PAB
Pesquisa Agropecuária Brasileira

Veterinary Science/Original Article

ISSN 1678-3921

Journal homepage: www.embrapa.br/pab

For manuscript submission and journal contents, access: www.scielo.br/pab

Performances of several machine learning algorithms and of logistic regression to predict *Fasciola hepatica* in cattle

Abstract – The objective of this work was to compare the performances of logistic regression and machine learning algorithms to predict infection caused by *Fasciola hepatica* in cattle. A dataset on 30,151 bovines from Uruguay was used. Logistic regression (LR) and the algorithms k-nearest neighbor (KNN), classification and regression trees (CART), and random forest (RF) were compared. The interquartile range (IQR) and z-score were used to improve the classification and compared to each another. Sex, age, carcass conformation score, fat score, productive purpose, and carcass weight were used as independent variables for all algorithms. Infection by *F. hepatica* was used as a binary dependent variable. The accuracies of LR, KNN, CART, and RF were 0.61, 0.57, 0.57, and 0.58, respectively. The variable importance of LR showed that adult cattle tended to be infected by *F. hepatica*. All models showed low accuracy, but LR successfully distinguished variables related to *F. hepatica*. Both the IQR and z-score show similar results in improving the classification metrics for the used dataset. In the dataset, data related to climate or factors such as body weight can improve the reliability of the model in future studies.

Index terms: *Fasciola hepatica*, classification, data mining, fluke, machine learning.

Malik Ergin[1] ✉, and
Özgür Koşkan[1],

[1] Department of Animal Science, Faculty of Agriculture, University of Isparta, Isparta 32000, Türkiye.
E-mail: malikergin@isparta.edu.tr, ozgurkoskan@isparta.edu.tr

✉ Corresponding author

Received
November 01, 2023

Accepted
May 29, 2024

How to cite
ERGIN, M.; KOŞKAN, Ö. Classification performances of several machine learning algorithms and logistic regression for *Fasciola hepatica* in cattle. **Pesquisa Agropecuária Brasileira**, v.59, e03563, 2024. DOI: https://doi.org/10.1590/S1678-3921.pab2024.v59.03563.

Desempenho de vários algoritmos de aprendizado de máquina e regressão logística para prever *Fasciola hepatica* em bovinos

Resumo – O objetivo deste trabalho foi comparar os desempenhos da regressão logística e de algoritmos de aprendizado de máquina para prever infecção por *Fasciola hepatica* em bovinos. Um conjunto de dados de 30.151 bovinos do Uruguai foi usado no estudo. Foram comparados a regressão logística (RL) e os algoritmos *k-nearest neighbor* (KNN), árvores de decisão (CART) e *random forest* (RF). O intervalo interquartil (IQR) e o escore-z foram usados para melhorar a classificação e comparados entre si. Sexo, idade, escore de conformação de carcaça, escore de gordura, propósito produtivo e peso da carcaça foram usados como variáveis independentes para todos os algoritmos. A infecção por *F. hepatica* foi usada como variável dependente binária. Os níveis de precisão de RL, KNN, CART e RF foram 0.61, 0.57, 0.57 e 0.58, respectivamente. A variável importância do modelo de RL mostrou que bovinos adultos tenderam à infecção por *F. hepatica*. Todos os modelos apresentaram baixa precisão, mas a RL distinguiu com sucesso as variáveis relacionadas a *F. hepatica*. Tanto o IQR quanto o escore-z mostram resultados semelhantes quanto à melhoria da métrica de classificação para o conjunto de

dados utilizadoo. No conjunto de dados, dados relacionados ao clima ou a fatores como peso corporal, podem melhorar a confiabilidade do modelo em estudos futuros.

Termos para indexação: *Fasciola hepatica*, classificação, verme trematódeo, aprendizado de máquina, mineração de dados.

Introduction

Livestock diseases – together with associated treatment costs and decreased productivity – causes significant economic and physiological losses to farmers (Yadav et al., 2023).

Recently, *Fasciola hepatica* has been recently reported as occurring in cattle of all continents except for Antarctica (Drescher et al., 2023). More than 70 countries have been affected by this problem (Centers for Disease Control and Prevention, 2016). A review carried out between 2000 and 2015 showed the prevalent countries for fasciolasis in cattle are as follows: 11 ones in Africa; 5, in Asia; 13, in the Americas; 2, in Australia/Oceania; and 11, in Europe (Mehmood et al., 2017).

Fasciolasis is an infectious disease that is devastating in livestocks such as cattle, sheep, goat, horse, rabbit, and camel (Charlier et al., 2014; Howell et al., 2015; Beesly et al., 2018; Costa et al., 2019). It is also known as a type of liver fluke. In cattle, this disease leads to a decrease of body weight, carcass weight, milk yield, and reproductive performance, causing the failure of organs (Rashid et al., 2019).

The infection is transmitted orally via the ingestion of metacercaria. The parasite depends on the presence of suitable intermediate hosts, such as snails. The young parasites penetrate the intestinal wall and progress to the liver. After residing in the bile ducts for a period of time, the adult parasite releases their eggs into the environment through feces. Symptoms manifest 11–12 weeks after infection. The life cycle of *F. hepatica* spans for about 18–24 weeks. (Kaplan, 2001; Urquhart et al., 2002; Balkaya et al., 2010).

Data mining, a.k.a. "knowledge mining from data", have impacted many scientific fields. It has been used in many data sources as historical records (Ahmed, 2016), stock exchange (Patel et al., 2021), time series (Sabu & Kumar, 2020), biological sequence (Liao et al., 2018), sensors (Porto et al., 2015), spatial and geographical data (Ducheyne et al., 2015; Kaya et al., 2023), and social media (Zuliani et al., 2021). Big data provides massive information for machine learning algorithms, to develop predictive models (Zhou et al., 2017).

Advancements in machine learning have made possible to develop automated diagnostic technologies (Cihan et al., 2017). The use of big datasets has enabled the detection of potential diseases in livestock, and the development of early diagnosis approaches through machine learning algorithms (Ghosh & Dasgupta, 2022). Many studies have been published on animal science using such technology. These studies include the characterization of harmful bacteria in livestock (Hermann-Bank et al., 2015), identification of factors affecting pregnancy in cattle (Caraviello et al., 2006), classification of some cattle breeds based on morphological characteristics (Parés Casanova et al., 2012), detection of mastitis (Tanyıldızı & Yıldırım, 2019; Altay & Delialioğlu, 2022), and the use of various algorithms to improve the accuracy of detecting bovine bluetongue disease (Gouda et al., 2022).

The objective of this work was to compare the performances of logistic regression and machine learning algorithms to predict infection caused by *F. hepatica* in cattle.

Materials and Methods

The open access data released by Corbellini et al. (2019) were used. This dataset consists of 30,151 rows, each one representing the individual record of bovines from a slaughterhouse in Uruguay. The columns of the dataset contain different variables such as date, sex, animal age (five categories), carcass weight (CW), carcass conformation code, carcass conformation score (CCS), fat score (FS) (fat coverage), productive purpose (PP), *F. hepatica* status, farm code, and department number. The variables date, carcass conformation code, farm code, and department number were not used in the present study. Detailed information on the variables employed in the present work is given (Table 1).

The statistical importance of the variables was analyzed using the chi-square test. All independent variables were significantly dependent on the binary dependent variable, at 5% probability. In data preparation, two different outlier detection techniques were performed, to determine observations which

seemed inconsistent with the remainder of the dataset. Z-score transformation positively contribute to increase the accuracy of the classification models (Karo & Hendriyana, 2022). Therefore, z-score transformation was used to eliminate outliers in the dependent variable, according to following equation:

$$Z = x - u/s \quad (1)$$

where: x is the current sample value; u is the overall mean of the sample; and s is the overall standard deviation of the sample.

Values out of the -3 < Z < +3 interval were considered outliers. The values of a total of 30,151 sample values were filtered down to 29,986, after the z-score transformation.

Quartiles are especially used, to avoid the effects of variation caused by outliers in the dataset (Sokal & Rohlf, 1969). In the present study, the first (Q_1) and third (Q_3) quartiles were used to detect outliers. Equations 2 and 3 give the values of the first and third quartiles, respectively.

$$Q_1 = (n+1)/4 \quad (2)$$

$$Q_3 = 3 \times (n+1)/4 \quad (3)$$

The interquartile range (IQR) value indicates the range where 50% of the data changes and can be calculated using the equation (4).

$$IQR = Q_3 - Q_1 \quad (4)$$

Outliers are then detected by equations 5 and 6.

$$x = Q_1 - 1.5 \times IQ \quad (5)$$

$$y = Q_3 + 1.5 \times IQ \quad (6)$$

Values smaller than x or larger than y, can be considered outliers. After outlier detection, 29,873 observations remained.

The total dataset was randomly separated into two sets: training (85%) and testing (15%). In literature, the ratios of the training and testing sets are usually 70% and 30%, respectively. In order to determine the accuracy of the model, the training set was chosen to be kept as high as possible. Therefore, instead of 70% of the dataset, 85% were selected for training. Five replicates and five-fold repeated cross validation

Table 1. Descriptive statistics of the independent variables according to the binary dependent variable.

Categorical variable	Category	Negative for *F.hepatica*		Positive for *F.hepatica*	
		n	(%)	n	(%)
Sex	Female	9,143	47.0	7,310	68.4
	Male	10,322	53.0	3,376	31.6
Age	0 (0–23 months)	598	3.1	146	1.4
	2 (23–30 months)	3,159	16.2	666	6.2
	4 (30–37 months)	3,471	17.8	1,020	9.5
	6 (37–42 months)	2,691	13.8	1,268	11.9
	8 (>42 months)	9,546	49.0	7,586	71.0
Carcass conformation score (CCS)	Low quality	641	3.3	755	7.1
	Regular and good quality	16,008	82.2	9,050	84.7
	Excellent quality	2,816	14.5	881	8.2
Fat score (FS) (Fat coverage)	Very low	1,047	5.4	1,013	9.5
	Low	4,655	23.9	2,079	19.5
	Regular	13,106	67.3	7,429	69.5
	Excessive	657	3.4	165	1.5
Productive purposes (PP)	Milk	5,050	25.9	2,251	21.1
	Beef (meat)	8,193	42.1	4,646	43.5
	Cross-breed	6,222	32.0	3,789	35.5

Continuous variable	Negative for *F. hepatica*				Positive for *F. hepatica*			
	n	Min	Max	$X \pm S_X$*	n	Min	Max	$X \pm S_X$*
Carcass weight (CW)	19,465	81.5	601.6	258.6±0.32a	10,686	74.4	560.4	247.1±0.44b

*Significant at 5% probability.

were performed on the training dataset to tune hyperparameters. Detailed information on the tuning parameters was presented (Table 2).

Due to the imbalance of the dataset for class variable, an undersampling strategy was performed. The *downSample* function in the *caret* package of the statistical software R was used for the undersampling of the classes. When the training dataset was split as 85% of the total dataset, the expected value of the dependent variable was determined to be "0", for 16,453 observations, and it was determined to be "1" for 9,035 observations. Undersampling was applied to the training set by undersampling the majority class, without replacement in the class attribute.

Once data were properly adjusted, the models were built. The first model was the logistic regression. A logistic regression estimates the probability of a binary categorical dependent variable to be "1" (Eyduran, 2005; Hosmer et al., 2013; Altay et al., 2019). The mathematical model of the logistic regression (logit model) is described in equation 7:

$$P(Y=1|X_i) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}} \quad (7)$$

where: X is the independent variable probability of *F. hepatica*, when it receives the x value; P(Y=1) is the probability of the occurrence of Y = 1, when X = x; β_0 is the constant of the regression; and e is the natural logarithm (approximately 2.718).

The effects of the independent variables on the binary dependent variable can be explained by the logistic regression analysis using this model.

After setting up the logistic model, the first algorithm implemented was the k-nearest neighbor (KNN). This algorithm is a supervised one, well-known for its simplicity. It can be used to predict a target value. Unlike traditional algorithms, it does not define a model, but it represents each observation as a standard Euclidean distance within an n-dimensional space (instance). When new instances are considered, the KNN algorithm calculates the distance belonging to each training instance (Mitchel, 1997; Uğuz, 2019). In the KNN algorithm, K is the only parameter that determines how many neighbors will be evaluated, to decide the classification of a new observation.

The second implemented algorithm was the classification and regression tree (CART); developed by Breiman et al. (1984), CART is a tree-based algorithm, and it is not presented in a mathematical form. One of the most significant advantages of the CART algorithm is that it does not require assumptions of normality, homogeneity of variances, and independency of observations, which are assumptions for multiple regression. The second important advantage is that during the tree construction stage, statistically insignificant independent variables are excluded from the tree diagram (Kayri & Boysan, 2008; Coşkun et al., 2023).

The third algorithm used was the random forest (RF). The RF algorithm creates different decision trees by subsampling different observations in the dataset. This prevents overfitting and provides a greater accuracy than a single decision tree such as CART (Breiman, 2001). The RF is a robust algorithm for overfitting, in comparison with other machine learning algorithms. It enables to working with as many independent decision trees as desired and it is quite fast because it does not perform any pruning (Breiman & Cuttler, 2005). In the present study, the criteria used to choose the branch in each node is the Gini index. The Gini index measures the homogeneity ("purity") of randomly selected variables that form the best branches among all variables, that is, the probability of misclassification. (Akar & Gungor, 2012; Daniya et al., 2020; Tangirala, 2020). The Gini index is calculated using equation (8).

$$\text{Gini index}(L) = 1 - \sum_{i=1}^{j} p_i^2 \quad (8)$$

where: L states a dataset containing j different classes ("0" and "1" for the present study); express the relative frequency or the probability of an object being classified into a particular class.

When the Gini index decreases, the homogeneity of the class increases. For a branch to be selected as the best one, the Gini index of a child node should be lower than that of its parent node. To terminate the tree, the Gini index should reach zero, and branching

Table 2. Tuning parameters used in algorithms.

Algorithms	Tuning parameters (z-score)	Tuning parameters (IQR)
KNN	k = 13	k = 13
CART (DT)	cp = 0.001881363	cp = 0.001583685
RF	mtry = 2 ntree = 500	mtry = 2 ntree = 500
LR	No need to tune	No need to tune

stops when each child node contains only one class ("0" or "1").

After setting the LR and the algorithms, five metrics were chosen to compare them. The performance metrics used for classification were the confusion matrix, accuracy, sensitivity, precision, and F1 score (Dişçi, 2012). The confusion matrix illustrates how many samples are in the right or wrong classes. The samples were classified as true positive (TP) or true negative (TN), when the number of samples are successfully classified as "1" (presence of *F. hepatica*) or "0" (absence of *F. hepatica*), respectively; and false positive (FP) or false negative (FN), when the samples were unsuccessfully classified as "1" or "0", respectively.

Accuracy is a percentage of correctly classified individuals within the total predictions made by the machine learning algorithm. Equation (9) represents the accuracy formula.

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \quad (9)$$

Sensitivity indicates the percentage of correctly predicted animals using the machine learning algorithm in all true positive cases. The formula for sensitivity is given in equation (10).

$$\text{Sensitivity} = \text{TP}/\text{TP} + \text{FN} \quad (10)$$

Precision measures the performance of classification algorithms as the proportion of true positive predictions among all cases predicted as positive. Equation (11) presents the formula for precision.

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \quad (11)$$

The F1 score measures the accuracy of the test. It is useful for unbalanced datasets (Vujović, 2021); its calculation is presented in equation (12).

$$\text{F1 score} = 2 \times \text{TP}/2 \times \text{TP} + \text{FN} + \text{FP} \quad (12)$$

All analyses were performed using the R programing language (version 4.3.1). The *caret* package (version 6.0.94), which consists of several machine learning functions, was used to build predictive models (Kuhn, 2008). The *glm*, *knn*, *rpart*, and *rf* parameters were used for building logistic regression, k-nearest neighbor, CART, and random forest, respectively. The variable importance of the algorithms was visualized using the *varImp* function. The variable importance considers certain coefficients to determine the relationship between the dependent and independent variables. In linear regression algorithms, each independent variable is ranked according to correlation coefficients, to detect which independent variable is more important for the prediction of the dependent variable. Hence, it serves for the dimensionality reduction and for the feature selection that improve the predictive performance of the model. Selecting the most important independent variables that explain a large portion of the variance of the dependent variable can be crucial for identifying and building high predictive performance models.

Results and Discussion

In the training set subjected to the z-score outlier detection (Table 3), the accuracy scores of LR, KNN, CART, and RF were calculated as 0.61, 0.66, 0.62, and 0.63, respectively. The KNN algorithm achieved the highest accuracy (66%). The accuracy scores of LR, KNN, CART, and RF were slightly decreased in the testing set and were calculated as 0.61, 0.57, 0.57, and 0.58, respectively. In the testing phase, the LR model showed the best results for accuracy (0.61). Considering the TP values, LR, KNN, CART, and RF classified 1098, 1046, 1299, and 1287 samples, respectively. The number of FNs in the LR, KNN, CART, and RF were 496, 548, 295, and 307, respectively. In machine learning algorithms, the sensitivity (S) values ranged between 0.65 and 0.81 in the testing set. The CART and RF algorithms showed higher sensitivity, indicating that these algorithms were successful in correctly detecting fasciolasis in cattle. The F1 scores (F1) of the testing set for LR, KNN, CART, and RF were 0.56, 0.52, 0.58, and 0.58, respectively. Except for the KNN, the F1 scores are quite similar. Models with high precision values also yield relatively high F1 scores.

The performance metrics of the training and testing sets subjected to IQR outlier detection are presented (Table 4). In the training set, the accuracy score of LR, KNN, CART, and RF algorithms were 0.62, 0.65, 0.62, and 0.63, respectively. The accuracy results were almost the same. A similar pattern of scores was observed for sensitivity and precision. The sensitivity scores were 0.66, 0.73, 0.80, and 0.80 for LR, KNN, CART, and RF, respectively. The precision scores were 0.60, 0.63, 0.59, and 0.59, respectively. In the testing set, there was no remarkable differences for the outlier detection

Table 3. Results of accuracy (AC), confusion matrix (CM), sensitivity (S), precision (P), and F1 score (F1) of the training and testing sets subjected to z-score outlier detection of the logistic regression (LR) and the algorithms k-nearest neighbor (KNN), classification and regression tree (CART), and random forest (RF).

Algorithms	Training set						Testing set					
	AC	CM		S	P	F1	AC	CM		S	P	F1
LR	0.61	5,976	3,931	0.66	0.60	0.63	0.61	1,098	1,226	0.68	0.47	0.56
		3,060	5,105					496	1,677			
KNN	0.66	6,476	3,553	0.71	0.64	0.68	0.57	1,046	1,365	0.65	0.43	0.52
		2,560	5,483					548	1,583			
CART (DT)	0.62	7,281	5,104	0.80	0.58	0.68	0.57	1,299	1,605	0.81	0.44	0.58
		1,755	3,932					295	1,298			
RF	0.63	7,218	4,915	0.80	0.60	0.68	0.58	1,287	1,560	0.81	0.45	0.58
		1,818	4,121					307	1,343			

Table 4. Results of the training and testing datasets by IQR outlier detection.

Algorithm[(1)]	Train dataset[(2)]						Test dataset[(2)]					
	AC	CM		S	P	F1	AC	CM		S	P	F1
LR	0.62	5,952	3,870	0.66	0.60	0.64	0.61	1,068	1,230	0.67	0.46	0.55
		3,046	5,128					519	1,663			
KNN	0.65	6,579	3,893	0.73	0.63	0.68	0.57	1,088	1,441	0.68	0.43	0.53
		2,419	5,105					499	1,452			
CART (DT)	0.62	7,206	5,009	0.80	0.59	0.68	0.57	1,273	1,594	0.80	0.44	0.57
		1,792	3,989					314	1,299			
RF	0.63	7,198	4,903	0.80	0.59	0.68	0.58	1,265	1,566	0.80	0.45	0.57
		1,800	4,095					322	1,327			

[(1)]LR, logistic regression; KNN, k-nearest neighbor; CART (DT), classification and regression tree; RF, random forest. [(2)]AC, accuracy; CM, confusion matrix; S, sensitivity; P, precision; F1, F1 score.

method used. In the testing set, the F1 scores were 0.55, 0.53, 0.57, and 0.57 for LR, KNN, CART, and RF, respectively. These results suggest that there were no significant differences between LR and the algorithms.

These results also suggest that LR has the highest performance, in comparison with any machine learning algorithms performed in the present study. In the literature, there is no study on the prediction or classification of *Fasciola hepatica* in cattle, using machine learning algorithms. The RF and CART algorithms are tree-based algorithms; therefore, remarkably similar performances and variable importance of the RF and CART are acceptable.

Conclusions

1. According to the bovine carcass dataset from Uruguay, the LR algorithm slightly outperforms for accuracy, sensitivity, precision, and F1 score.

2. For outlier detection, the IQR and z-score techniques give quite similar results and do not show any remarkable effects to improve classification metrics for this dataset.

Acknowledgements

To Luis Corbellini, Ricardo Almeida da Costa, Franklin Riet-Correa, and Eleonor Castro-Janer, for sharing their data on Mendeley Data, making it open and accessible.

References

AHMED, T.M. Using data mining to develop model for classifying diabetic patient control level based on historical medical records. **Journal of Theoretical and Applied Information Technology**, v.87, p.316-323, 2016.

AKAR, Ö.; GÜNGÖR, O. Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. **Jeodezi ve Jeoinformasyon Dergisi**, v.106, p.139-146, 2012. DOI: https://doi.org/10.9733/jgg.241212.1t.

ALTAY, Y.; DELİALİOĞLU, R.A. Diagnosing lameness with the Random Forest classification algorithm using thermal cameras and digital colour parameters. **Mediterranean Agricultural Sciences**, v.35, p.47-54, 2022. DOI: https://doi.org/10.29136/mediterranean.1065527.

ALTAY, Y.; KILIÇ, B.; AYTEKİN, İ.; KESKİN, İ. Determination of factors affecting mastitis in Holstein Friesian and Brown Swiss by using logistic regression analysis. **Selcuk Journal of Agriculture and Food Sciences**, v.33, p.194-197, 2019. DOI: https://doi.org/10.15316/SJAFS.2019.175.

BALKAYA, I.; KAPAKIN, K.A.T.; ATASEVER, I. Morphological and histopathological examination of bovine livers naturally infected with *Fasciola hepatica*. **Veterinary Sciences and Practices**, v.5, p.7-11, 2010.

BEESLY N.J.; CAMINADE, C.; CHARLIER, J.; FLYNN, R.J.; HODGKINSON, J.E.; MARTINEZ-MORENO, A.; MARTINEZ-VALLADARES, M.; PEREZ, J.; RINALDI, L.; WILLIAMS, D.J.L. *Fasciola* and fasciolosis in ruminants in Europe: identifying research needs. **Transboundary and Emerging Diseases**, v.65, p.199-216, 2018. DOI: https://doi.org/10.1111/tbed.12682.

BREIMAN, L. Random forests. **Machine Learning**, v.5, p.5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

BREIMAN, L.; CUTLER, A. **Random forests**. 2005. Available at <http://www.stat.berkeley.edu/~breiman/RandomForests/>. Accessed on: Apr. 2 2024.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. **Classification and regression trees**. Boca Raton: Chapman and Hall, 1984. 368p. DOI: https://doi.org/10.1201/9781315139470.

CARAVIELLO, D.Z.; WEIGEL, K.A.; CRAVEN, M.; GIANOLA, D.; COOK, N.B.; NORDLUND, K.V.; FRICKE, P.M.; WILTBANK, M.C. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. **Journal of Dairy Science** v.89, p.4703-4722, 2006. DOI: https://doi.org/10.3168/jds.S0022-0302(06)72521-8.

CENTERS FOR DISEASE CONTROL AND PREVENTION. **Fasciola biology**. 2016. Available at <www.cdc.gov/parasites/fasciola/biology.html>. Accessed on: Apr. 2 2024.

CHARLIER, J.; VERCRUYSSE, J.; MORGAN, E.; VAN DIJK, J.; WILLIAMS, D.J.L. Recent advances in the diagnosis, impact on production and prediction of *Fasciola hepatica* in cattle. **Parasitology**, v.141, p.326-335, 2014. DOI: https://doi.org/10.1017/S0031182013001662.

CIHAN, P.; GOKCE, E.; KALIPSIZ, O. A review of machine learning applications in veterinary field. **Journal of the Faculty of Veterinary Medicine, Kafkas University**, v.23, p.673-680, 2017. DOI: https://doi.org/10.9775/kvfd.2016.17281.

CORBELLINI, L.; COSTA, R.A.; FRANKLIN, R.C.; ELEONOR, C.J. Bovine carcasses Uruguay. **Mendeley Data**, V3, 2019. DOI: https://doi.org/10.17632/3jnn876my4.3.

COSTA, R.A. da; CORBELLINI, L.G.; CASTRO-JANER, E.; RIET-CORREA, F. Evaluation of losses in carcasses of cattle naturally infected with *Fasciola hepatica*: effects on weight by age range and on carcass quality parameters. **International Journal for Parasitology**, v.49, p.867-872, 2019. DOI: https://doi.org/10.1016/j.ijpara.2019.06.005.

COŞKUN, G.; ŞAHIN, Ö.; DELIALIOĞLU, R.A.; ALTAY, Y.; AYTEKIN, I. Diagnosis of lameness via data mining algorithm by using thermal camera and image processing method in Brown Swiss cows. **Tropical Animal Health and Production**, v.55, art.50, 2023. DOI: https://doi.org/10.1007/s11250-023-03468-9.

DANIYA, T.; GEETHA, M.; KUMAR, K.S. Classification and regression trees with Gini index. **Advances in Mathematics**: **Scientific Journal**, v.9, p.8237-8247, 2020. DOI: https://doi.org/10.37418/amsj.9.10.53.

DİŞÇİ, R. **Basic and clinical biostatistics**. 2nd ed. Istanbul: Istanbul Medicine Bookstore, 2012. 313p.

DRESCHER, G.; VASCONCELOS, T.C.B. de; BELO, V.S.; PINTO, M.M. da G.; ROSA, J. de O.; MORELLO, L.G.; FİGUEIREDO, F.B. Serological diagnosis of fasciolosis (*Fasciola hepatica*) in humans, cattle, and sheep: A meta-analysis. **Frontiers in Veterinary Science**, v.10, art.1252454, 2023. DOI: https://doi.org/10.3389/fvets.2023.1252454.

DUCHEYNE, E.; CHARLIER, J.; VERCRUYSSE, J.; RINALDI, L.; BIGGERI, A.; DEMELER, J.; BRANDT, C.; WAAL, T. de; SELEMETAS, N.; HÖGLUND, J.; KABA, J.; KOWALCZYK, S.J.; HENDRICKX, G. Modelling the spatial distribution of *Fasciola hepatica* in dairy cattle in Europe. **Geospatial Health**, v.9, p.261-270, 2015. DOI: https://doi.org/10.4081/gh.2015.348.

EYDURAN, E.; OZDEMIR, T.; ÇAK, B.; ALARSLAN, E. Using of logistic regression in animal science. **Journal of Applied Sciences**, v.5, p.1753-1756, 2005. DOI: https://doi.org/10.3923/jas.2005.1753.1756.

GHOSH, S.; DASGUPTA, R. **Machine learning in biological sciences**: updates and future prospects. Singapore: Springer Nature, 2022. 336p. DOI: https://doi.org/10.1007/978-981-16-8881-2.

GOUDA, H.F.; HASSAN, F.A.; EL-ARABY, E.E.; MOAWED, S.A. Comparison of machine learning models for bluetongue risk prediction: a seroprevalence study on small ruminants. **BMC Veterinary Research**, v.18, art.394, 2022. DOI: https://doi.org/10.1186/s12917-022-03486-z.

HERMANN-BANK, M.L.; SKOVGAARD, K.; STOCKMARR, A.; STRUBE, M.L.; LARSEN, N.; KONGSTED, H.; INGERSLEV, H.-C.; MOLBAK, L.; BOYE, M. Characterization of the bacterial gut microbiota of piglets suffering from new neonatal porcine diarrhoea. **BMC Veterinary Research**, v.11, art.139, 2015. DOI: https://doi.org/10.1186/s12917-015-0419-4.

HOSMER JR, D.W.; LEMESHOW, S.; STURDIVANT, R.X. **Applied logistic regression**. Hoboken: J. Wiley & Sons, 2013. 518p. DOI: https://doi.org/10.1002/9781118548387.

HOWELL, A.; BAYLIS, M.; SMITH, R.; PINCHBECK, G.; WILLIAMS, D. Epidemiology and impact of *Fasciola hepatica* exposure in high-yielding dairy herds. **Preventive Veterinary Medicine**, v.121, p.41-48, 2015. DOI: https://doi.org/10.1016/j.prevetmed.2015.05.013.

KAPLAN, R.M. *Fasciola hepatica*: a review of the economic impact in cattle and considerations for control. **Veterinary Therapeutics**, v.2, p.40-50, 2001.

KARO, I.M.K.; HENDRIYANA. Klasifikasi penderita diabetes menggunakan algoritma machine learning dan z-score. **Jurnal Teknologi Terpadu**, v.8, p.94-99, 2022. DOI: https://doi.org/10.54914/jtt.v8i2.564.

KAYA, F.; MISHRA, G.; FRANCAVIGLIA, R.; KESHAVARZI, A. Combining digital covariates and machine learning models to predict the spatial variation of soil cation exchange capacity. **Land**, v.12, art.819, 2023. DOI: https://doi.org/10.3390/land12040819.

KAYRI, M.; BOYSAN, M. Assesment of relation between cognitive vulnerability and depression's level by using classification and regression tree analysis. **Hacettepe University Journal of Education**, v.34, p.168-177, 2008.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v.28, p.1-26, 2008. DOI: https://doi.org/10.18637/jss.v028.i05.

LIAO, Z.; LI, D.; WANG, X.; LI, L.; ZOU, Q. Cancer diagnosis through isomiR expression with machine learning method. **Current Bioinformatics**, v.13, p.57-63, 2018. DOI: https://doi.org/10.2174/1574893611666160609081155.

MEHMOOD, K.; ZHANG, H.; SABIR, A.J.; ABBAS, R.Z.; IJAZ, M.; DURRANI, A.Z.; SALEEM, M.H.; REHMAN, U.M.; IQBAL, M.K.; WANG, Y.; AHMAD, H.I.; ABBAS, T.; HUSSAIN, R.; GHORI, M.T.; ALI, S.; KHAN, A.U.; LI, J. A review on epidemiology, global prevalence and economical losses of fasciolosis in ruminants. **Microbial Pathogenesis**, v.109, p.253-262, 2017. DOI: https://doi.org/10.1016/j.micpath.2017.06.006.

MITCHEL, T.M. **Machine learning**. [S.l.]: McGraw-Hil, 1997. 432p.

PARÉS CASANOVA, P.-M.; SINFREU BASI, I.; VILLALBA MATA, D. Principal component analysis of cephalic morphology to classify some Pyrenean cattle. **Animal Genetic Resources**, v.50, p.59-64, 2012. DOI: https://doi.org/10.1017/S2078633611000385.

PATEL, S.K.; PRAJAPATI, J.B.; PATEL, H.R. A study on developing effective option trading strategy on NIFTY index in national stock exchange using data mining. In: INTERNATIONAL CONFERENCE ON CLOUD COMPUTING, DATA SCIENCE & ENGINEERING, 11., Noida, 2021. **Proceedings**. Piscataway: IEEE, 2021. p.298-303. Confluence 2021. DOI: https://doi.org/10.1109/Confluence51648.2021.9377179.

PORTO, S.M.; ARCIDIACONO, C.; ANGUZZA, U.; CASCONE, G. The automatic detection of dairy cow feeding and standing behaviours in free-stall barns by a computer vision-based system. **Biosystems Engineering**, v.133, p.46-55, 2015. DOI: https://doi.org/10.1016/j.biosystemseng.2015.02.012.

RASHID, M.; RASHID, M.I.; AKBAR, H.; AHMAD, L.; HASSAN, M.A.; ASHRAF, K.; SAEED, K.; GHARBI, M. A systematic review on modelling approaches for economic losses studies caused by parasites and their associated diseases in cattle. **Parasitology**, v.146, p.129-141, 2019. DOI: https://doi.org/10.1017/S0031182018001282.

SABU, K.M.; KUMAR, T.K.M. Predictive analytics in agriculture: forecasting prices of Arecanuts in Kerala. **Procedia Computer Science**, v.171, p.699-708, 2020. DOI: https://doi.org/10.1016/j.procs.2020.04.076.

SOKAL, R.R.; ROHLF, F.J. **Biometry**: the principles and practices of statistics in biological research. New York: W.H. Freeman and Company, 1969. 915p.

TANGIRALA, S. Evaluating the impact of Gını index and information gain on classification using decision tree classifier algorithm. **International Journal of Advanced Computer Science and Applications**, v.11, p.612-619, 2020. DOI: https://doi.org/10.14569/ijacsa.2020.0110277.

TANYILDIZI, E.; YILDIRIM, G. Performance comparison of classification algorithms for the diagnosis of mastitis disease in dairy animals. In: 7th INTERNATIONAL SYMPOSIUM ON DIGITAL FORENSICS AND SECURITY, 7., 2019, Barcelos. P**roceedings**. Piscataway: IEEE, 2019. p.1-4. DOI: https://doi.org/10.1109/ISDFS.2019.8757469.

UĞUZ, S. **Makine öğrenmesi - Teorik yönleri ve Python uygulamaları ile bir yapay zekâ ekolü**. Isparta: Nobel Akademik Yayıncılık, 2019. 312p.

URQUHART, G.M.; ARMOUR, J.; DUNCAN, J.L.; DUNN, A.M.; JENNINGS, F.W. **Veterinary Parasitology**. 2nd ed. New Jersey: Blackwell Publishing, 2002. 300p.

VUJOVIĆ, Ž. Đ. Classification model evaluation metrics. **International Journal of Advanced Computer Science and Applications**, v.12, p.599-606, 2021. DOI: https://doi.org/10.14569/IJACSA.2021.0120670.

YADAV, A.K.; VERMA, D.; SOLANKI, P.R. Introduction to numerous diseases of the livestock. In: SINGH, R.P.; ADETUNJI, C.O.; SINGH, R.L.; SINGH, J.; SOLANKI, P.R.; SINGH, K.R.B. (Ed.). **Nanobiotechnology for the livestock industry**. Amsterdam: Elsevier, 2023. p.141-156. DOI: https://doi.org/10.1016/B978-0-323-98387-7.00020-3.

ZHOU, L.; PAN, S.; WANG, J.; VASILAKOS, A.V. Machine learning on big data: opportunities and challenges. **Neurocomputing**, v.237, p.350-361, 2017. DOI: https://doi.org/10.1016/j.neucom.2017.01.026.

ZULIANI, A.; CONTIERO, B.; SCHNEIDER, M.K.; ARSENOS, G.; BERNUÉS, A.; DOVC, P.; GAULY, M.; HOLAND, Ø.; MARTIN, B.; MORGAN-DAVIES, C.; ZOLLITSCH, W.; COZZI, G. Topics and trends in mountain livestock farming research: a text mining approach. **Animal**, v.15, art.100058, 2021. DOI: https://doi.org/10.1016/j.animal.2020.100058.