasons over one year, totaling 320 observations. The genotypes were Brahman (BH), Brangus (BG), Nelore (NE), and additionally, two small local genetic groups called Criollo Pilcomayo (CP) and Criollo Ñeembucú (CÑ), corresponding to 16 animals each genotype.

In each observation unit, 32 variables were evaluated, which were related to animal welfare, sanitary condition, and metabolic profile, being the following: ruminal frequency, respiratory frequency, heart rate, body temperature (°C), cortisol concentration (nmol L^{-1}), length (cm) and hair density (quantity per cm^2), live weight (kg), tick count in different body regions (flank, head, neck, armpit, and groin) and the total frequency, concentrations of hemoglobin (g L^{-1}), hematocrit (%), cholesterol (mmol L^{-1}), triglyceride (mmol L^{-1}), GOT - glutamic oxalic transaminase (UI mL^{-1}), alkaline phosphatase (UI mL^{-1}), GGT - gamma-glutamyl transferase (UI mL^{-1}), CPK - creatine phosphokinase (UI mL^{-1}), calcium (mmol L^{-1}), phosphorus (mmol L^{-1}), magnesium (mmol L^{-1}), sodium (mmol L^{-1}), urea (mmol L^{-1}), creatinine (µmol L^{-1}), total proteins (g L^{-1}), albumin (g L^{-1}), and globulin (g L^{-1}), the presence of endoparasite (yes/no) and the quantity of endoparasites identified. All variables were measured according to the procedures described by Martínez-López (2020) and Martínez-López et al. (2021). Fieldwork was conducted quarterly, totaling four measurements, coinciding with the spring, summer, autumn, and winter seasons 2016.

The combination between the five cattle genotypes and the year's four seasons was also calculated, constituting the 33rd variable. The weight data as a function of the 33 described variables were adjusted to the following machine learning algorithms: deep learning, distributed random forest (DRF), extremely randomized trees (XRT), generalized linear model with regularization, gradient boosting machine (GBM), and stacked ensembles. These algorithms are available in the h2o library through the h2o AutoML interface. For this purpose, the dataset was divided into training (60%) and testing (40%). All 33 variables were initially included, selecting those with at least 2% relative importance for the final model. All variables were normalized to the range of 0 to 1 to prevent a certain variable from influencing the model due to its greater variability, applying the following Eq. 1:

$$\text{Normalized}_{\text{data}} = \frac{\left(\text{data} - \min_{\text{values}}\right)}{\text{ranges}} \quad (1)$$

where:

data - observed values;

$\min_{\text{values}}$ - minimal values; and,

ranges - the difference between the largest and smallest value of each variable.

Table 1. Geographical coordinates of cattle ranches in the Ñeembucú wetlands (Paraguay) and areas of influence included in this study

Ranches	Geographical coordinates	Genotypes included
Isla Umbú	27° 5' 13.87" S and 58° 24' 53.48" W, altitude 52 masl	Criollo Neembucú and Brangus
Nueva Italia	25° 30' 49.87" S and 57° 28' 32.47" W, altitude 113 masl	Brangus and Criollo Pilcomayo
San Miguel	26° 37' 30.96" S and 57° 3' 0.94" W, altitude 113 masl	Brahman and Brangus
Caapucú	26° 16' 30.06" S and 57° 9' 51.71" W, altitude 65 masl	Nelore and Brangus

masl - Meters above sea level

The following performance indexes were considered for model selection: Eqs. 2 and 3 (Janssen & Heuberger, 1995):

- Root mean square error (RMSE):

$$RMSE = \sqrt{\left[\frac{\sum_{i=1}^{n}(S_i - O_i)^2}{n}\right]} \quad (2)$$

- Mean absolute error (MAE):

$$MAE = \left[\frac{\sum_{i=1}^{n}|S_i - O_i|}{n}\right] \quad (3)$$

where:

S_i - estimated values;

O_i - observed values; and,

n - number of samples.

Pearson's correlation was calculated and graphed between the observed and estimated values, considering the testing and training datasets.

SHAP (Shapley additive explanations) values were considered, a technique inspired by game theory that seeks to improve the interpretability of machine learning models to interpret the effects of the variables included in the final model. These values, known as SHAP values, provide a unified measure of feature importance. SHAP is effective in differentiating between different output classes and aligns its results better with human intuition compared to other methods (Linardatos et al., 2021).

The three desirable properties of SHAP are: a) Local accuracy ensures that the sum of the SHAP values for all features equals the model's prediction. This means that the explanations accurately reflect the model's behavior; b) Missingness establishes that an absent feature receives a zero attribution. This prevents assigning importance to irrelevant or missing features; c) Consistency states that if a model is modified so that the marginal contribution of a feature increases or remains the same, the SHAP value also increases or remains the same. This ensures that the explanations are consistent with changes in the model. Together, these properties make SHAP a very reliable and useful method of explanation for understanding the influence of features in ML model predictions (Linardatos et al., 2021).

Specifically, in this paper, these three SHAP techniques were used: a) SHAP feature importance: calculates the average importance of each feature by analyzing the absolute value of its Shapley contribution in all instances. It is visualized with a bar chart ordered by decreasing importance; b) SHAP summary: combines importance with effects. Each point shows the Shapley contribution of a feature for a specific instance. This allows us to see how the value of a feature impacts the individual prediction; c) SHAP dependence: shows the relationship between the value of a feature and its impact on the prediction. It analyzes each instance individually, generating a graph with the variable on the horizontal axis and the Shapley contribution on the vertical. The result is similar to a scatter plot that reflects the variation of importance according to the value of the feature (Molnar, 2022).

The h2o library (Fryda et al., 2023) available in the R language (R Core Team, 2023) was used to adjust the ML models.

Results and Discussion

In Table 2, all the predictive variables included in the initial model are observed and listed in a hierarchical scale. The combined factor between seasons of the year with genotypes included in the study presents the greatest scaled importance, with a value of 1 and 29.42% of the total deviance value, followed by hair length with 11.25% of the total value. Creatinine is the last variable chosen to be part of the final model, with 2.48% of the total value, accumulating with the 11 variables 73.24% of the total deviance value of the dataset. The other variables were excluded from the final model due to their low importance. It should be noted that a

Table 2. Scaled importance and percentage expressed of all variables included in the initial machine learning model to estimate the live weight of five cattle genotypes in four seasons of the year in the Ñeembucú wetlands

Variable	Scaled importance	%	Variable	Scaled importance	%
Season_genotype	1.00	29.42	Albumin (g L^{-1})	0.05	1.52
Hair length (cm)	0.38	11.25	Cholesterol (mmol L^{-1})	0.05	1.49
Hematocrit (%)	0.19	5.66	Resp freq	0.05	1.47
Phosphatase (UI Ml^{-1})	0.16	4.86	Tick neck	0.05	1.46
Endoparasites quantity	0.15	4.42	GGT (UI Ml^{-1})	0.04	1.40
Phosphorus (mmol L^{-1})	0.12	3.67	Globulins (g L^{-1})	0.04	1.35
Cortisol (nmol L^{-1})	0.10	2.96	Body temperature (°C)	0.04	1.32
Calcium (mmol L^{-1})	0.10	2.95	Hair density (cm^2)	0.03	1.09
Total proteins (g L^{-1})	0.09	2.80	Tick armpit	0.03	1.06
CPK (UI mL^{-1})	0.09	2.77	Urea (mmol L^{-1})	0.03	1.02
Creatinine (μmol L^{-1})	0.08	2.48	Total tick	0.03	0.88
Hemoglobin (g L^{-1})	0.06	1.91	Heart rate	0.02	0.63
Sodium (mmol L^{-1})	0.06	1.87	Ruminal freq	0.01	0.57
Magnesium (mmol L^{-1})	0.06	1.81	Presence of endoparasite	0.01	0.53
Triglycerides (mmol L^{-1})	0.05	1.69	Tick head	0.01	0.38
Tick groin	0.05	1.57	Tick flank	0.01	0.19
GOT (UI mL^{-1})	0.05	1.54			

CPK - Creatine phosphokinase; GOT - Glutamic oxalic transaminase; GGT - Gamma-glutamyl transferase

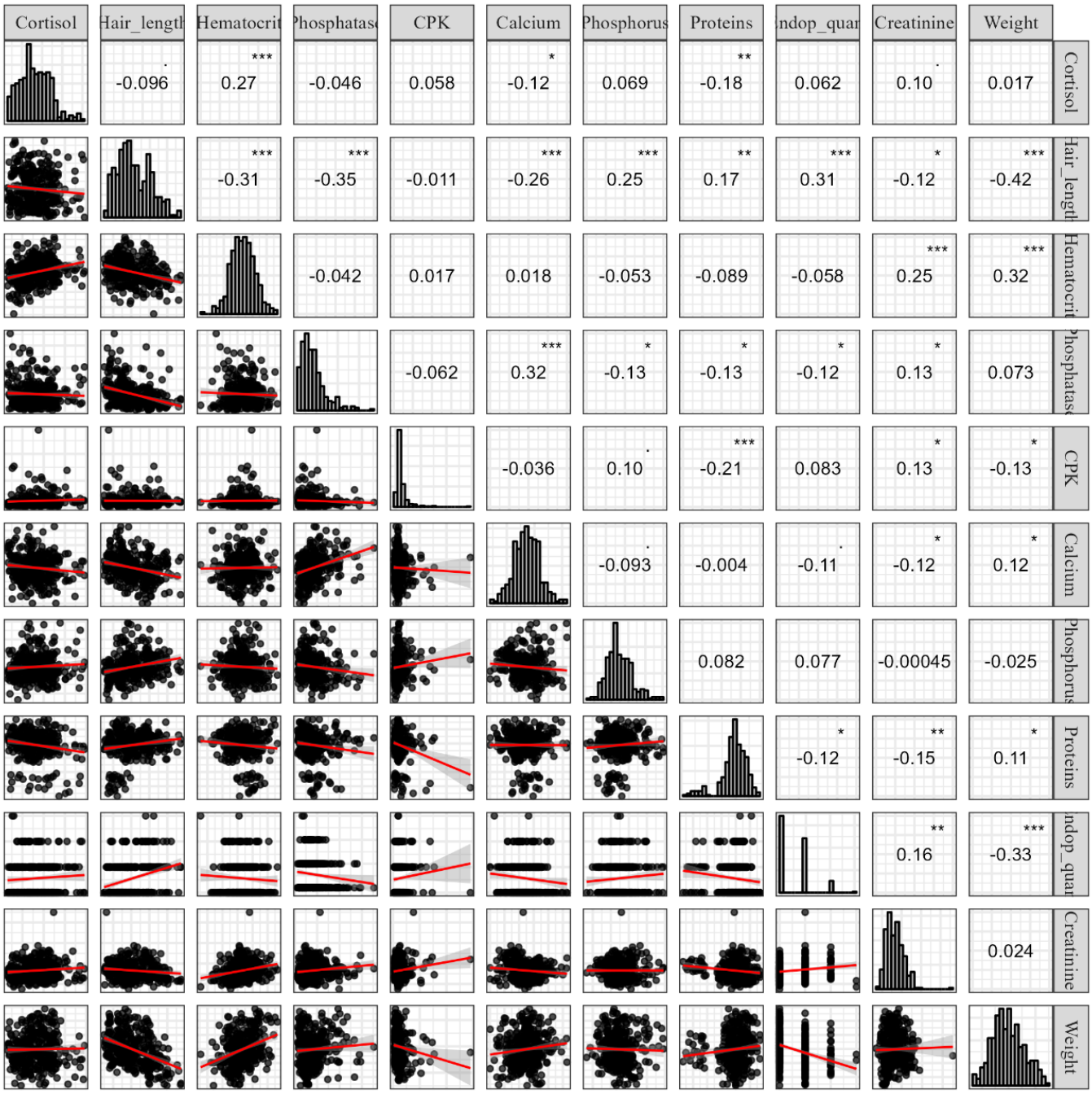
"threshold" of inclusion in importance was considered, studied phenotypic variables that have a minimum of 2% quantitative preponderance calculated upwards.

In this regard, in machine learning, the analysis of the individual contribution of the variables to the model is a very relevant procedure in observance of the consideration of more precise models (Shahzad et al., 2022).

In Figure 1, the correlations between the pairs of variables evaluated in the five cattle genotypes are visualized, considering only the variables with relative importance greater than 2%, as described in Table 2.

Considering the coefficients obtained between the cattle weight associated with the other variables considered, significant relationships are verified only with hair density, hematocrit, calcium, total proteins, CPK, and quantity of endoparasites (Figure 1). According to Rusakov (2022), these values suggest a weak ($r < 0.3$) to moderate ($0.3 < r < 0.7$) relationship. However, the causal effect is emphasized in the referred variables according to the statistical inference used. In this context, Ruchay et al. (2022) point out that machine learning methods can generate more precise estimates compared to the linear regressions usually used in weight prediction, a postulation verified by them when analyzing the relationship of this variable with morphometric variables and age in cows, in the cited case, of the Hereford breed.

According to Yavuz Ozalp et al. (2023), the ability to handle high-dimensional and complex nonlinear data is one of the main advantages of ML compared to traditional statistical techniques. Thus, in line with the low correlations of weight with the other variables evidenced in this study, machine



*, **, ***: Significant at $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$ by t-test. CPK - Creatine phosphokinase; Endop_quant - Endoparasites quantity; Proteins - Total proteins

Figure 1. Correlation of the selected variables based on relative importance (greater than 2%) for the final machine learning model

learning algorithms are justified to estimate body live weight based on the calculated predictor variables.

Table 3 exposes some basic descriptive statistics of the variables considered in the final model. In this context, it was evidenced that the average live weight of cattle was 298.73 kg, with a maximum value equal to 443.00 kg. The variable with the smallest amplitude is the calcium concentration, with a value of 1.32 mmol L^{-1}, while the concentration of CPK presents the greatest amplitude, equivalent to 5746 UI mL^{-1}. This evidences the need for normalization in the range of 0 to 1 to prevent variables with greater amplitude from presenting greater importance in the model solely due to their great variability.

Table 4 shows the performance of the six prediction models of live weight associated with the phenotypic variables considered in the study, with the best adjustments in the dataset, quantified through RMSE and MAE.

This study selected model 144, corresponding to the GBM algorithm, considered one of the latest generation ensemble methods based on trees (Sahin, 2020). Although the model did not report the smallest values of RMSE and MAE among the first five mentioned in Table 4, considering that there is a minimal difference, it allows obtaining the relative importance of each variable included in the model for the prediction of live weight, which was marked as one of the objectives of this research. Thus, with the selected model, observing Figure 2, a

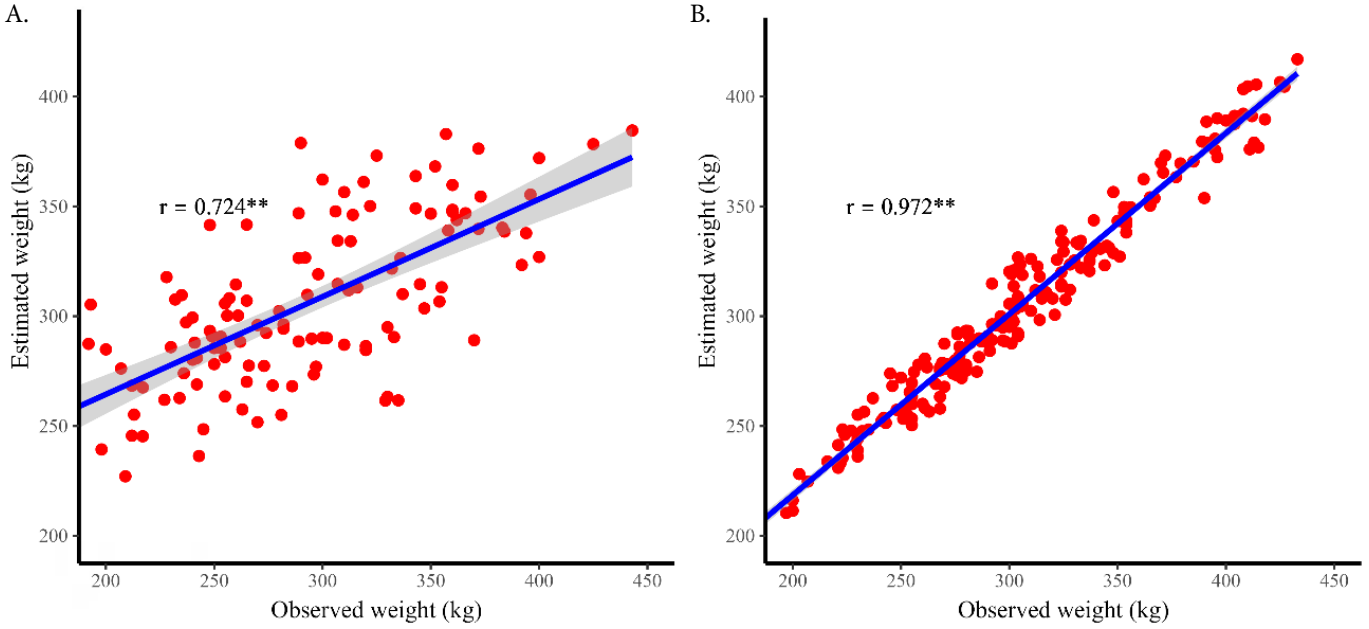
Table 3. Descriptive measures for the variables selected in the final machine learning model to estimate the weight of five cattle genotypes in four seasons of the year

Variables	Minimum	First quartile	Median	Mean	Third quartile	Maximum	SEM (n=320)
Weight (kg)	170.00	255.00	296.00	298.73	337.00	443.00	3.24
Cortisol (nmol L^{-1})	32.84	119.99	175.54	181.69	241.09	433.32	4.62
Hair length (cm)	0.50	1.35	1.91	2.08	2.82	4.75	0.05
Hematocrit (%)	17.70	28.90	31.65	31.68	34.50	43.30	23.00
Phosphatase (UI mL^{-1})	53.00	224.20	351.00	417.26	526.50	1767.00	15.53
CPK (UI mL^{-1})	42.00	279.50	375.00	491.28	495.20	5,788.00	28.49
Calcium (mmol L^{-1})	1.75	2.21	2.36	2.36	2.51	3.07	0.01
Phosphorus (mmol L^{-1})	0.53	1.61	1.91	2.01	2.34	3.99	0.03
Total proteins (g L^{-1})	43.00	76.00	80.00	78.94	85.00	97.00	0.53
Endoparasites quantity	0.00	0.00	0.00	0.56	1.00	3.00	0.03
Creatinine (μmol L^{-1})	66.30	117.60	143.70	146.67	168.80	397.80	2.19

CPK - Creatine phosphokinase; SEM – Standard error of the mean

Table 4. Performance indexes of the six models with the best adjustments based on root mean square error (RMSE) and mean absolute error (MAE)

Models	RMSE	MAE
StackedEnsemble_BestOfFamily_3_AutoML_1_20231215_173609	38.62	32.08
StackedEnsemble_BestOfFamily_4_AutoML_1_20231215_173609	39.57	32.44
StackedEnsemble_BestOfFamily_6_AutoML_1_20231215_173609	39.85	33.11
StackedEnsemble_AllModels_2_AutoML_1_20231215_173609	39.92	32.86
StackedEnsemble_BestOfFamily_7_AutoML_1_20231215_173609	39.94	33.10
GBM_grid_1_AutoML_1_20231215_173609_model_144	40.16	33.37



** - Significant at $p \leq 0.01$ by t-test

Figure 2. Correlation coefficient (r) between observed and estimated values by the selected model, gradient boost machine (GBM), considering testing (A) and training datasets (B)

high predictive capacity of the same is verified, with correlation coefficients between estimated and observed weight values quite strong (r > 0.7) (Rusakov, 2022; Zúñiga et al., 2022) and statistically significant, both for the data used in identifying the behavioral pattern given in the training dataset (Figure 2A) and for the test dataset (Figure 2B); with high values of r (0.724 and 0.972, respectively).

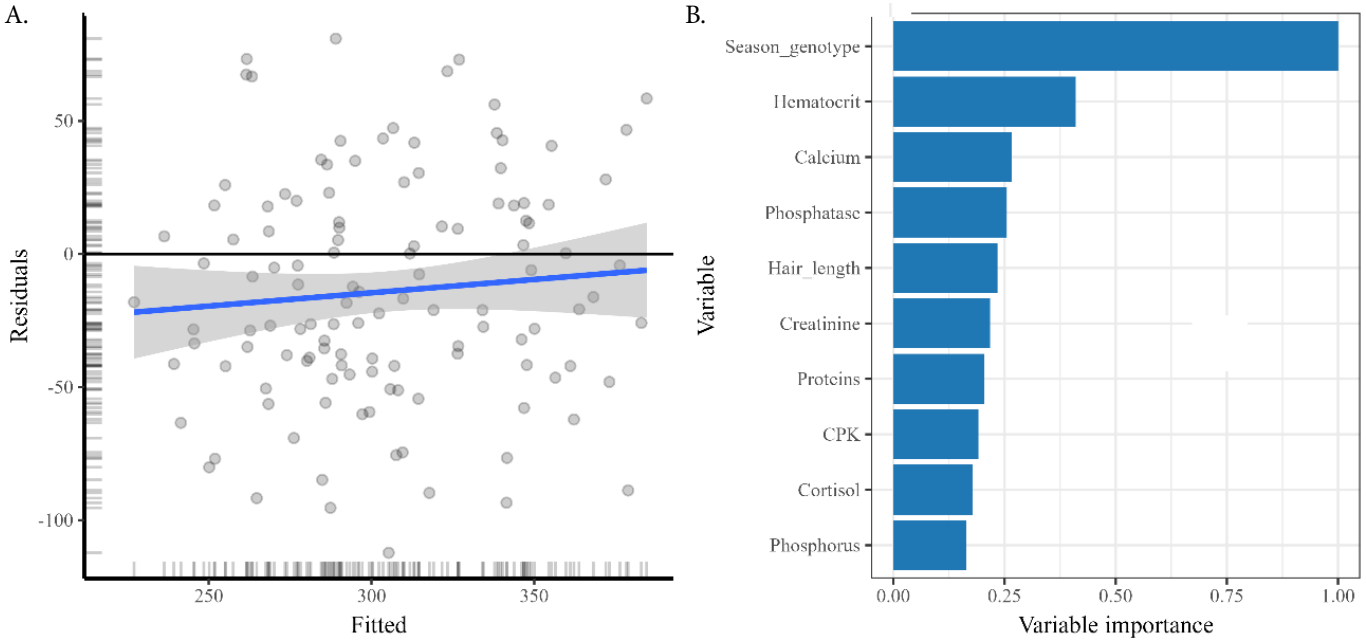
In Figure 3A, the distribution of the residuals in the final model is observed, verifying random distribution as a function of the estimated values without any trend or major noises. Figure 3B shows the relative importance of the predictor variables of live weight in cattle. The most influential factor was the combination of the five cattle genotypes with the four seasons of the year analyzed, presenting a value of 1, immediately followed by hair density as well as blood hematocrits. Likewise, the phosphatase, cortisol, CPK, and creatinine factors contributed substantially to the estimation of live weight, with values between 0.125 and 0.25. In a similar study on weight prediction reported by Ruchay et al. (2022), they found that the ExtraTreesRegressor algorithm was the best prediction model of the referred variable based on morphometric traits and age of cows, where the latter was the most influential variable, with the minimum proportion of the relative importance in percentage of the metric body characteristics, presenting values between 4% and 6%, similarly to those verified in this research.

The relationship between the predictive variables in each animal and their impact on the prediction of live weight can be visualized in Figure 4. Considering the greater relative importance of the combination of seasons of the year and cattle genotypes, this variable presents the greatest variability in SHAP contribution values, fluctuating between -40 to 45 kg approximately.

It is also verified that the lower the hair length, the higher the live weight increases, whereas parallelly, a greater number of animals with longer hair length and lower weight are evidenced, remembering that these animals are raised in the humid heat of this specific ecosystem (freshwater wetlands). Another aspect to highlight is that low concentrations of phosphatase and CPK imply an increase in weight in animals, while creatinine has the opposite effect. Likewise, when the quantity of endoparasites increases, the weight decreases in all animals, as can be verified by the blue coloration of the SHAP, without exception, recorded in its contribution (Figure 4), increasing its value.

On the other hand, phosphorus associated with weight did not show statistical importance, remembering that the studied cows were adults, suggesting that these levels of phosphorus, low in blood, evidenced in most of the animals (Figure 4), are related to their age condition, as stated by Bonifaz et al. (2023).

Shyma et al. (2015) reported that cattle breeds with thicker coats and long hair are more susceptible to tick infestation than those with shorter hair. This leads to thinking that it also affects live weight. Gebremedhin et al. (2023) postulated that it is more likely that this coat characteristic is expressed in *Bos taurus* cattle breeds (long hair), where they mentioned the importance of this character in the exchange of heat and humidity between cattle and their environment. Returning to the parasitic issue, in a more recent study, Feltes et al. (2021) reported a very low genetic correlation between tick count and coat traits in a mixed Angus x Nelore population, which suggests that different genes regulate these traits. However, racial origin and coat characteristics are strongly dominant in cattle performances for meat in hot and humid environments.



CPK - Creatine phosphokinase; Endop_quant - Endoparasites quantity; Proteins - Total proteins

Figure 3. Residuals as a function of the estimated weight values (A) and relative importance of the variables included in the final machine learning model (B)

Endop_quant - Endoparasites quantity; CPK - Creatine phosphokinase; Proteins - Total proteins

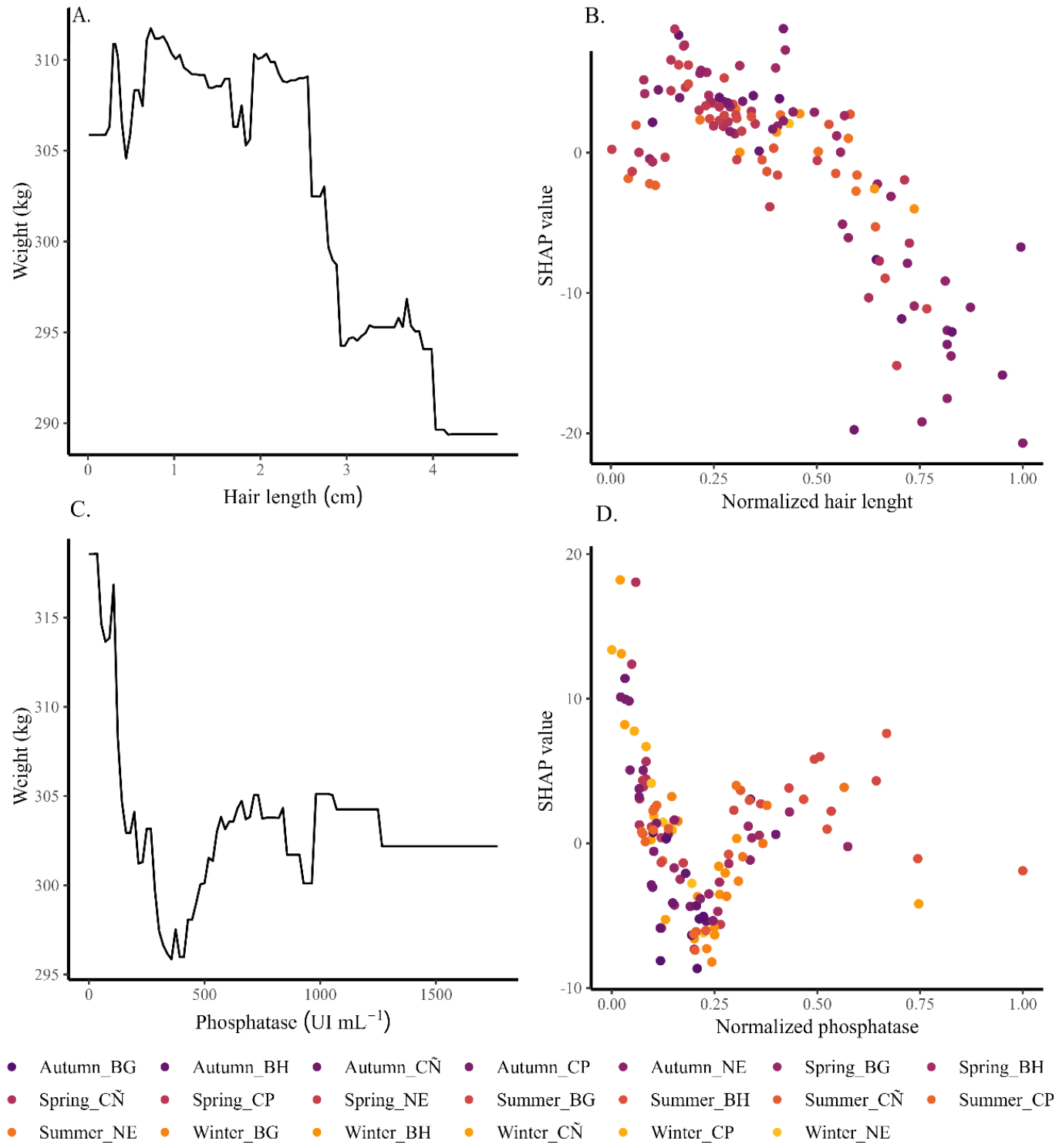
Figure 4. SHAP (Shapley additive explanation) contribution, in kg, of the variables included in the final machine learning model, for the increase or decrease in the weight of five cattle genotypes analyzed with data from four seasons of the year

Comparing Figures 5A with B, it is possible to verify the relationship between the weight values and the corresponding SHAP dependence values and the relationship between the original hair length values and their normalized values, varying from 0 to 1. The same happens with phosphatase (Figures 5C and D) and all the other predictor variables. These two examples facilitate appropriately interpreting the other variables' SHAP dependence figures in the subsequent figures without presenting the original weight values.

The increase in hair length, up to approximately 2.50 cm, equivalent to the normalized value of 0.50, does not affect live weight. With higher values, a decrease of up to 40 kg in weight is observed (Figures 5A and B), with more cases in autumn and winter. On the other hand, with the increase in phosphatase concentration in the blood, an exponential decrease in body weight is verified, stabilizing from approximately 250 UI mL^{-1}, corresponding to approximately 0.15 on the normalized scale (Figures 5C and D).

In Figures 6A, B, C, D, 7A, B, C, and D, the SHAP dependence values corresponding to the other variables selected in the model are shown, namely calcium concentrations, cortisol, CPK, creatinine, hematocrit, total proteins, phosphorus, and quantity of endoparasites identified in the five cattle genotypes during the four seasons of the year.

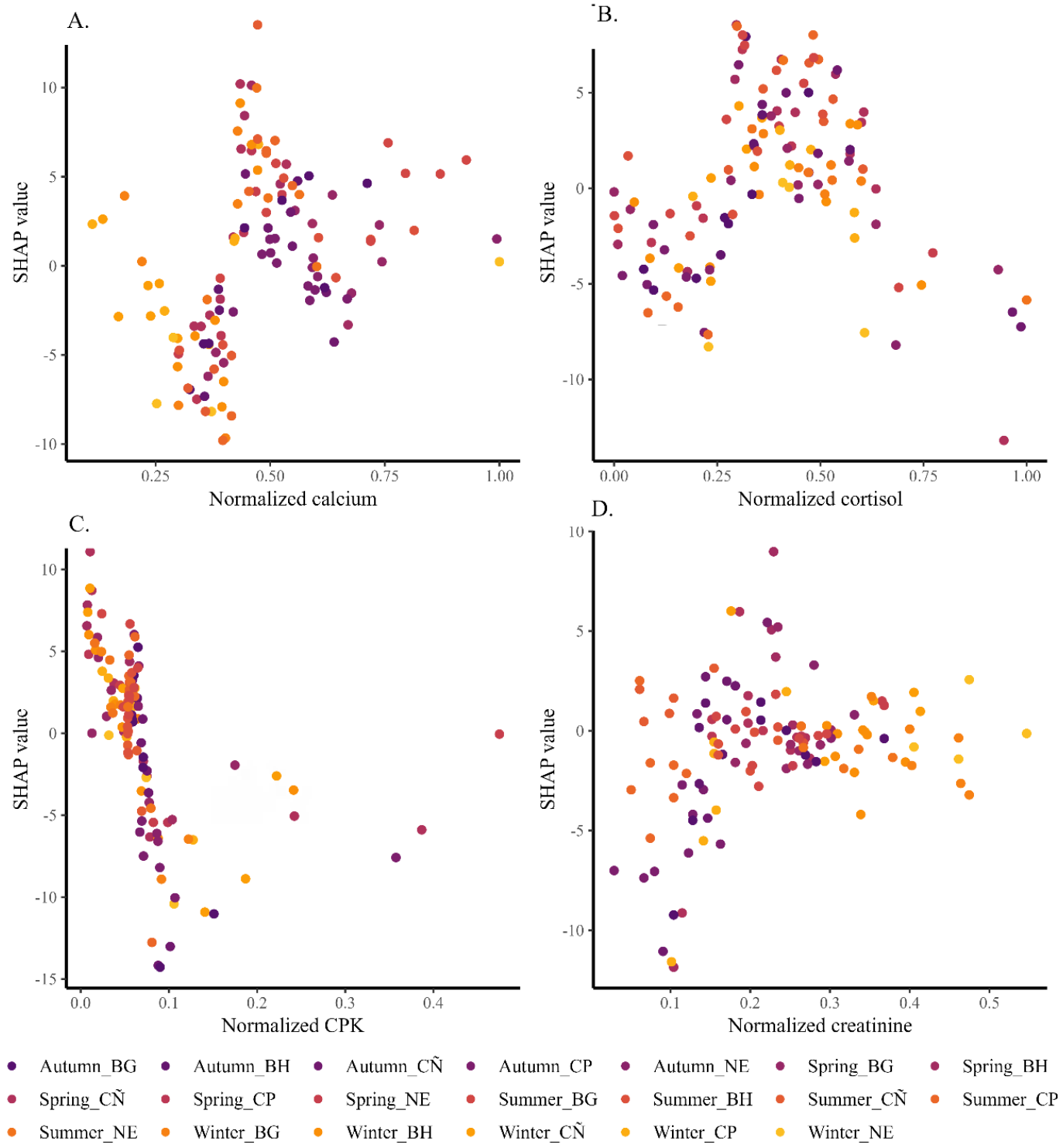
In a study conducted by Martínez-López et al. (2021), it was concluded that exotic genotypes in unfavorable environments (such as the humid heat of the study



BG - Brangus; BH – Brahman; CN - Criollo Ñeembucú; CP - Criollo Pilcomayo; NE - Nelore

Figure 5. Estimated weight as a function of hair length and phosphatase concentration and their equivalence with the SHAP (Shapley additive explanation) values (kg) and the normalized values in the range of 0 to 1, respectively, in five cattle genotypes and four seasons of the year

area), high concentrations of creatinine would be a potential indicator of the activation of mechanisms that compromise tissue protein integrity and renal function. This coincides with the observed results regarding the behavior of creatinine and protein. Likewise, the same study showed the relevance of the "season of the year" and "cattle genotype" factors with concentrations of total proteins, which would be associated with better (or not) adaptation processes from the metabolic pathways of nitrogen compounds in the blood. On the other hand, elevated blood concentrations of CPK, urea, and creatinine may be associated with some viral disease of the animal, possibly causing, in addition, hepatic, muscular, and renal lesions (Jalali et al., 2017), which may be relevant for the progression of the disease and weight loss (Kamr et al., 2022). Regarding cortisol, it is clearly an indicator of animal welfare, which, at high blood levels, is directly linked to animal stress, considerably impacting live weight



BG - Brangus; BH – Brahman; CN - Criollo Ñeembucú; CP - Criollo Pilcomayo; NE - Nelore

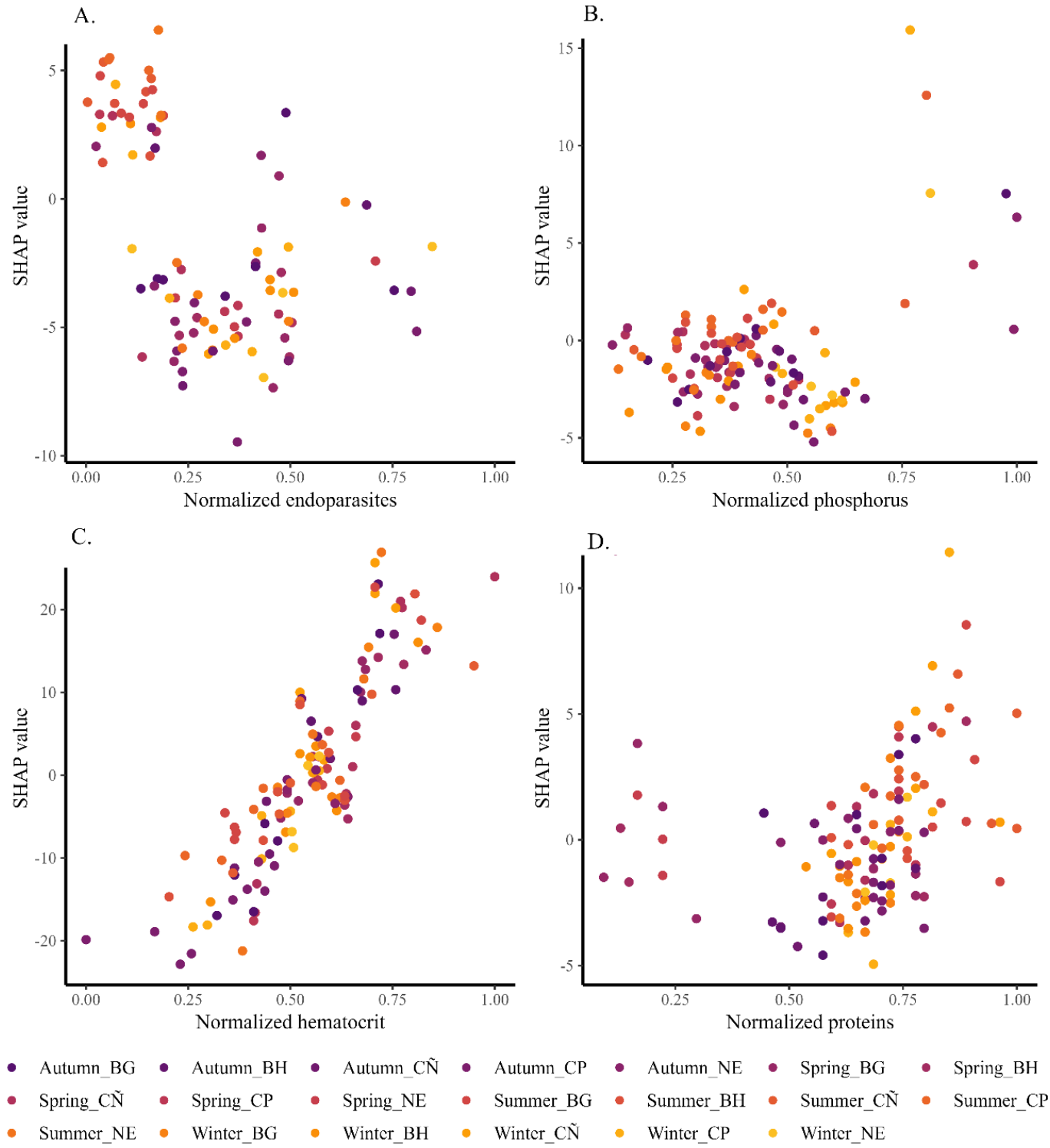
Figure 6. SHAP dependence values (kg) as a function of the normalized values in the range of 0 to 1 of calcium, cortisol, CPK (Creatine phosphokinase), and creatinine in five cattle genotypes during four seasons of the year

(Larios-Cueto et al., 2019), evidenced in the results of this study, marking the importance of this variable as a strategy for identifying the adaptability of a genotype in special and specific environments.

According to Ott et al. (2023), the breed factor in breeding cows is highly dominant in the variation of blood calcium concentrations, which, in a certain way, coincides with what was observed in this study.

Figure 8 shows the partial dependence graph between the estimated weight and the five cattle genotypes during the year's four seasons; higher values were verified in autumn and summer.

According to the results observed in the present study, it is necessary to point out that the combination of genotype and season of the year marks the guideline on the potential of cattle in the breeding condition of calves for meat production in the humid heat of this special and specific environmental zone, on the ability to present better live weights of these animals and cattle capacity.



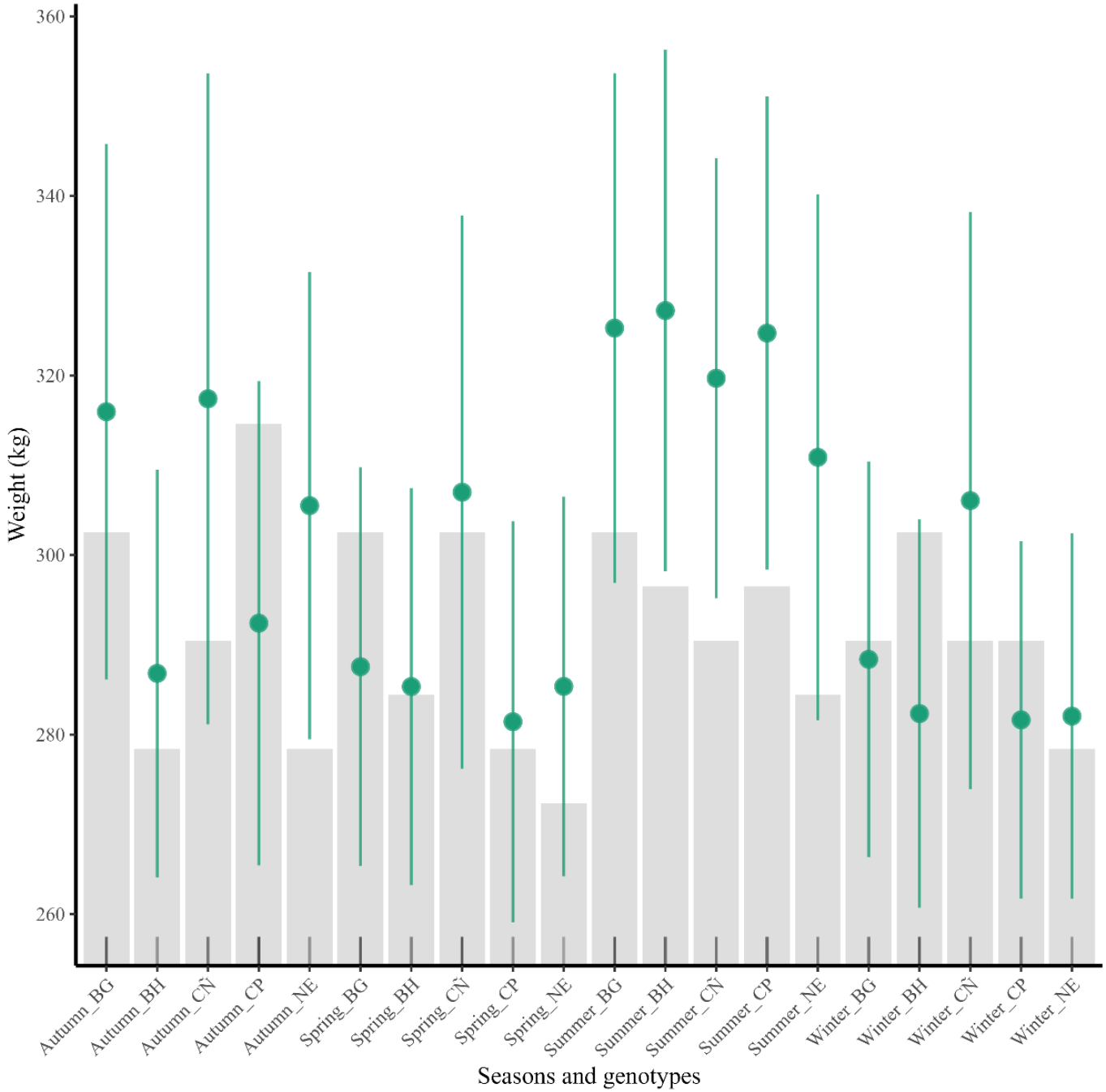
BG - Brangus; BH – Brahman; CN - Criollo Ñeembucú; CP - Criollo Pilcomayo; NE - Nelore

Figure 7. SHAP dependence values (kg) as a function of the normalized values in the range of 0 to 1 of the number of endoparasites, phosphorus, hematocrit, and total protein in five cattle genotypes during four seasons of the year

It is important to mention that conclusions that do not include studies of the association of multivariate phenotypic variables of cattle related to their correct genetic identification, obtained with traditional analysis methods, would represent a potential risk of a substantial error when establishing selection criteria and predictive strategies for animal improvement programs.

It is necessary to work with estimates with different machine learning models to establish and categorize a ranking of phenotypic variables to consider in a genetic selection program, a path of extreme relevance to develop, explore, and concretize in front of traditional strategies currently used in the environmental and zootechnical field.

In future research, it is recommended to use other reference variables, such as body condition, prolificacy, or categorization of a ranking of elemental importance, analyzed by predictive models of ML associated with multiple cattle phenotypic variables; it is undoubtedly a path to follow with less bias, to choose individuals adapted to the breeding and production environment.



BG - Brangus; BH – Brahman; CN - Criollo Ñeembucú; CP - Criollo Pilcomayo; NE - Nelore

Figure 8. Partial dependence graph of the estimated weight and the five cattle genotypes during four seasons of the year

Conclusions

1. The variables hair length, hematocrit, phosphatase, phosphorus, creatine phosphokinase, creatinine, protein, cortisol, calcium, and the number of endoparasites present in the animal demonstrated a robust hierarchical categorization, becoming potential selection indexes associated with live weight.

2. The hierarchical contribution provided by the use of Shapley additive explanation (SHAP) proved to be quite useful and strategic in estimating the importance of the variables included in this study.

3. The machine learning models that showed the best performance and suitability for this study and purpose were the stacked ensemble and the gradient boosting machine (GBM), with the latter being the chosen method for the final modeling.

4. The machine learning approach has proven useful and strategic for determining relevance hierarchies from multiple variables in this study.

Contribution of authors: Martínez-López, R.: Performed the experiments, collected data, prepared the first version of the manuscript, and conducted the literature review. Pereira, W. E.: Performed the data analysis, implemented the computational models, and prepared the first version of the manuscript. Centurión, L. M. and Valdez, C.: Conducted the literature review and made corrections to the manuscript.

Supplementary documents: There are no supplementary documents.

Conflict of interest: The authors declare no conflict of interest.

Financing statement: This research was financed by the National Council of Science and Technology of Paraguay (CONACYT), protocol BINV02-84.

Acknowledgments: To the National Council of Science and Technology of Paraguay (CONACYT), through the financing of the Research Stay (BINV02-84) corresponding to the Paraguayan Program for the Development of Science and Technology (PROCIENCIA II). To the Multidisciplinary Center for Technological Research of the National University of Asunción of Paraguay (CEMIT/UNA). To the Federal University of Paraíba, Brazil (UFPB). To the University Program of Scholarships for Research, "Andrés Borgognon Montero" (PUBIABM).

Literarure Cited

Bonifaz, V. de los Á.; Morán, D. L. T.; Morales, L. I. C.; García, Y. I. M. Efeito do fósforo (P) e do cálcio (Ca) na produção de laticínios. Brazilian Journal of Animal and Environmental Research, v.6, p.3251-3267, 2023. https://doi.org/10.34188/bjaerv6n4-014

Chen, C.; Zhu, W.; Norton, T. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. Computers and Electronics in Agriculture, v.187, e106255, 2021. https://doi.org/10.1016/j.compag.2021.106255

Feltes, G. L.; Michelotti, V. T.; Prestes, A. M.; Rorato, P. R. N.; Mello, F. C. B.; Oliveira, M. M.; Roso, V. M. Genetic and environmental factors that affect tick infestation in Nelore x Angus crossbreed cattle. Ciência Rural, v.51, e20200677, 2021. https://doi.org/10.1590/0103-8478cr20200677

Fryda, T.; LeDell, E.; Gill, N.; Aiello, S.; Fu, A.; Candel, A.; Click, C.; Kraljevic, T.; Nykodym, T.; Aboyoun, P.; Kurka, M.; Malohlava, M.; Poirier, S.; Wong, W. h2o: R Interface for the 'H2O' Scalable Machine Learning Platform_. R package version 3.44.0.2, 2023. Available on: <https://github.com/h2oai/h2o-3>. Accessed on: Nov. 2023.

Gebremedhin, K. G.; Fonseca, V. D. F.; Maia, A. S. C. Methods, thermodynamic applications, and habitat implications of physical and spectral properties of hair and haircoats in cattle. *Animals*, v.13, e3087, 2023. https://doi.org/10.3390/ani13193087

Jalali, S. M.; Rasooli, A.; Seifi Abad Shapuri, M.; Daneshi, M. Clinical, hematologic, and biochemical findings in cattle infected with lumpy skin disease during an outbreak in southwest Iran. Archives of Razi Institute, v.72, p.255-265, 2017. https://doi.org/10.22092/ari.2017.113301

Janssen, P. H. M.; Heuberger, P. S. C. Calibration of process-oriented models. Ecological Modelling, v.83, p.55-56, 1995. https://doi.org/10.1016/0304-3800(95)00084-9

Kamr, A.; Hassan, H.; Toribio, R.; Anis, A.; Nayel, M.; Arbaga, A. Oxidative stress, biochemical, and histopathological changes associated with acute lumpy skin disease in cattle. Veterinary World, v.15, p.1916-1923, 2022. https://doi.org/10.14202/vetworld.2022.1916-1923

Larios-Cueto, S.; Ramírez-Valverde, R.; Aranda-Osorio, G.; Ortega-Cerrilla, M.; García-Ortiz, J. Indicadores de estrés en bovinos por el uso de prácticas de manejo en el embarque, transporte y desembarque. Revista Mexicana de Ciencias Pecuarias, v.10, p.885-902, 2019. https://doaj.org/article/76f22014b1174c50a99e5fd87c578b70

Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, v.23, e18, 2021. https://dx.doi.org/10.3390/e23010018

Martínez-López, R.; Centurión-Insaurralde, L. M.; Núñez-Yegros, O. L.; Sponenberg, D. P. Protein Status in Cattle raised in the Wetlands of Paraguay during three periods of the year. Revista Científica de la Facultad de Ciencias Veterinarias de la Universidad del Zulia, v.32, p.1-9, 2021. https://doi.org/10.52973/rcfcv-e32081

Martínez-López, R. Estudio de parámetros adaptativos de diferentes genotipos bovinos criados en los humedales del Ñeembucu y su área de influencia, 14-INV-140. Paraguay: Consejo Nacional de Ciencia y Tecnología, 2020. 45p.

Molnar, C. Interpretable machine learning: A guide for making black box models explainable. 2.ed. Leanpub, 2022. 329 p.

Ott, D.; Manneck, D.; Schrapers, K. T.; Rosendahl, J.; Aschenbach, J. R. Blood calcium concentration and performance in periparturient and early lactating dairy cows is influenced by plant bioactive lipid compounds. Journal of Dairy Science, v.106, p.3706-3718, 2023. https://doi.org/10.3168/jds.2022-22387

RAMSAR. The List of Wetlands of International Importance. 2023. Available on: < https://www.ramsar.org/sites/default/files/documents/library/sitelist.pdf >. Accessed on: Nov. 2023.

R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2023. Available on: < https://www.R-project.org/ >. Accessed on: Nov. 2023.

Ruchay, A.; Kober, V.; Dorofeev, K.; Kolpakov, V.; Dzhulamanov, K.; Kalschikov, V.; Guo, H. Comparative analysis of machine learning algorithms for predicting live weight of Hereford cows. Computers and Electronics in Agriculture, v.195, e106837, 2022. https://doi.org/10.1016/j.compag.2022.106837

Rusakov, D. A. A misadventure of the correlation coefficient. Trends in Neurosciences, v.46, p.94-96, 2022. https://doi.org/10.1016/j.tins.2022.09.009

Sahin, E. K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. SN Applied Sciences, v.2, e1308, 2020. https://doi.org/10.1007/s42452-020-3060-1

Sarker, I. H. Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. Annals of Data Science, v.10, p.1473-1498, 2023. https://doi.org/10.1007/s40745-022-00444-2

Shahzad, N.; Ding, X.; Abbas, S. A Comparative Assessment of Machine Learning Models for Landslide Susceptibility Mapping in the Rugged Terrain of Northern Pakistan. Applied Sciences, v.12, e2280, 2022. https://doi.org/10.3390/app12052280

Shyma, K. P.; Gupta, J. P.; Singh, V. Breeding strategies for tick resistance in tropical cattle: A sustainable approach for tick control. Journal of parasitic diseases: official organ of the Indian Society for Parasitology, v.39, p.1-6, 2015. https://doi.org/10.1007/s12639-013-0294-5

Slobe, N.; Catal, C.; Kassahun, A. Application of machine learning to improve dairy farm management: A systematic literature review. Preventive Veterinary Medicine, v.187, e105237, 2021. https://doi.org/10.1016/j.prevetmed.2020.105237

Xu, B.; Mao, Y.; Wang, W.; Chen, G. Intelligent weight prediction of cows based on semantic segmentation and back propagation neural network. Frontiers in Artificial Intelligence, v.7, e1299169, 2024. https://doi.org/10.3389/frai.2024.1299169

Yavuz Ozalp, A.; Akinci, H.; Zeybek, M. Comparative Analysis of Tree-Based Ensemble Learning Algorithms for Landslide Susceptibility Mapping: A Case Study in Rize, Turkey. Water, v.15, e2661, 2023. https://doi.org/10.3390/w15142661

Zúñiga, E. A.; Chambi, S.C.; Carbajal, C.C.; Meléndez, F.A.; Figueroa, I.T.; Viveros, W.Y.; Coaquira, J. Q. La correlación de Pearson o de Spearman en caracteres físicos y textiles de la fibra de alpacas. Revista de Investigaciones Veterinarias del Perú, v.33, e22908, 2022. https://doi.org/10.15381/rivep.v33i3.22908