Enfoque UTE, V.14-N.2, Abr. 2023, pp. 20-34
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542
CC BY 3.0 EC

Received: 2022/10/22
Accepted: 2023/01/25
Online: 2023/01/25
Published: 2023/04/01
https://doi.org/10.29019/enfoqueute.861

Adaptability of regression algorithms to the behavior of protein plants

(Adaptabilidad de algoritmos de regresión al comportamiento de las plantas proteicas)

Pedro M. Estrada-Jiménez[1]; Hernán A. Uvidia-Cabadiana[2]; Rocío del Carmen Herrera-Herrera[3], Luís G. Hernández-Montiel[4]; Danis M. Verdecia-Acosta[5]; Jorge L. Ramírez-de la Ribera[6]; Pedro J. Noguera-López[7]; Edilberto Chacón-Marcheco[8]

Abstract

The behavior of components of protein plant is of vital importance for animals that consume them in their diet. The objective of this research is to evaluate regression algorithms, to determine the behavior of the expressions that best adapt to the procedures of a traditional laboratory and to estimate the chemical components of protein plants, in this sense the MULAN library of java has been used, that contain automatic learning algorithms capable of adapting to dissimilar problems. Three data set were created for each species treated in this study; each of these include the main elements to be evaluate in each experiment, these are delimitings by: secondary metabolites, cell wall components and digestibility element for training files one, two and three, respectively; subsequently, they were evaluated through learning supervised and cross-validation of each to determine the best fit by aRMSE (Average Root Mean Square Error). The learning results were compare with previous experiments, where there was a learning variant that contained in a single dataset all the components to be evaluates in a single prediction. The result of the comparison shows that the lazy algorithms based on instances have a better learning behavior than the others evaluate.

Keywords

Secondary metabolites; regression models, cell wall; nutritional value

Resumen

El comportamiento de los componentes de las plantas proteicas es de vital importancia para los animales que los consumen en su dieta. La presente investigación tiene como objetivo evaluar algoritmos de regresión para determinar la conducta de las expresiones que mejor se adaptan a los procedimientos de un laboratorio tradicional y estimar los componentes químicos de plantas proteicas, en este sentido, se ha utilizado la biblioteca MULAN de java, que contiene algoritmos de aprendizaje automático capaces de adaptarse a disímiles problemas. Para ello, se crearon tres conjuntos de datos para cada especie estudiada en este trabajo; cada uno de estos incluye los elementos principales a ser evaluados en cada experimento, que están delimitados por: Metabolitos secundarios, componentes de la pared celular y digestibilidad para los ficheros de entrenamiento uno, dos y tres, respectivamente. Posteriormente, fueron evaluados por medio del aprendizaje supervisado y una validación cruzada de cada uno para determinar el mejor ajuste por aRMSE (Error cuadrático medio de la raíz). Los resultados del aprendizaje fueron comparados con experimentos anteriores, donde se tenía una variante de aprendizaje que contenía en un solo dataset todos los componentes a evaluar en una sola predicción. El resultado de la comparación muestra que los algoritmos vagos basados en instancias tienen un mejor comportamiento en el aprendizaje que los otros evaluados.

Palabras clave

Metabolitos secundarios; modelos de regresión, pared celular; valor nutritivo.

[1] Universidad de Granma, Cuba [pestrada@udg.co.cu, https://orcid.org/0000-0001-8759-9500]
[2] Universidad Estatal Amazónica, Ecuador [huvidia@uea.edu.ec, https://orcid.org/0000-0002-2961-6963]
[3] Universidad Nacional de Loja, Ecuador [rocio.herrera@unl.ec, https://orcid.org/0000-0002-4136-4746]
[4] Centro de Investigaciones Biológicas del Noroeste, Baja California Sur, México [lhernandez@cibnor.mx, https://orcid.org/0000-0002-8236-1074]
[5] Universidad de Granma, Cuba [dverdeciaacosta@gmail.com, https://orcid.org/0000-0002-4505-4438]
[6] Universidad de Granma, Cuba [jramirezrivera1971@gmail.com, https://orcid.org/0000-0002-0956-0245]
[7] Universidad de Granma, Cuba [pedrojnl88@gmail.com, https://orcid.org/0000-0002-0966-5266]
[8] Universidad Técnica de Cotopaxi, Ecuador. [edilberto.chacon@utc.edu.ec, https://orcid.org/0000-0001-9590-6451]

1. Introduction

The study of the nutrient components of plants used for animal feed has gained importance in scientific research with the aim of improving nutrition in both ruminants and non-ruminants. The different applications of artificial intelligence in the different fields of life and science are an advance in research. Solutions in medicine with contributions in most specialties can be mentioned, as in the case of application development in the field of imaging where there is software that makes use of pattern recognition to detect pathological anomalies (Verdecia et al., 2018).

The protein plants used in animal nutrition are of great importance for the livestock community as a substitute for concentrates that are becoming more expensive every day. Within these the so-called of excellence reach more relevance every day among farmers for their properties; that is why it is necessary to know the behavior of its components in order to form a quality diet (Díaz et al., 2007; Otegui & Totaro, 2007; Alonso-Peña, 2011; Verdecia et al., 2018). Hence, it is important that several authors dedicate time and resources to the investigation of the behavior of metabolites, components of the cell wall and digestibility of these plants used in livestock (Rincón-Tuexi et al., 2006; Ramírez-Lozano, 2010; T. Ruiz et al., 2011; T. E. Ruiz et al., 2014).

Agriculture is currently committed to the so-called efficient agriculture, which is the one that is equipped with research and applications in the field of artificial intelligence to improve yields. In the present research, the lazy algorithms are analyzing with the learning bases of four plant varieties to determine which of these algorithms is better adapted when simulating the laboratory results in the determination of secondary metabolites, cell wall components and quality components of the species under study (Herrera et al., 2017).

Meat and milk production in ruminants is conditions using forage plants in their diet. In the tropics, the use of leguminous has increased in search of better production indicators, as well as other feeding alternatives, obtaining indicators similar to conventional systems in several cases (Mahecha & Rosales, 2005; Mahecha et al., 2007). Forage plants, beyond being one of the main and excellent components in ruminant nutrition, offer various advantages, among which it is worth noting that they prevent soil erosion, maintain humidity, and provide organic matter; *Gliricidia sepium Erythrina variegata, Leucaena leucocephala* and *Tithonia diversifolia* are among those preferred and used in the tropics (Cabrera, 2008).

The aim of the present research is based on the prediction of the phytochemical components, cell wall components and digestibility of four varieties of protein plants. For this, the adaptability of the multiple regression algorithms to the database provided by the specialists in pastures and forages of the University of Granma has been determined as the main problem.

In previous research, many regressive algorithms have been tested in order to evaluate their behavior with the databases obtained. The result of these analyses has shown that lazy algorithms are the ones that best adapt to these data (Barrios et al., 2015).

The present research studies the lazy algorithms present in the MULAN (Tsoumakas et al., 2011) library developed by the University of Waikato. In this, the aRMSE (Average Root Mean Square Error) is evaluate as the main performance measure to determine the one that best suits the database. The peculiarity of these types of algorithms is that since they work with little data, they are then base on the probability that an object may resemble another to estimate or predict a value. Hence, the objective of this research is to evaluate regression algorithms, to determine the behavior of the expressions that best adapt to the procedures of a traditional laboratory and to estimate the chemical components of protein plants, in this sense the MULAN library of java has been used, that contain automatic learning algorithms capable of adapting to dissimilar problems.

2. Metodology

Regression and Classification Task

The one a most important problem into the Machine Learning is define a type of solutions. Is necessary to have a count the types of variables or types of the data into the data set (Alzubi et al., 2018; Coraddu et al., 2016; González, 2015). Thus is very important to define the types of machine learning tasks. To give solutions to the problem firstly we define a classification and regression:

- Classification is the task of predicting a discrete class label.
- Regression is the task of predicting a continuous quantity.

There is some overlap between the algorithms for classification and regression, for example:

- A classification algorithm may predict a continuous value, but the continuous value is in the form of a probability for a class label.
- A regression algorithm may predict a discrete value, but the discrete value in the form of an integer quantity.

Some algorithms can be used for both classification and regression with small modifications, such as decision trees and artificial neural networks (Alebele et al., 2020), (Mastelini et al., 2020). Some algorithms cannot, or cannot easily be used for both problem types, such as linear regression for regression predictive modeling and logistic regression for classification predictive modeling. Importantly, the way that we evaluate classification and regression predictions varies and does not overlap, for example:

- Classification predictions can be evaluated using accuracy, whereas regression predictions cannot.
- Regression predictions can be evaluated using root mean squared error, whereas classification predictions cannot.

Multi-Target Regression task

In Machine Learning to predict a vector of values of any task, first it is need give to the model a dataset with all examples to system to create with this a system, it is composing to three steps, training, evaluation the training task to define the quality of the model, later to test a model given a vector of real values to obtain a vector of real values that affect a result of prediction (Džeroski et al., 2000; Despotovic et al., 2016; Waegeman et al., 2019; Chen et al., 2021).

The learning process are realized using a learning algorithm. They algorithms are capable to the learn to the dataset to return a vector result. There are many algorithms, it are classify in based on rules, based on decision tree, lazy, vector regression, etc, each one with its specific characteristics (Nogueira & Koch, 2019).

At present, the problems solved by means of regression have reached high levels of applicability. In various life scenarios, these are decision-makers in the behavior of systems or help in rational decision-making. Current models have reached levels of complexity by having problems where several dependent and several independent variables concur, a challenge that has drawn significant attention from researchers. Among the most current regression techniques is the Multiple Target Regression (MTR) where the main task is to simultaneously predict each objective variable from several independent variables.

Among the latest contributions to this technique is the proposal by (Borchani et al., 2015) that establishes two forms or ways of solution according to the approach of the problem. The problems of transformation of the problem and those of adaptation of the method are then raised (Chen et al., 2021). These differ by themselves in exploiting the interrelationship between variables to make a prediction. Adaptation-based problems take into account the relationships between the output variables, while transformation-based problems decompose the multi-objective problem into several output variables (Fang et al., 2015; Zhang et al., 2017; Zhen et al., 2017; Wang et al., 2018; Joshi et al., 2020).

According to Spyromitros-Xioufis et al., (2016), when an MTR problem is modeled, it is taken into account that the input is made up of two vectors, one input X and the other output Y, where each one consists of n variables, one X can then be defined as a set of input variables $X_1, \dots, X_i$ and Y as the set of target variables $Y_1, \dots, Y_j$, therefore vectors can then be defined as $X = [X_1, \dots, X_i]$ and $Y = [Y_1, \dots, Y_j]$.

Once an MTR model is conceive, it is evaluate to verify to what extent it fits the training data. Learning in a model of this type is carried out precisely with the use of regressive algorithms so that the model learns from the knowledge base and then can predict a given value. To evaluate a model, then, one of the methods used is cross validation (Kohavi, 1995; Refaeilzadeh et al., 2016; Berrar, 2019).

Cross validation or cross-validation is a technique used to evaluate the results of a statistical analysis and ensure that they are independent of the partition between training and test data. It consists of repeating and calculating the arithmetic mean obtained from the evaluation measures on different partitions. It is used in environments where the main objective is prediction and the accuracy of a model that will be carried out in practice is to be estimated. It is a technique widely used in artificial intelligence projects to validate generated models. Cross-validation is a way to predict the fit of a model to a hypothetical set of test data (Refaeilzadeh et al., 2016; Berrar, 2019).

Regression algorithms

The regressor algorithms studied in this research are found in the WEKA library, these are:

The IBk algorithm does not build a model, instead it generates a prediction for a test instance just-in-time. The IBk algorithm uses a distance measure to locate k instances in the training data for each test instance and uses those selected instances to make a prediction (Amin & Habib, 2015).

Locally Weighted Regression (LWL) or LOWESS. LOESS or LOWESS are nonparametric regression methods that combine multiple regression models in k-nearest-neighbor based model. Most of the algorithms such as classical feedforward neural network, support vector machines, nearest neighbor algorithms etc (Cambronero & Moreno, 2006; Mariño, 2015).

The principal difference of K* against other IB algorithms is the use of the entropy concept for defining its distance metric, which is calculated by mean of the complexity of transforming an instance into another; so it is taken into account the probability of this transformation occurs in a random walk away manner. The classification with K* is made by summing the probabilities from the new instance to all of the members of a category (Cleary & Trigg, 1995).

This must be done with the rest of the categories, to finally select that with the highest probability (Barrios et al., 2015).

Regression metrics of evaluation

Now, to evaluate regression models are exists some metrics:

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \quad Predicted_j\text{-}Current \quad _j)^2}$$

one way to assess how well a regression model fits a dataset is to calculate the root mean square error, which is a metric that tells us the average distance between the predicted values from the model and the actual values in the dataset (Despotovic et al., 2016).

Average Root Mean Square Error (Average RMSE) is the average of the RMSE of the data set.

$$AverageRMSE = \frac{\sum_{j=1}^{n} \quad MSE/}{n}$$

Relative Root Mean Squared Error (RRMSE) this indicator is calculate by dividing RMSE with average value of measured data. (Despotovic et al., 2016).

$$RRMSE = \frac{\sqrt{\frac{1}{n}\ \sum_{j=1}^{n}(Predicted_j\text{-}Actual_j)^2}}{\sum_{j=1}^{n} Actual_j}100$$

Mean Absolute Error (MAE) is the magnitude of difference between the prediction of an observation and the true value of that observation. MAE takes de average of absolutes errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

$$MAE = \frac{1}{n}\sum_{j=1}^{n} \quad |Predictedj\_ \quad Current\_j|$$

Average Mean Absolute Error (Average MAE) is the average of the MAE of the data set.

$$AverageMAE = \frac{\sum_{j=1}^{n} \quad MAE/}{n}$$

Relative Mean Absolute Error (Relative MAE) resents the ratio of the error between the measured value and the predicted dataset to the measured value of all the points (Li et al., 2018).

$$RelativeMAE = \frac{1}{n}\sum_{j=1}^{n} \frac{|Predictedj\_\ \ Current\_j|}{Predicted_j}$$

Average Relative Absolute Error (Average Relative MAE) is the average of the RMSE of the data set.

$$AverageRelativeMAE = \frac{\sum_{j=1}^{n} M\,AE/}{n}$$

3. Results and discussion

Because the principal objective of the research is based on to, find the algorithm that shows the best results in learning the databases of the treated species, all the algorithms in the MULAN library were tested to observe the learning behavior of the algorithms. For this task, were selected all algorithms can be used on regression problems. We specifically focused on instance-based ones, because in previous research these are show the best results in learning. The results showed in the next table, reflect the learning of the algorithms put in competition, that showed a better performance in learning were the lazy algorithms, in this case only the IBK and KStar algorithms was tested, the result is because they are instance-based algorithms, they can work with little data and use the probability that one object is similar to another to predict values given an input. In the case of predicting a numerical value, it is express in the way that a number can be approach to another, for this are used the approximation measures, as the nearest neighbor technique used by IBK while the measure used by KSTAR is based on the relative entropy between objects.

Table 1: Results of learning of machine learning

Algorithm	Erythrina variegata			Gliricidia sepium		
	Train 1	Train 2	Train 3	Train 1	Train 2	Train 3
M5P	0.6344	0.4075	3.6233	0.8816	1.2631	1.4116
M5PRULER	0.4004	0.2912	5.7588	0.6733	1.1885	1.2510
LINEARREGRESSION	0.1893	0.6994	24.9956	1.0639	0.9014	2.0969
IBK	0.0803	0.0809	0.1404	0.0957	0.8193	0.1498
KSTAR	0.0752	0.0809	0.1226	0.0823	0.7216	0.1369
ZERO	5.4043	1.9451	6.6927	6.3610	3.0554	5.4754
KSVM	0.2155	0.1556	0.4791	0.7948	1.0604	0.6712
SMOREG	0.3559	0.8400	5.4309	1.4938	1.2586	0.4980
DS	1.6679	1.1915	2.4516	2.4842	2.0812	3.5238

Algorithm	Erythrina variegata			Gliricidia sepium		
	Train 1	Train 2	Train 3	Train 1	Train 2	Train 3
MP	0.1122	0.1071	0.9798	0.2292	1.0130	0.3179
GP	1.2388	0.7448	3.7722	1.9998	1.3003	1.7266
RT	0.1785	0.2200	0.6036	0.4012	1.0413	0.8650
	Leucaena leucocephala			Tithonia diversifolia		
M5P	0.9367	0.8492	1.0961	0.1682	0.4329	1.0953
M5PRULER	0.5168	0.5157	0.9820	0.1170	0.3033	0.8341
LINEARREGRESSION	1.0316	1.0990	1.3070	0.1533	0.6576	1.5512
IBK	0.1115	0.0775	0.2271	0.0783	0.1068	0.1426
KSTAR	0.0893	0.0775	0.2257	0.0733	0.1068	0.1210
ZERO	4.5769	2.6097	4.4963	1.8289	2.5271	5.1321
KSVM	0.2565	0.2482	0.6244	0.1064	0.1954	0.5048
SMOREG	1.4007	1.4061	0.6379	0.1572	0.7590	0.8611
DS	2.9063	1.7746	2.5740	0.9467	1.1404	2.7933
MP	0.1318	0.1288	0.3835	0.0889	0.1392	0.4133
GP	1.6843	1.2084	1.4599	0.3435	0.8445	0.8441
RT	0.2217	0.2559	1.2949	0.1150	0.2516	0.5714

In the table 1, includes results of a study of the regressor algorithms of the MULAN tool, in this all were put to compete, to observe the learning behavior for each database of all of the species' studies. As previously explained, in this case the ones that showed the best results were those based on instances (IBK and KSTAR). In this case, was created three databases for each species, that respond to secondary metabolites, cell wall components and digestibility, all algorithms were tested on each database of each variety. Hence, the instance-based algorithms (IBK and KSTAR) are the ones that showed the best performance in learning with databases. This is since they learn from previous cases and with the use of distance measurements between two points (protein values, digestibility and composition of the cell wall) to be able to predict the real values of the different constituents of the nutritional value of each one of the forage species or protein plants studied.

Pascual et al. (2016) when using artificial neural networks to estimate the components of yield and nutritional value of three species of pasture grasses. The use of this technique through a multilayer perceptor network, which allowed estimating these indicators for the first time in Cuba, from databases for learning with information collected from scientific publications and data from referenced laboratories of the Humboldt University, Germany. Studies that served as the basis for our research since, despite being neural networks and regression models, the most recommended ways when you want to predict a numerical value from the set of real values. From the comparison of the two solutions in our study, better values were obtaine with the regression models of multiple objectives. These results are because the networks during the learning process can appear the so-called false positives, which generates an over-prelearning of the model. In these models based on neural networks, it is very complex to control the internal process and the interaction between the neurons that make up the model, which is why it is often unlikely to detect an over-learning phenomenon.

While, Estrada-Jiménez et al. (2018), through a comparison between the regressor algorithms of the WEKA tool, they reported that the estimation of the phytochemical components of *Leucaena leucocephala* and *Tithonia diversifolia* from the variables of climate, rebound age and primary compounds (nitrogen and sugars) products of the photosynthetic activity of the plant, there was a better response for the KSTAR algorithm when evaluating the performance of the predictions using the aRRMSE function. Initial tests that served to establish for the present study the division of the training sets by processes, which favored a higher performance of the algorithms based on instances.

After to realize the preliminary study with all the algorithms and to see that the ones with the best performance were those based on instances, the work was centered on evaluating only those of this type in the tool. The Table 2 shows the principal measures of evaluation for this type of task. As you can see, this tool has 3 algorithms for regression, to putting them to compete, is observed that the most efficient training was about KStar. With this result then the other evaluation measures described above can also be seen. Now, the principal purpose is to find the one that learns the best to later create a tool, that automatically predict these values from input data which are, soil data, rebound age, primary metabolites, and climate.

In previous research (Spyromitros-Xioufis et al., 2016; Santana et al., 2017; Estrada-Jiménez et al., 2019; Waegeman et al., 2019; Estrada-Jiménez et al., 2020; Chen et al., 2021), various machine learning algorithms have also been tested, as well as regressor, with satisfactory results. In this only the lazy algorithms were put to the test since these are based on the probability that an object can resemble others; in these investigations the datasets had not been divided, this responded to a variant to test the behavior of all types of algorithms to select the best one through aRMSE, in these the best ones always turned out to be the lazy ones, therefore hence the decision to in this investigation test only the sloths.

Also, several models have been compared that included in most cases a single dataset, this contained in a first experimentation all the data of the studied varieties without considering the flow of the processes that we tried to simulate, even so the lazy algorithms always showed a better adaptation to them even with this drawback.

Then, in consultation with specialists from the department of pastures and forages of the University of Granma, it was decided to separate the dataset of how the process flow is carried out in a laboratory, therefore the datasets were separated, and 3 datasets were created datasets that respond to phytochemical components, cell wall components and digestibility components, which is as shown in Table 2. When comparing. In the aRMSE values, it can be observed that the aRMSE decreases with respect to Table 2, so it can be affirmed that with the new variant of separating the data set, the algorithms simulate with better quality the behavior of the plants studied in this research.

Table 2. Results of learning of machine learning

Erythrina variegata				
Algorithm	**Metric**	**Train1**	**Train2**	**Train3**
LWL	AverageRMSE	1.0262±0.3615	0.6441±0.1141	1.3213±0.2831
	AverageRelativeRMSE	0.3197±0.1438	0.3760±0.1484	0.3297±0.2227
	AverageMAE	0.8548±0.2322	0.5622±0.1058	1.1121±0.2124
	AverageRelativeMAE	0.3097±0.1487	0.3842±0.1747	0.3127±0.1951
IBK	AverageRMSE	1.0262±0.3615	0.0543±0.0167	0.138±0.0383
	AverageRelativeRMSE	0.3197±0.1438	0.0483±0.0222	0.0835±0.0728
	AverageMAE	0.8548±0.2322	0.0405±0.0115	0.1036±0.0267
	AverageRelativeMAE	0.3097±0.1487	0.0444±0.0211	0.0769±0.0660
KSTAR	AverageRMSE	1.0262±0.3615	0.054±0.0167	0.1227±0.0368
	AverageRelativeRMSE	0.3197±0.1438	0.0483±0.0222	0.0724±0.0574
	AverageMAE	0.8548±0.2322	0.0405±0.0115	0.0927±0.0238
	AverageRelativeMAE	0.3097±0.1487	0.0444±0.0211	0.0642±0.04691
Gliricidia sepium				
LWL	AverageRMSE	1.7472±0.5762	1.9449±0.5887	1.7135±0.3222
	AverageRelativeRMSE	0.3699±0.1067	0.6949±0.1912	0.3725±0.1124
	AverageMAE	1.3327±0.4196	1.5623±0.5724	1.4414±0.2912
	AverageRelativeMAE	0.3479±0.1147	0.6254±0.2084	0.3689±0.1208
IBK	AverageRMSE	0.0880±0.0311	1.8194±0.6851	0.1398±0.0716
	AverageRelativeRMSE	0.0602±0.0650	0.6522±0.2119	0.0294±0.0119
	AverageMAE	0.0634±0.0207	1.3083±0.6846	0.0953±0.0442
	AverageRelativeMAE	0.0466±0.0478	0.5288±0.2372	0.0236±0.0099
KSTAR	AverageRMSE	0.0762±0.0266	1.8194±0.6851	0.1323±0.0709
	AverageRelativeRMSE	0.0518±0.0423	0.6522±0.2119	0.0287±0.0115
	AverageMAE	0.0552±0.0171	1.3083±0.6846	0.0924±0.0449
	AverageRelativeMAE	0.0414±0.0317	0.5288±0.2372	0.0233±0.0090
Leucaena leucocephala				
LWL	AverageRMSE	0.8748±0.1156	0.7728±0.1055	1.2730±0.3951
	AverageRelativeRMSE	0.3102±0.1861	0.4689±0.1521	0.3550±0.1450
	AverageMAE	0.7878±0.1127	0.6805±0.1104	1.0245±0.3081
	AverageRelativeMAE	0.3293±0.1930	0.4817±0.1820	0.3269±0.1398
IBK	AverageRMSE	0.1047±0.0290	0.0690±0.0257	0.2106±0.1811
	AverageRelativeRMSE	0.0702±0.0345	0.0923±0.0369	0.0677±0.0616
	AverageMAE	0.0775±0.0165	0.0519±0.0184	0.1463±0.0983
	AverageRelativeMAE	0.0592±0.0289	0.0825±0.0311	0.0577±0.0535
KSTAR	AverageRMSE	0.0873±0.0217	0.0690±0.0257	0.2041±0.1729
	AverageRelativeRMSE	0.0574±0.0311	0.0923±0.0369	0.0620±0.0511
	AverageMAE	0.0646±0.0138	0.0519±0.0184	0.1406±0.0924
	AverageRelativeMAE	0.0484±0.0284	0.0825±0.0311	0.0524±0.0432

Tithonia diversifolia				
Algorithm	**Metric**	**Train1**	**Train2**	**Train3**
LWL	AverageRMSE	0.3859±0.1235	0.5390±0.1134	1.0671±0.3649
	AverageRelativeRMSE	0.3138±0.1277	0.3930±0.1444	0.3778±0.2450
	AverageMAE	0.3313±0.1011	0.4700±0.1186	0.8699±0.3224
	AverageRelativeMAE	0.3232±0.1410	0.4175±0.1476	0.3911±0.2510
IBK	AverageRMSE	0.0686±0.0224	0.0966±0.0304	0.1385±0.0653
	AverageRelativeRMSE	0.0962±0.0620	0.1372±0.0336	0.0688±0.0408
	AverageMAE	0.0523±0.0140	0.0707±0.0187	0.0974±0.0361
	AverageRelativeMAE	0.0775±0.0411	0.1314±0.0274	0.0603±0.0358
KSTAR	AverageRMSE	0.0661±0.0224	0.0966±0.0304	0.1240±0.0398
	AverageRelativeRMSE	0.0950±0.0583	0.1372±0.0336	0.0664±0.0369
	AverageMAE	0.0494±0.0142	0.0707±0.0187	0.0894±0.0255
	AverageRelativeMAE	0.0759±0.0391	0.1314±0.0274	0.0594±0.0355

With the data studied by learning the algorithms, it was possible to verify that the algorithms that had the best adaptation to the data were obviously the lazy algorithms. The performance measure used responds to the fact that in regression problems these are the measures to be used, but the same does not happen when the problem to be treated is classification. As has been used by several authors (Karalič & Bratko, 1997; Tuia et al., 2011; Osojnik et al., 2017; Reyes et al., 2018; Camejo-Corona et al., 2019), this aRMSE measure is the most representative in a model, even though it is known that among its limitations is precisely that the average encompasses all the values that are included within it, therefore, if at any time there is any high very high or low very low value may directly affect the average.

In comparison with the research developed by Estrada-Jiménez et al, (2018), it was possible to include the data referring to soil components, digestibility, and cell wall components. He proposed a model that predicted only the secondary metabolites from the primary metabolites, climate and rebound age. A significant detail is the reduction of the error evaluated to select the regressor algorithm. In addition, in the present paper the aRMSE is optimized by the creation of the databases by processes to be determined, that is, a data set to learning to secondary metabolites, cell wall components and digestibility respectively.

Painuli et al. (2014) reported the effectiveness of the KSTAR algorithm in predicting the wear of agricultural machinery parts, based on the collection of a set of data and characteristics of these parts, with which the data set for learning was formed and with With the application of this algorithm, the effectiveness of the predictions could be evaluated at 78%, a value that is considered high due to the adaptability of the algorithm to the data set (Painuli et al., 2014).

The use of artificial intelligence as a powerful tool to predict different life processes and different branches of science is a practice that has gained popularity in recent years due to its practical utility and high levels of precision. In this sense, Erdal et al. (2018) developed studies with algorithms based on instances (lazy) to simulate the evaluation of concrete quality. At first, all the algorithms of the WEKA tool, which contains the relevant libraries for data mining, were evaluate. Then the data was evaluated only with the lazy algorithms, from the error it was possible to determine the high performance of the instance-based algorithms (LW, IBK and KSTAR). While Maliha et al. (2019) to predict the causes and appearance of cancer found when using algorithms J-48 and KSTAR that in logistic regression the accuracy is 99,3%; for KSTAR it was 99,5% and J-48 is 99,1 %.

However, Zighed and Bounour (2019) used the KSTAR algorithm to assess software maintenance; based on the Quantity of codes to be implement for the maintenance of a specific computer product. Prediction models based on data collected from two object-oriented systems were create. In addition, the models created with the linear regression algorithms, neural network, decision tree, SVM, were compare with the use of the WEKA tool; where comparisons of the prediction accuracy of all models were established using and cross-validation. As a result, it shown that KSTAR produces better results by predicting more accurately than the other techniques. It should be note that the present study, using this tool, eliminates the multicollinearity between the input variables, eliminating the correlation between them to avoid setbacks.

Khosravi et al. (2021), using field data at one station, succeeded in predicting flow depth, water surface width, and water surface longitudinal slope using independent data mining techniques: database learning. instances (IBK), KSTAR, locally weighted learning (LWL), Vote, Attribute Selected Classifier (ASC), Regression by Discretization (RBD) and Cross-validation Parameter Selection (CVPS) (Vote-IBK, Vote-KSTAR, Vote-LWL, ASC-IBK, ASC-KSTAR, ASC-LWL, RBD-IBK, RBD-KSTAR, RBD-LWL, CVPS-IBK, CVPS-KSTAR, CVPS-LWL). Through a comparison of predictive performance and a sensitivity analysis of the driving variables, the results reveal that among other features the Vote-KSTAR model had the highest performance in predicting depth and width, and ASC-KSTAR in the estimation of the slope.

The results obtained in this research attest to the good behavior of the adaptability of the algorithms and artificial intelligence to predict the components studied. In this way, it is evident that for future studies it is advisable to use these procedures based on instances, which is due to the behavior of these before others.

4. Conclusions and recommendations

- The aRRMSE was optimized with respect to previous investigations
- The species that showed the best behavior was *Leucaena leucocephala* with the KStar algorithm
- Three datasets were created for each plant variety to evaluate the behavior of lazy algorithms.
- The datasets of the plant varieties created were tested with the lazy algorithms of the WEKA tool developed by the University of Waikato to evaluate the adaptability of these with the datasets.
- The results of the training with each of the datasets were evaluated with the conventional metrics for the evaluation of the regression algorithms. It was verified that the algorithm that presented the best aRMSE turned out to be KStar, which shows that this is the one that can best simulate the behavior of the properties of the varieties studied.
- We recommended, test the proposed and trained models with test cases designed by specialists from the Center for Animal Production Studies of the University of Granma.
- Consider the results proposed for the use of these in a computational tool that can gather and learn from these databases with these algorithms to simulate the behavior of plant components. Evaluate the methodology used and the flow of processes in the study of other varieties of plants used for animal nutrition.

Bibliography

Alebele, Y., Zhang, X., Wang, W., Yang, G., Yao, X., Zheng, H., Zhu, Y., Cao, W. & Cheng, T. (2020). Estimation of Canopy Biomass Components in Paddy Rice from Combined Optical and SAR Data Using Multi-Target Gaussian Regressor Stacking. *Remote Sensing*, *12*(16), 2564. https://doi.org/10.3390/rs12162564/

Alzubi, J., Nayyar, A. & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. Journal of physics: conference series, *1142*(1), 012012. https://doi.org/10.1088/1742-6596/1142/1/012012/

Amin, M. N. & Habib, A. (2015). Comparison of different classification techniques using WEKA for hematological data. *American Journal of Engineering Research*, *4*(3), 55-61. http://www.ajer.org/papers/v4(03)/H043055061.pdf/

Barrios, H. D., Rivas, Y. A., Hernández, L. C., Hernández, A. M., Cárdenas, M. del C. C. & Cardoso, G. M. C. (2015). Algoritmos de aprendizaje automático para clasificación de Splice Sites en secuencias genómicas. *Revista Cubana de Ciencias Informáticas*, *9*(4), 155-170. http://scielo.sld.cu/pdf/rcci/v9n4/rcci12415.pdf/

Berrar, D. (2019). Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology,1, 542–545.. https://doi.org/10.1016/B978-0-12-809633-8.20349-X/

Borchani, H., Varando, G., Bielza, C. & Larranaga, P. (2015). A survey on multi-output regression. *Wires Data Mining and Knowledge Discovery*, *5*(5), 216-233. https://doi.org/10.1002/widm.1157/

Cabrera, D. (2008). Manejo y uso de pastos y forrajes en ganadería tropical. *Universidad de Córdoba*, pp 40. http://www.uco.es/zootecniaygestion/img/pictorex/08_21_24_4.1.1.pdf/

Cambronero, C. G. & Moreno, I. G. (2006). Algoritmos de aprendizaje: Knn & kmeans. *Inteligencia en Redes de Comunicación, Universidad Carlos III,* Madrid, Spain. pp. 8. http://blogs.ujaen.es/barranco/wp-content/uploads/2012/02/Algoritmos-de-aprendizaje-knn-y-kmeans.pdf/

Camejo-Corona, J., Gonzalez, H. & Morell, C. (2019). Los principales algoritmos para regresión con salidas múltiples. Una revisión para Big Data. *Revista Cubana de Ciencias Informáticas*, *13*(4), 118-150. http://scielo.sld.cu/pdf/rcci/v13n4/2227-1899-rcci-13-04-118.pdf/

Chen, S., Gu, C., Lin, C. & Hariri-Ardebili, M. A. (2021). Prediction of arch dam deformation via correlated multi-target stacking. *Applied Mathematical Modelling*, *91*, 1175-1193. https://doi.org/10.1016/j.apm.2020.10.028/

Cleary, J. G. & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. En *Machine Learning Proceedings 1995*, 108-114. https://sci2s.ugr.es/keel/pdf/algorithm/congreso/KStar.pdf/

Coraddu, A., Oneto, L., Ghio, A., Savio, S., Anguita, D. & Figari, M. (2016). Machine learning approaches for improving condition-based maintenance of naval propulsion plants. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, *230*(1), 136-153. https://doi.org/10.1177/1475090214540874/

Despotovic, M., Nedic, V., Despotovic, D. & Cvetanovic, S. (2016). Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. *Renewable and Sustainable Energy Reviews*, *56*, 246-260. https://doi.org/10.1016/j.rser.2015.11.058/

Díaz, A., Cayón, G. & Mira, J. J. (2007). Metabolismo del calcio y su relación con la «mancha de madurez» del fruto de banano. Una revisión. *Agronomía Colombiana*, 25(2), 280-287. https://revistas.unal.edu.co/index.php/agrocol/article/view/14131/14886/

Džeroski, S., Demšar, D. & Grbović, J. (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, *13*(1), 7-17. https://doi.org/10.1023/A:1008323212047/

Erdal, H., Erdal, M., Simsek, O. & Erdal, H. I. (2018). Prediction of concrete compressive strength using non-destructive test results. *Computers and Concrete*, *21*(4), 407-417. https://doi.org/10.12989/cac.2018.21.4.407/

Estrada-Jiménez, P. M., Diez, H. R., Cabrera, A. V., Verdecia, D. M. & Ramírez, J. L. (2018). Modelos de predicción de metabolitossecundarios para dos variedades de plantas protéicas. Memorias del *VIII Congreso Iberoamericano de Ingeniería de Proyectos. Universidad de Ciencias Informáticas*, Cuba, pp1-9. https://repositorio.uci.cu/bitstream/123456789/9499/1/A205.pdf/

Estrada-Jiménez, P. M., González-Diez, H. R., Verdecia-Cabrera, A., Verdecia-Acosta, D. M. & Ramírez-de la Rivera, J. L. (2018). Modelos de predicción de metabolitos secundarios para dos variedades de plantas proteicas. *Libro de Memorias: VIII Congreso Iberoamericano de Ingeniería de Proyectos. Ediciones Futuro. Universidad de Ciencias Informáticas, Cuba,* pp 1-9. https://repositorio.uci.cu/jspui/bitstream/123456789/9499/1/A205.pdf/

Estrada-Jiménez, P. M., Noguera-López, P. J. & Recio-Avilés, R. (2020). Aplicación de la regresión de múltiples objetivos en la estimación de componentes fitoquímicos. *Pensamiento Matemático*, *10*(2), 7-14. https://dialnet.unirioja.es/servlet/articulo?codigo=7782227/

Estrada-Jiménez, P. M., Ramírez-de la Ribera, J. L., Verdecia-Acosta, D. M. & Soler-Pellicer, Y. (2019). Aplicación de la minería de datos en la estimación de componentes fotoquímicos (Original). *Roca. Revista científico-educacional de la provincia Granma*, *15*(2), 177-186. https://dialnet.unirioja.es/servlet/articulo?codigo=7013276/

Fang, J., Li, Y., Liu, R., Pang, X., Li, C., Yang, R., He, Y., Lian, W., Liu, A.L. & Du, G.H. (2015). Discovery of multitarget-directed ligands against Alzheimer's disease through systematic prediction of chemical–protein interactions. *Journal of chemical information and modeling*, *55*(1), 149-164. https://doi.org/10.1021/ci500574n/

González, F. A. (2015). Machine learning models in rheumatology. *Revista Colombiana de Reumatología*, *22*(2), 77-78. http://dx.doi.org/10.1016/j.rcreu.2015.06.001/

Herrera, R.S., Verdecia, D.M., Ramírez, J.L., García, M. & Cruz, A.M. (2017). Relation between some climatic factors and the chemical composition of *Tithonia diversifolia*. *Revista Cubana de Ciencia Agrícola*, *51*(2), 271-279. http://cjascience.com/index.php/CJAS/article/view/719/

Joshi, R. S., Jagdale, S. S., Bansode, S. B., Shankar, S. S., Tellis, M. B., Pandya, V. K., Chugh, A., Giri, A. P. & Kulkarni, M. J. (2020). Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. *Journal of Biomolecular Structure and Dynamics*, 1-16. https://doi.org/10.1080/07391102.2020.1760137/

Karalič, A. & Bratko, I. (1997). First order regression. *Machine learning*, *26*(2), 147-176. https://link.springer.com/content/pdf/10.1023/A:1007365207130.pdf/

Khosravi, K., Khozani, Z. S. & Cooper, J. R. (2021). Predicting stable gravel-bed river hydraulic geometry: A test of novel, advanced, hybrid data mining algorithms. *Environmental Modelling & Software*, *144*, 105165. https://doi.org/10.1016/j.envsoft.2021.105165/

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137-1145. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf/

Li, J., Zhang, L., He, C. & Zhao, C. (2018). A comparison of Markov chain random field and ordinary kriging methods for calculating soil texture in a mountainous watershed, northwest China. *Sustainability*, *10*(8), 2819. https://doi.org/10.3390/su10082819/

Mahecha, L. & Rosales, M. (2005). Valor nutricional del follaje de botón de oro *Tithonia diversifolia* (Hemsl.) Gray, en la producción animal en el trópico. *Livestock Research for Rural Development*, *17*(9), 1. https://www.lrrd.cipav.org.co/lrrd17/9/mahe17100.htm/

Mahecha, L., Escobar, J., Suárez, J. & Restrepo, L. (2007). *Tithonia diversifolia* (hemsl.) Gray (botón de oro) como suplemento forrajero de vacas F1 (Holstein por Cebú). *Livestock Research for Rural Development*, *19*(2), 1-6. https://lrrd.cipav.org.co/lrrd19/2/mahe19016.htm/

Maliha, S. K., Islam, T., Ghosh, S. K., Ahmed, H., Mollick, Md. R. J. & Ema, R. R. (2019). Prediction of Cancer Using Logistic Regression, K-Star and J48 algorithm. *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, 1-6. https://doi.org/10.1109/EICT48899.2019.9068790/

Mariño, A. P. (2015). GMLKNN: modelo basado en instancias para el aprendizaje multi-etiqueta utilizando la distancia VDM [PhD Thesis]. Universidad Central "Marta Abreu" de Las Villas. Facultad de Matemática. pp. 100. https://dspace.uclv.edu.cu/bitstream/handle/123456789/7551/Tesis%20Final.pdf?sequence=1&isAllowed=y/

Mastelini, S. M., Santana, E. J., Cerri, R. & Barbon Jr, S. (2020). DSTARS: a multi-target deep structure for tracking asynchronous regressor stacking. *Applied Soft Computing*, *91*, 106215. https://doi.org/10.1016/j.asoc.2020.106215/

Nogueira, M. S. & Koch, O. (2019). The development of target-specific machine learning models as scoring functions for docking-based target prediction. *Journal of chemical information and modeling*, *59*(3), 1238-1252. https://doi.org/10.1021/acs.jcim.8b00773/

Osojnik, A., Panov, P. & Džeroski, S. (2017). Multi-label classification via multi-target regression on data streams. *Machine Learning*, *106*, 745-770. https://doi.org/10.1007/s10994-016-5613-5/

Otegui, M.B. & Totaro, M. E. (2007). *Atlas de histología vegetal*. EDUNAM-Editorial Universitaria de la Universidad Nacional de Misiones. https://editorial.unam.edu.ar/images/documentos_digitales/978-950-579-064-7.pdf

Painuli, S., Elangovan, M. & Sugumaran, V. (2014). Tool condition monitoring using K-star algorithm. *Expert Systems with Applications*, *41*(6), 2638-2643. https://doi.org/10.1016/j.eswa.2013.11.005/

Pascual, I. de los A., Ramírez, J., & Ortiz, A. (2016). Métodos de Inteligencia Artificial para la predicción del rendimiento y calidad de gramíneas. *REDVET. Revista Electrónica de Veterinaria*, *17*(12). https://www.redalyc.org/pdf/636/63649052026.pdf/

Ramírez-Lozano, R. (2010). *Importancia de los taninos condensados en la nutrición del venado cola blanca*. Conferencia: 5° Simposio sobre Fauna Cinegética en México At: Puebla, México (1), 1-21. https://www.researchgate.net/publication/268207092_Importancia_de_los_taninos_condensados_en_la_nutricion_del_venado_cola_blanca/

Refaeilzadeh, P., Tang, L. & Liu, H. (2016). Cross-Validation. In Liu L & Özsu MT (Eds.), *Encyclopedia of Database Systems* (pp. 1–6). Springer New York. http://leitang.net/papers/ency-cross-validation.pdf/

Reyes, O., Cano, A., Fardoun, H. M. & Ventura, S. (2018). A locally weighted learning method based on a data gravitation model for multi-target regression. *International Journal of Computational Intelligence Systems*, *11*(1), 282-295. https://doi.org/10.2991/ijcis.11.1.22/

Rincón-Tuexi, J. A., Castro-Nava, S., López-Santillán, J. A., Huerta, A. J., Trejo-López, C. & Briones-Encinia, F. (2006). Temperatura alta y estrés hídrico durante la floración en poblaciones de maíz tropical. *Phyton (Buenos Aires)*, *75*, 31-40. http://www.scielo.org.ar/pdf/phyton/v75/v75a03.pdf/

Ruiz, T. E., Febles, G. J., Galindo, J. L., Savón, L. L., Chongo, B. B., Torres, V., Cino, D. M., Alonso, J., Martínez, Y., Gutiérrez, D., Crespo, G. J., Mora, L., Scull, I., La O, O., González, J., Lok, S., González, N. & Zamora, A. (2014). *Tithonia diversifolia*, sus posibilidades en sistemas ganaderos. *Revista Cubana de Ciencia Agrícola*, *48*(1), 79-82. https://www.redalyc.org/pdf/1930/193030122017.pdf/

Ruiz, T., Febles, G., Castillo, E., Jordan, H., Galindo, J., Chongo, B., Delgado, D., Mejías, R. & Crespo, G. (2011). Tecnología de producción animal mediante Leucaena leucocephala asociada con pastos en el 100% del área de la unidad ganadera. *Sitio Argentino de Producción Animal.* https://www.produccion-animal.com.ar/produccion_y_manejo_pasturas/pasturas_cultivadas_megatermicas/112-leucaena.pdf/

Santana, E. J., Mastelini, S. M. & Barbon Jr, S. (2017). Deep regressor stacking for air ticket prices prediction. In *Anais do XIII Simpósio Brasileiro de Sistemas de Informação*, (pp. 25-31). Porto Alegre: SBC. https://doi.org/10.5753/sbsi.2017.6022/

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W. & Vlahavas, I. (2016). Multi-target regression via input space expansion: Treating targets as inputs. *Machine Learning*, *104*, 55-98. https://doi.org/10.1007/s10994-016-5546-z/

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. & Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, *12*(Jul), 2411-2414. https://www.jmlr.org/papers/volume12/tsoumakas11a/tsoumakas11a.pdf/

Tuia, D., Verrelst, J., Alonso, L., Pérez-Cruz, F. & Camps-Valls, G. (2011). Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geoscience and Remote Sensing Letters*, *8*(4), 804-808. https://matlabtools.com/wp-content/uploads/p603.pdf/

Verdecia, D.M., Herrera, R.S., Ramírez, J.L., Bodas, R., Leonard, I., Giráldez, F., Andrés, S., Santana, A., Méndez-Martínez, Y. & López, S. (2018). Yield components, chemical characterization and polyphenolic profile of *Tithonia diversifolia* in Valle del Cauto, Cuba. *Cuban Journal of Agricultural Science*, *52*(4), 457-471. http://cjascience.com/index.php/CJAS/article/view/838/

Waegeman, W., Dembczyński, K. & Hüllermeier, E. (2019). Multi-target prediction: A unifying view on problems and methods. *Data Mining and Knowledge Discovery*, *33*(2), 293-324. https://arxiv.org/pdf/1809.02352.pdf/

Wang, X., Zhen, X., Li, Q., Shen, D. & Huang, H. (2018). Cognitive assessment prediction in Alzheimer's disease by multi-layer multi-target regression. *Neuroinformatics*, *16*(3-4), 285-294. https://doi.org/10.1007/s12021-018-9381-1/

Zhang, J., Li, Q., Caselli, R. J., Thompson, P. M., Ye, J. & Wang, Y. (2017). Multi-source multi-target dictionary learning for prediction of cognitive decline. *International Conference on Information Processing in Medical Imaging*, 10265, 184-197. https://doi.org/10.1007/978-3-319-59050-9_15/

Zhen, X., Yu, M., He, X. & Li, S. (2017). Multi-target regression via robust low-rank learning. *IEEE transactions on pattern analysis and machine intelligence*, *40*(2), 497-504. https://ieeexplore.ieee.org/ielaam/34/8249508/7888599-aam.pdf/

Zighed, N. & Bounour, N. (2019). On The Use Of KStar Algorithm For Predicting Object-Oriented Software Maintainability. Conference *Internationale sur intelligence Artificielle et les Technologies Information ICAIIT 2019.* Pp. 1-5. https://dspace.univ-ouargla.dz/jspui/bitstream/123456789/20983/1/Zighed%20Narimane.pdf/