Pantanal Editora
Trends in Agricultural and Environmental Sciences
10.46420/TAES. e230006

Classification of Tropical Forage Grass Varieties Under Moderate and Severe Water Stress using Naïve Bayes and Kernel Density Estimation

Bruno Rodrigues **De Oliveira**[1,*], Renato Lustosa **Sobrinho**[2] and Marco Aparecido Queiroz **Duarte**[3]

[1] Pantanal Editora, Rua Abaete, 83, Sala B, Centro. Nova Xavantina-MT, Brazil;
[2] Federal University of Technology of Paraná (UTFPR), Brazil;
[3] State University of Mato Grosso do Sul (UEMS), Cassilândia Unit, Brazil;
* Correspondence: bruno@editorapantanal.com.br

Received: 2023-11-21
Accepted: 2023-11-23
Published: 2023-11-24

Main Editor
Alan Mario Zuffo
Jorge González Aguilera

Abstract: The selection of forage grasses that are more adapted to adverse conditions, such as water scarcity or dry rain periods, is extremely important. Mainly due to the severe climate changes, and the search for more sustainable ways of farming. Forage grasses form the basis of the diet of beef cattle and are also used as a source of biofuels, for erosion control and soil improvement. This work presents a machine learning methodology to obtain classification models for nine forage cultivars, subject to moderate and severe water stress. The Naïve Bayes algorithm is used together with the Kernel Density Estimation method to obtain the densities used in the classification models. Before learning the models, the grouped cross-validation technique and also the grid search are used to search for the best set of hyperparameters. The best accuracy and precision results are 0.88 and 0.90, respectively. It is observed that the classification performance depends on the cultivars used in the training and test sets. At the end, the estimated probability densities are also analyzed by comparing them with some statistics obtained for each variable and water stress or control environments. The proposed methodology is a complementary approach to classical statistical methods. It provides abstract models for obtaining information about the cultivar's harvesting environment.

Keywords: machine learning; smart agriculture; physiological variables.

For citation: De Oliveira, B. R.; Sobrinho, R. L.; Duarte, M. A. Q. (2023). Classification of Tropical Forage Grass Varieties Under Moderate and Severe Water Stress using Naïve Bayes and Kernel Density Estimation. Trends in Agricultural and Environmental Sciences, (e230006), DOI: 10.46420/TAES.e230006

1. Introduction

Tropical forage grasses are an important food source for livestock in tropical and subtropical regions. They are an essential component for ruminant animal production around the world. They provide essential nutrients for animal health and performance, including protein, fiber, energy, vitamins and minerals (Iqbal et al., 2019). They are also used for erosion control (Ghimire et al., 2015), soil improvement and biofuel production (Freitas et al., 2021). They are suited to the hot, humid conditions of tropical and subtropical regions and are also relatively drought tolerant and can be grown on marginal lands. Its importance is increasing due to the growing demand for animal feed. The world population is expected to reach 9.7 billion by 2050, which will put pressure on the global food supply (FAO, 2023). Raising animals for food production is one way to meet this demand, but it requires a reliable source of high-quality forage. Furthermore, considering issues related to sustainability, it is essential to obtain varieties of forage grasses that produce more in smaller spaces (Simeão et al., 2021), thus avoiding the deforestation of new areas. Although forage grasses have advantages for tropical regions, there are still some challenges to overcome in their production. One of them is the need to develop more efficient production methods. Another is the need to develop new varieties of forage grasses in genetic improvement programs that are resistant to pests, water stress and diseases (Jank et al., 2021). Climate change has increased the likelihood of extreme weather events,

presenting substantial obstacles to agricultural production. Agriculture relies heavily on consistent seasonal weather patterns, particularly with regard to rainfall patterns, which are among the main factors directly affecting plant cultivation (Luiz Piati et al., 2023). In this sense, the selection of forage grasses that are best adapted to these adverse conditions, such as water scarcity or periods of excess rainfall, is very important.

In the Brazilian Cerrado region the most tropical forage grasses used in livestock and agriculture systems are *Urochloa*, *Cynodon*, *Panicum*, *Paspalum* and *Pennisetum*. Zuffo et al. (2022) studied the cultivars *Urochloa brizantha* cv. BRS Piatã, *U. brizantha* cv. Marandu, *U. brizantha* cv. Xaraés, *U. ruziziensis* cv. Common, *Pennisetum glaucum* cv. ADR 300, *Panicum maximum* cv. Aruana, *P. maximum* cv. Mombaça, *P. maximum* cv. Tanzania, *Paspalum atratum* cv. Pojuca, to identify indices of tolerance to water stress. Since the quality of forage is an important factor for animal productivity, being influenced by several factors, including micronutrients (Iqbal et al., 2019) and the water stress (Habermann et al., 2019), such studies are essential. For this reason, a recent study conducted by de De Oliveira et al. (2023b) reevaluated the data presented by Zuffo et al. (2022) and presented a new approach for selecting forage cultivars subject to water stress.

Machine learning methods have been widely employed in agriculture in recent years (Meshram et al., 2021; Benos et al., 2021; Sharma et al., 2020). Applications include: analysis of pre-harvest parameters such as seeds and soil, and also post-harvest such as productivity, plant height, among others (Meshram et al., 2021). Benos et al. (2021) in their meta-analysis found that 68% of applications are related to crop management, 12% to livestock management, 10% to water management and the remaining 10% to soil management. Crop management applications are related to yield forecast, disease and weed detection, and crop and quality recognition.

Given the relevance of using machine learning methods in agriculture and also the importance of the research carried out by Zuffo et al. (2022), in this work we propose the use of such methods to obtain classification models for forage grasses subject to water stress. The analyzes performed by Zuffo et al. (2022) consisted of classical statistical, such as ANOVA, canonical correction analysis, correlation networks and also the use of tolerance indices proposed by Farshadfar et al. (2012). And in the selection method presented by De Oliveira et al. (2023b), Manhattan distances and TOPSIS were used. Therefore, machine learning methods have not yet been applied to data from the research conducted by Zuffo et al. (2022).

The main objective of this research is to use machine learning methods to obtain classification models. These models must classify samples of forage cultivars into three soil water regimes (class): "Control", "Moderate" and "Severe". To this end, we chose to use the Naïve Bayes algorithm, due to its advantages mentioned later. The generated models are probability distributions for each class and each variable. In addition to this approach, the Kernel Density Estimation (KDE) method is also used. With KDE we were able to generate more complex probability densities, enabling more accurate modeling.

The remainder of this work is divided as follows: in section 2 we present the experimental data obtained by Zuffo et al. (2022) and we also explain the concepts related to machine learning, so that the reader understands how the data was modeled; in section 3 we present the results obtained in the adjustment stage of the classification models, and also the performance results of these models; in addition, discussions of these results and possible applications of the proposed methodology are performed, as well as discussions of future works.

2. Material and Methods

2.1 Experimental data

The data were obtained from the experiment conducted at Cassilândia, Mato Grosso do Sul, Brazil (19º05'29"S and 51°48'50"W, and altitude of 540 m) from May to August 2019. In the

experiment, nine cultivars of tropical forage grasses were used: *Urochloa brizantha* cv. BRS Piatã, *U. brizantha* cv. Marandu, *U. brizantha* cv. Xaraés, *U. ruziziensis* cv. Comum, *Pennisetum glaucum* cv. ADR 300, *Panicum maximum* cv. Aruana, *P. maximum* cv. Mombaça, *P. maximum* cv. Tanzânia, *Paspalum atratum* cv. Pojuca, and three soil water regimes: high soil water regime (Control), medium soil water regime (Moderate) and low soil water regime (Severe). The experiment was organized in an experimental design in completely randomized blocks in a 3×9 factorial arrangement with four replications. There were used seeds of nine tropical forage cultivars, three commercial cultivars of *Urochloa brizantha* (Hochst. Ex A. Rich.) R.D.Webster ('BRS Piatã', 'Marandu', and 'Xaraés'), three commercial cultivars of *Panicum maximum* Jacq. ('Aruana', 'Mombaça', and 'Tanzânia'), one commercial cultivar of *Pennisetum glaucum* (L.) R. Br. ('ADR 300'), one commercial cultivar of *Urochloa ruziziensis* (R. Germ. & C.M. Evrard) Crins ('Comum'), and a commercial cultivar of *Paspalum atratum* Swallen ('Pojuca') (Zuffo et al., 2022).

2.2 Machine Learning, Naïve Bayes and Kernel Density Estimation

Machine learning is the field of artificial intelligence that allows computers to learn without being explicitly programmed. One of its main objectives is pattern recognition. In supervised learning, each data sample (pattern) $\boldsymbol{d}$ is associated with a label l. These samples are vectors, i.e., $\boldsymbol{d} = [d_1\ d_2\ \dots d_N]$ (Haykin, 2009; Theodoridis and Koutroumbas, 2006). The aim is then to learn a function (model) $\hat{f}(d) = L$ that is an estimate of a real function $f(d) = L$, which is unknown (given the complex nature of association and interaction of the data). In other words, we want a model that, for a given input $\boldsymbol{d}$ with N attributes, returns an output (label or class) l. Dataset D contains all $\boldsymbol{d}$ patterns from the analyzed data. To learn these models, D is divided into two subsets: training and testing (Bishop, 2006). The first is used to learn the models. The second is used to test/validate the learned models. To ensure that this choice does not bias the models, the K-fold cross-validation methodology is used (Unpingco, 2016). This methodology consists of subdividing the data set into K partitions (folds). One of them is used for testing and the others for training. This is done K times, until all subsets have been used for testing. Finally, the average of the evaluation results is used to estimate the generalization ability of the learned model. The algorithms used to learn these models are quite diverse, for example: Logistic Regression, Naïve Bayes, Decision Tree, Artificial Neural Networks, K-Nearest Neighbors, Support Vector Machines, Artificial Immune System, among others (Haykin, 2009; Theodoridis and Koutroumbas, 2006; Bishop, 2006).

Naïve Bayes is one of the most popular machine learning algorithms (Reddy et al., 2022). According to Settouti et al. (2016), Naïve Bayes is one of the top performing algorithms for data mining. This algorithm has been used in the most diverse areas, such as: spam detection, product recommendations, medical diagnosis, identification software bugs, healthcare, cyber security, education, agriculture services, soil mapping, crop prediction and autonomous system (De Oliveira, Duarte, and Vieira Filho, 2022; Shreya et al., 2022; Wickramasinghe and Kalutarage, 2021; Yudhana, Sulistyo, and Mufandi, 2021; Priya, Ramesh, and Khosla, 2018). The Naïve Bayes algorithm has several advantages. Firstly, its simplicity and computational efficiency stand out (Kotsiantis, Zaharakis and Pintelas, 2006). Therefore, it is suitable for voluminous and real-time datasets. Furthermore, it is particularly useful when dealing with categorical or textual data. Another notable advantage is its ability to deal with imbalanced datasets, where classes have very different sizes. Furthermore, it is a good starting point in classification tasks, allowing rapid prototyping and performance benchmarking (Wickramasinghe and Kalutarage, 2021; Kotsiantis, Zaharakis and Pintelas, 2006). The models generated by the algorithm are transparent, as they are obtained from the probability distributions of the attributes (Al-Aidaroos, Bakar and Othman, 2010). Naïve Bayes is also robust to errors during execution (Wickramasinghe and Kalutarage, 2021). However, the algorithm also has some disadvantages, such as the assumption of conditional independence between the predictor variables, which can lead to suboptimal results when this assumption is not valid. However, in practice it has been

noticed that violating this assumption does not affect performance results greatly (Hand and Yu, 2001). Some authors have suggested that the hypothesis of conditional independence is a sufficient condition, but not necessary for the optimal application of the Naïve Bayes algorithm (Zhang, 2004; Hand and Yu, 2001; Rish, 2001; Domingos and Pazzani, 1996).

Learning the parameters of the Naïve Bayes algorithm is based only on the calculation of probabilities. Therefore, it is a low computational cost algorithm. It is based on two assumptions (John & Langley, 1995):

I) Bayes rule, i.e.,

$$p(y|\boldsymbol{x}) = p(\boldsymbol{x}|y)p(y)/p(\boldsymbol{x}),$$

where $p(\boldsymbol{x})$ and $p(y)$ are prior probabilities of features vector ($\boldsymbol{x} = [x_1, x_2, \dots, x_Q]$) and classes, respectively, and $p(y|\boldsymbol{x})$ and $p(\boldsymbol{x}|y)$ are posterior probabilities;

II) statistical independence of features, i.e.,

$$p(\boldsymbol{x}|y) = \prod_{q=1}^{Q} p(x_q|y),$$

where $x_q \in \mathbb{R}^N$ is the q-th feature with N samples.

One choice that must be made is the density function to be used to calculate $p(x_q|y_k)$ for each y_k class. If the probability distribution of the variable (feature) is not known *a priori* (or cannot be approximated by a known distribution such as the Gaussian which is often used), then the Kernel Density Estimation (KDE) method can be employed. It is a non-parametric method that estimates the probability distribution from the dataset, without any assumptions regarding this distribution. This flexibility makes KDE a very popular method (Chen, 2017). The objective of this method is to estimate a density function using a kernel function $K(t)$, such that

$$\hat{f}(x_q) = \frac{1}{Nh}\sum_{n=1}^{N} K\left(\frac{t - x_{qn}}{h}\right)$$

is an estimation of $f(x_q)$, where h is the smoothing parameter called bandwidth that influences in the shape of the estimated kernel. In addition, there is the hyperparameter "metric" that accepts distance measurements and is used by the Scikit-learn algorithm (Pedregosa et al., 2011). Tables 1 and 2 show respectively the main distance metrics and the most common kernels with their formulations (Pedregosa et al., 2011).

Table 1. Distance formulas, considering two vectors $\boldsymbol{x} = [x_1\ x_2\ \dots x_N]$ e $\boldsymbol{y} = [y_1\ y_2\ \dots y_N]$ in the N-dimentional space.

Distance name	**Formula**
Euclidian	$\text{dist}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{n=1}^{N}(x_n - y_n)^2}$
Manhattan	$\text{dist}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{n=1}^{N}\lvert x_n - y_n\rvert$
Chebyshev	$\text{dist}(\boldsymbol{x}, \boldsymbol{y}) = \max(\lvert x_1 - y_1\rvert, \lvert x_2 - y_2\rvert, \dots, \lvert x_N - y_N\rvert)$

Table 2. Kernel formulas. Source: (Węglarczyk, 2018).

Kernel name	**Formula**
Gaussian	$K(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(t-\mu)^2/2\sigma^2}$
Tophat	$K(t) = \frac{1}{2h}$
Epanechnikov	$K(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t\right)^2, & \lvert t\rvert < \sqrt{5} \\ 0, & \lvert t\rvert \geq \sqrt{5} \end{cases}$
Exponential	$K(t) = e^{-t}$
Linear	$K(t) = 1 - t$
Cosine	$K(t) = \cos\left(\frac{\pi}{2}t\right)$

Hyperparameters must be fixed before presenting data to the machine learning algorithm. That is, they are not actually learned, unlike the model parameters. However, you can use the data to check which set of hyperparameters provides the best results, fix them, and then learn the model parameters.

In summary, KDE smooths data points using a kernel function. It then sums the bumps to estimate the density of the data. Thus, regions with many observations will have a high-density value, while regions with few observations will have a low-density value (Chen, 2017). Figure 1 shows application of KDE (using a bandwidth equal to 0.5) to estimate a probability density function with a complex format as it has two modes (bimodal). To this end, two types of kernel functions were used, namely: Gaussian and Exponential, which generate very similar results for the example shown.

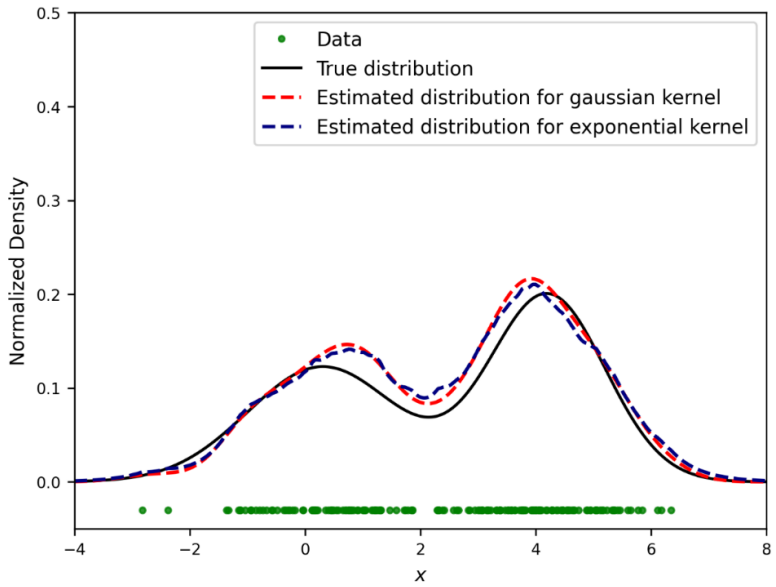

Figure 1. Application of the KDE method to estimate a probability density (black line) using the Gaussian (red dotted line) and Exponential (blue dotted line) kernel functions, for data points in green.

2.3 Proposed approach

First, the Shapiro-Wilk test is used to assess whether some of the variables have a Gaussian distribution, to verify the need of KDE use to obtain probability densities. Next, to implement KDE it is necessary to choose three hyperparameters: bandwidth, kernel and distance. So that this choice is not arbitrary, the Grid search algorithm (Pedregosa et al., 2011) is applied, which tests all combinations of these hyperparameters. For bandwidth, 100 values between 1 and 10^2 are tested. For kernel and distance, those in Tables 1 and 2 are respectively evaluated. Grouped cross-validation is used, ensuring that data from the same cultivar are not used in training and testing simultaneously, avoiding biasing the models. Five folds (subsets) are used in each of the 5 iterations, 4 of which are used for training and 1 for testing.

After obtaining the best set of hyperparameters, grouped cross-validation with 5 folds is used again to finally obtain the classification models. For each model obtained in each of the 5 iterations, accuracy (Acc), precision (Pr), and F1 score measures are calculated, according to the respective formulations:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, \quad Pr = \frac{TP}{TP+FP}, \text{ and } \quad F1 = \frac{2PrRe}{Pr+Re},$$

where TP, TN, FP and FN are true positive, true negative, false positive, and false negative, respectively, and $Re = TP/(TP + FN)$. The average confusion matrix is also calculated. Finally, the estimated probability densities for each class and variable are detailed.

3. Results and Discussion

In the previous section it was mentioned that the choice of the probability density function is essential in implementing the Naïve Bayes algorithm. The parameters of the classification models will be fitted according to this choice. One of the most used density functions is the Gaussian (Normal). To find out whether the variables have a normal distribution, the Shapiro-Wilk test is applied for each water stress class: "Control", "Moderate" and "Severe". Table 3 shows the statistic of the Shapiro-Wilk test, the p-values, the Fisher's kurtosis and the Skewness (based on Fisher-Pearson coefficient), for each variable, and the conclusion whether or not the variable has a normal distribution (column "Normal?"). These tests are performed for all data.

Table 3. Shapiro-Wilk test results for each variable, Fisher's kurtosis and the Skewness.

Class	Variable	Statistic	p-value	Fisher's kurtosis	Skewness	Normal?
Control	PH	0.7908	0.0000	2.1299	1.6312	No
	NT	0.9607	0.3845	0.0961	0.2456	Yes
	NGL	0.7624	0.0000	1.9722	1.6790	No
	RV	0.9610	0.3905	-0.8979	0.2423	Yes
	LA	0.9698	0.5982	-0.1728	-0.1925	Yes
	SDM	0.9557	0.2948	0.8316	0.8267	Yes
	RDM	0.9455	0.1669	0.4463	0.4410	Yes
Moderate	PH	0.7547	0.0000	2.3672	1.8044	No
	NT	0.8364	0.0006	5.4796	1.8327	No
	NGL	0.7437	0.0000	1.7697	1.6883	No
	RV	0.9758	0.7586	-0.2837	-0.2038	Yes
	LA	0.9672	0.5322	-0.7611	-0.0441	Yes
	SDM	0.9624	0.4203	-0.0966	-0.0242	Yes
	RDM	0.9496	0.2107	0.7465	0.6442	Yes
Severe	PH	0.9452	0.1644	-0.1216	0.7378	Yes
	NT	0.9735	0.6971	-0.4310	0.3314	Yes
	NGL	0.8424	0.0008	1.8199	1.4282	No
	RV	0.9684	0.5619	0.0894	0.5545	Yes
	LA	0.9299	0.0690	-0.6930	-0.5398	Yes

Class	Variable	Statistic	p-value	Fisher's kurtosis	Skewness	Normal?
	SDM	0.9122	0.0258	-0.9952	0.5094	No
	RDM	0.9774	0.7996	-0.3958	0.1258	Yes

"Normal?" column tells whether the distribution is normal or not, depending on the p-value.

From the results in Table 3, considering a significance level of 5% and that the null hypothesis is that the distribution of the variable is Gaussian, depending on the water stress class, we accept or reject this hypothesis, based on p-values. In addition to the p-values, Fisher's kurtosis and skewness statistics also inform about the distribution of the data. The closer its values are to zero, the more the distribution approaches Gaussian. Therefore, for variables whose column value "Normal?" in Table 3 is equal to "No", KDE should be used to model the probability density functions employed in the classification. For the other variables, the use of the Gaussian distribution for modeling is appropriate.

Using the Grid search algorithm for 5 folds returns the results in Table 4, where it appears that the "Euclidean" distance, the "Gaussian" kernel and the bandwidth equal to 1.0, generated the highest accuracy in the test data. The kernel choice is expected, as it was noted in Table 4 that most variables have a normal (Gaussian) distribution. Therefore, choosing the Gaussian kernel is the one that will bring the best results. Although KDE could be used exclusively for those variables that do not have a normal distribution, as seen in the results in Table 4 there is no consistency between the classes. In other words, some variables have a normal distribution for one class but not for another. Therefore, due to issues related to the computational implementation of machine learning algorithms, we chose to use KDE for all variables. However, as the kernel choice is the Gaussian model, for those variables with normal distribution, the KDE estimation will be very close to what we would obtain using the Gaussian model directly.

Table 4. Best hyperparameters obtained for each fold in grouped cross-validation. Accuracy is measured on the test set.

Fold	Bandwidth	Distance	Kernel	Mean accuracy ± Std.
1	1.0	Euclidean	Gaussian	0.7555 ± 0.0902
2	1.0	Euclidean	Tophat	0.3333 ± 0.0000
3	1.0	Euclidean	Epanechnikov	0.3333 ± 0.0000
4	1.0	Euclidean	Exponential	0.7111 ± 0.0888
5	1.0	Euclidean	Linear	0.3333 ± 0.0000

Std. means Standard deviation.

Table 4 shows only the average results of applying cross-validation on the training set considering non-overlapping groups (cultivars). To check how the hyperparameters change, two of the best hyperparameters are fixed and the other varies. For each tested value, the accuracy is computed. Figures 2, 3, and 4 display the average results of this implementation for the training and testing sets separately. In all cases considering grouped cross-validation with 5 folds.

In Figure 2 it is observed that the accuracy reaches the maximum (0.7666) in the test set when the bandwidth is equal to 1.1497. While the accuracy on this set fluctuates from this value, for the training set it always decreases. This value is different from that presented in Table 4, because the table shows the average value among the folds.

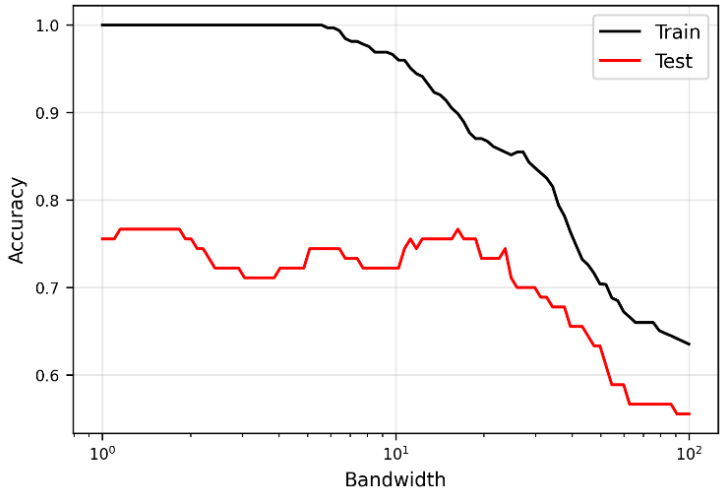

Figure 2. Graph of bandwidth versus accuracy for training (Train) and test (Test) data, keeping the other hyperparameters fixed.

From Figures 3 and 4 it is observed that any choice of kernel and distance metric results in maximum accuracy in the training set. On the other hand, in the test set the Gaussian kernel and the Chebyshev metric present higher accuracies than the other choices. What is new in this result is the Chebyshev metric that did not appear in Table 4. It is noted that in the test set, it results in slightly higher accuracy than the Euclidean distance metric, with a value of 0.7777.

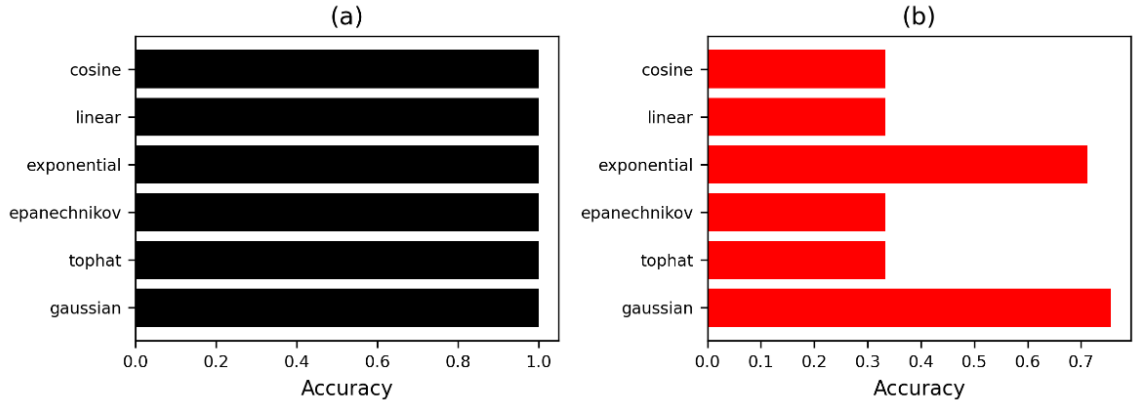

Figure 3. Accuracy for different kernel choices, for the (a) training and (b) test sets, keeping the other hyperparameters fixed, according to Table 4.

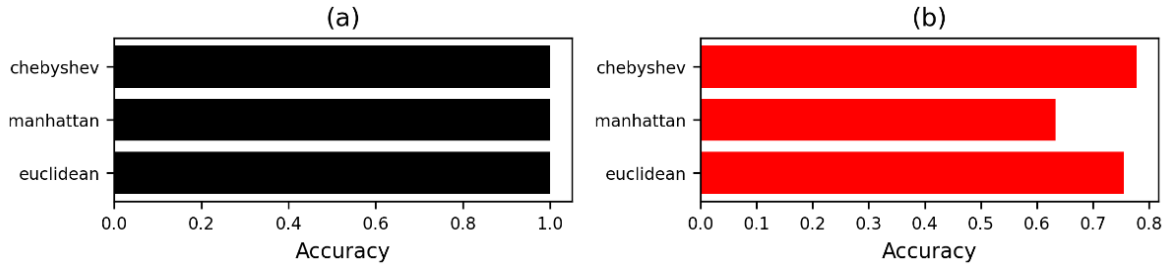

Figure 4. Accuracy for different distance function choices, for the (a) training and (b) test sets, keeping the other hyperparameters fixed, according to Table 4.

Based on the results presented in Table 4 and Figures 2, 3 and 4, we chose to set the hyperparameters as: bandwidth at 1.1497, Gaussian kernel and the Chebyshev metric as a distance measure.

The boxplot graph in Figure 5 displays the performance measures for the training [Figure 5 (a)] and test [Figure 5 (b)] sets, when employing grouped cross-validation with 5 folds. From this graph it can be seen that the patterns in the training set were learned perfectly, regardless of the data used (fold). On the other hand, as in the test set, we have values varying from 0.66 to 0.91, depending on the chosen fold, showing that the generalization capacity is variable.

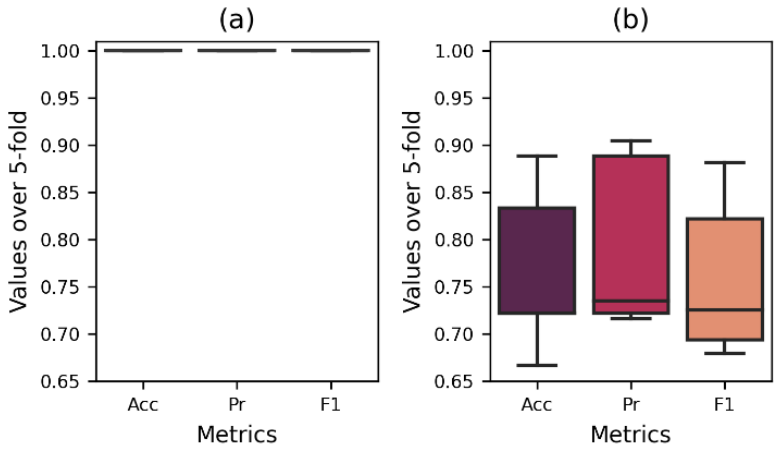

Figure 5. Boxplot showing the accuracy (Acc), precision (Pr) and F1 score measurements for the (a) training and (b) test sets.

This variation in the test set can be explained by the fact that certain cultivars present unrepresentative patterns. As the grouped cross-validation approach was used, the patterns of some cultivars were not adequately learned in the training stage. Table 5 details the cultivars used in the training and testing stages, and displays the accuracy of this last stage. The "Errors in Test" column displays the cultivar and the number of prediction errors. It can be seen that there was a decrease in accuracy for fold 5, as in this case only the 9 samples from "Marandu" cultivar were used for testing. It can also be seen that the "Pojuca" cultivar was the one that generated the most prediction errors, indicating that the patterns of this cultivar were not learned adequately from the other cultivars employed in training stage.

Table 5. Cultivars used in the training and testing stages and results in the test stage.

Fold	Training	Test	Accuracy (Test)	Errors in Test
1	ADR 300, Aruana, BRS Piatã, Marandu, Mombaça, Pojuca, Tanzânia	Comum, Xaraés	0.8889	Comum: 1 Xaraés: 1
2	ADR 300, Aruana, Comum, Marandu, Mombaça, Pojuca, Xaraés	BRS Piatã, Tanzânia	0.7222	BRS Piatã: 2 Tanzânia: 3
3	ADR 300, BRS Piatã, Comum, Marandu, Mombaça, Tanzânia, Xaraés	Aruana, Pojuca	0.7222	Aruana: 1 Pojuca: 4
4	Aruana, BRS Piatã, Comum, Marandu, Pojuca, Tanzânia, Xaraés	ADR 300, Mombaça	0.8333	ADR 300: 3 Mombaça: 0
5	ADR 300, Aruana, BRS Piatã, Comum, Mombaça, Pojuca, Tanzânia, Xaraés	Marandu	0.6667	Marandu: 3

In the results presented by De Oliveira et al. (2023b) the authors conclude that the "ADR 300" cultivar presents less variation in water stress environments, when compared to the "Control" water regime. Meanwhile, the "Tanzânia" cultivar is the one that presents the greatest change.

Regardless of the weights assigned to stressed environments in TOPSIS, the "Tanzânia" is always the one that most varies the most, while "ADR 300" is always the one that varies the least, in relation to the "Control" water regime. The "Pojuca" and "Marandu" cultivars alternate between second and third position, depending on the weights set. While the cultivars "Comum", "Pojuca", "Marandu" and "Tanzânia" present extreme values. These results partially explain the errors returned in the test set (last column of Table 2). Because, these cultivars present patterns that are more difficult to learn than other cultivars.

The average confusion matrices in Figure 6 show the comparison of predictions between classes. In Figure 6 (a) it can be seen that there was no confusion in the classifications, because as previously observed, in the training stage, the patterns were learned perfectly. But in the testing stage, Figure 6 (b), it is observed that there was greater confusion in classifying the examples into the "Control" and "Moderate" classes. On average, 1.6 examples from the "Control" class were classified as being from the "Moderate" class. On the other hand, no example of the "Control" class was mistakenly classified as the "Severe" class. While, on average 0.6 of the examples from the "Moderate" class were misclassified as being from the "Control" or "Severe" classes. These results are consistent with the characteristics of soil water regimes, as the "Moderate" regime is intermediate between "Control" and "Severe". Therefore, the changes caused in the measured variables, and consequently in the modeled patterns, partly reflect this relationship.

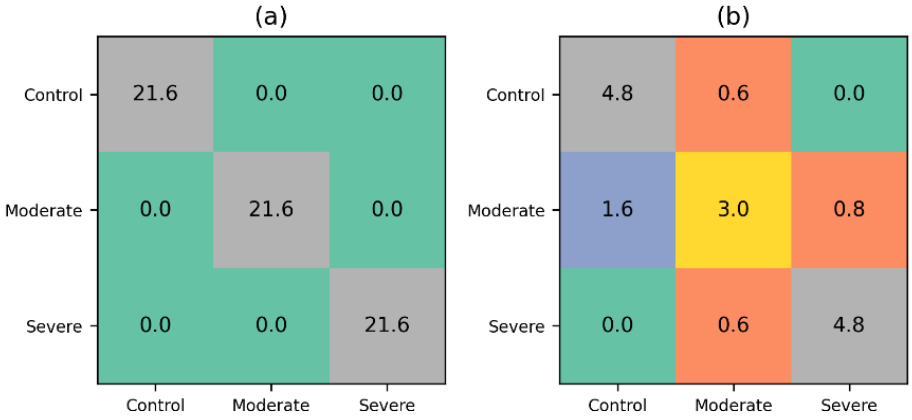

Figure 6. Average confusion matrix for the 5 folds, showing the comparison of predictions per class, in the (a) training (b) testing steps.

Finally, Figure 7 shows the probability densities obtained for each variable and each class, separately. It is noted that the estimated densities for some variables have complex forms. For example, for the PH variable, probability densities with two main lobes (bimodal) are noted for the "Control" and "Moderate" water stress classes, as according to Table 3 (Shapiro-Wilk test), this variable does not have a normal distribution for these classes. Furthermore, the values of Fisher's kurtosis and Skewness statistics are further from zero.

Another example is the NGL variable. It has a probability distribution different from Gaussian for all classes of water stress or control (see Table 3). Analyzing the probability densities estimated by KDE, it is observed that for the "Control" class the density has a heavy right tail. While for the other classes there is another main lobe around the value 100. And, regardless of class, all estimated densities are asymmetric. This fact is also shown in Table 3 by the Fisher's kurtosis and Skewness, whose values are greater than 1.4 for all classes.

An interesting result can be observed for the NT variable. According to Table 3, it has a normal distribution for the "Control" and "Severe" classes only. For the "Moderate" class, the Fisher's kurtosis value is the highest calculated, far exceeding the values for the other variables (regardless of the class). Analyzing the estimated densities in Figure 7, for the NT variable, we

note that this is explained due to its bimodal distribution. Because there is another lobe around the value 50. Therefore, it is a leptokurtic distribution. The values of this second mode (or outliers) may be due to some intrinsic characteristic of the "Moderate" water regime, or also noise in the data, requiring additional research to conclude. These results (and others that can be seen in Figure 7) show the importance of using KDE to estimate probability densities, mainly for those variables that do not have a normal distribution, or close to it.

As mentioned by de De Oliveira et al. (2023a) the machine learning models obtained can be used to check soil water conditions just by analyzing the variables measured from plants. Thus, with the models detailed in Figure 7, it is possible to estimate whether a given sample was obtained from a soil that has suffered severe or moderate water stress, or no water stress (control class). This can also be done using just one of the variables, whichever is easier to measure. Of course, in this case the forecast may be reduced depending on the pattern analyzed.

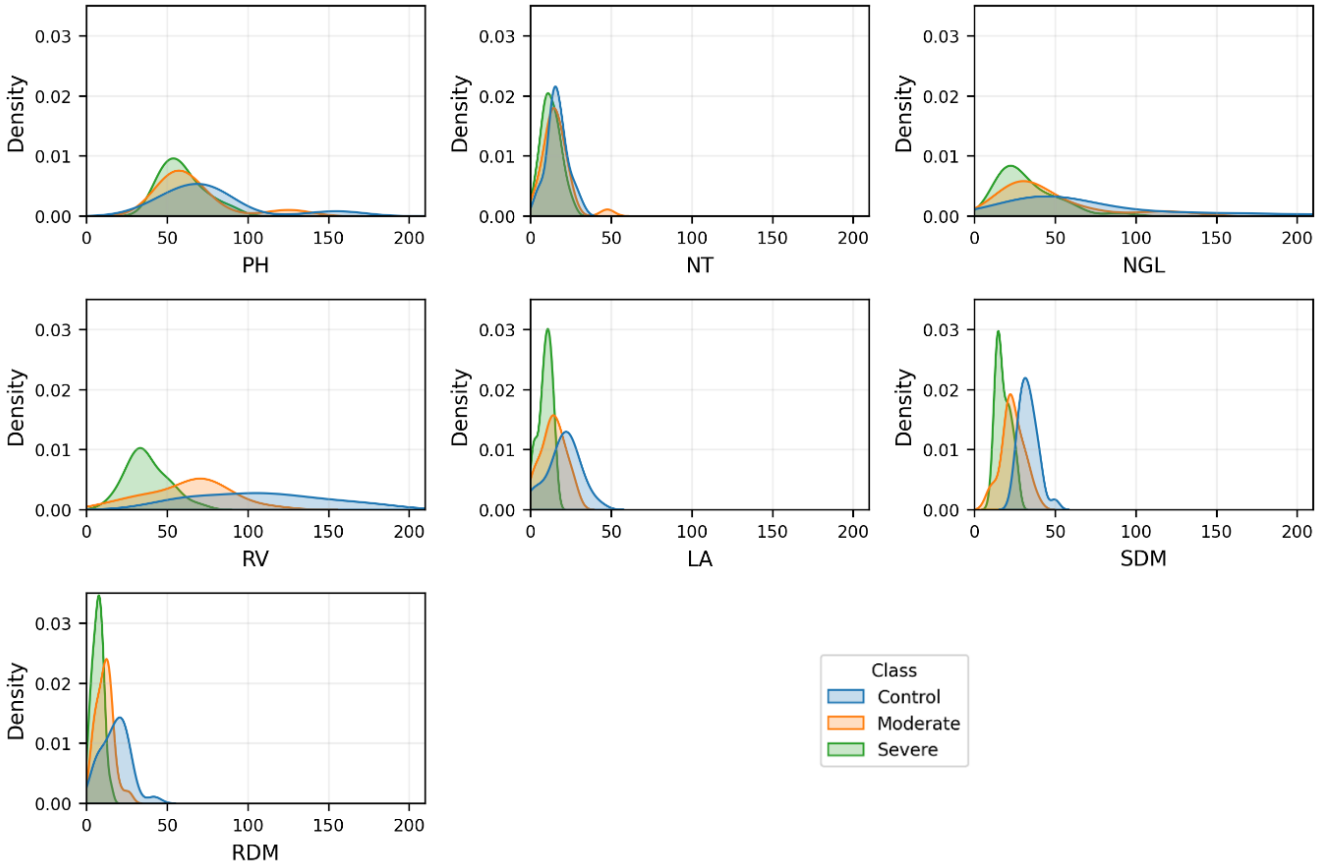

Figure 7. Probability densities obtained for each variable and each class, using grouped cross-validation with 5 folds.

The performance measures accuracy, precision and F1-score have maximum values equal to 0.88, 0.90 and 0.88, respectively, for fold 1 (see Table 2 and Figure 5). On the other hand, the lowest values obtained were 0.66, 0.72, 0.67, respectively, for fold 5. These results show significant variation depending on the choice of training and testing data. Therefore, to achieve better results, more data would be necessary, as in the used database there are only three instances of each cultivar in each water stress or control environment. Another way to try to improve performance consists of using other types of machine learning algorithms, comparing with the results obtained here.

Finally, the models obtained here can be used to classify forage samples, for the cultivars studied here, to understand which soil water regimes (related to water stress) they come from. And this can be done with a precision ranging from 72% to 90%, depending on the used cultivar. As the

models obtained are based on probability distribution, it can also be assessed whether the cultivar comes from an intermediate stress water regime.

4. References

Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2010). Naive Bayes variants in classification learning. In 2010 international conference on information retrieval & knowledge management (CAMP) (pp. 276-281). DOI: 10.1109/INFRKM.2010.5466902

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. Sensors, 21(11), 3758. DOI: 10.3390/s21113758

Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

Chen, Y. C. (2017). A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology, 1(1), 161-187. DOI: 10.1080/24709360.2017.1396742

de Oliveira, B. R., Duarte, M. A. Q., Vieira Filho, J. (2022). Premature ventricular contraction recognition using blind source separation and ensemble gaussian naive bayes weighted by analytic hierarchy process. Acta Scientiarum. Technology, 44, e60386-e60386. DOI: 10.4025/actascitechnol.v44i1.60386

de Oliveira, B. R., Zuffo, A. M., Steiner, F., Aguilera, J. G., & Gonzales, H. H. S. (2023a). Classification of soybean genotypes during the seedling stage in controlled drought and salt stress environments using the decision tree algorithm. Journal of Agronomy and Crop Science. DOI: 10.1111/jac.12654

de Oliveira, B. R., Duarte, M. A. Q., Zuffo, A. M., Steiner, F., Aguilera, J. G., Dutra, A. F., Alcântara Neto, F., Leite M. R. L., Silva, N. S. G., Salcedo, E. P., Morales-Aranibar, L., Chura, R. C. A., & Contreras, W. C. (2023b). Selection of forage grasses for cultivation under water-limited conditions using Manhattan distance and TOPSIS. Plos One. [Manuscript not yet published, but accepted for publication]. DOI: 10.1371/journal.pone.0292076

Domingos, P., & Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple bayesian classi er. In Proc. 13th Intl. Conf. Machine Learning (pp. 105-112).

FAO (2023). FAOSTAT. Available at: https://www.fao.org/faostat/en/#data/QCL [Accessed October 17, 2023].

Farshadfar, E., PoursiahbidI, M. M., & Abooghadareh, A. R. P (2012). Repeatability of drought tolerance indices in bread wheat genotypes. Intl. J. Agri. Crop Sci. 4, 891–903.

Feltran-Barbieri, R., & Féres, J. G. (2021). Degraded pastures in Brazil: improving livestock production and forest restoration. Royal Society Open Science, 8: e201854. DOI: 10.1098/rsos.201854

Freitas, E. N. D., Alnoch, R. C., Contato, A. G., Nogueira, K. M. V., Crevelin, E. J., Moraes, L. A. B. D., Silva, R. N., Martínez, C. A., & Polizeli, M. D. L. T. (2021). Enzymatic pretreatment with laccases from Lentinus sajor-caju induces structural modification in lignin and enhances the digestibility of tropical forage grass (Panicum maximum) grown under future climate conditions. International Journal of Molecular Sciences, 22(17), 9445. DOI: 10.3390/ijms22179445

Ghimire, S. R., Njarui, D. M., Mutimura, M., Cardoso Arango, J. A., Johnson, L., Gichangi, E., & Djikeng, A. (2015). Climate-smart Brachiaria for improving livestock production in East Africa: Emerging opportunities.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all?. International statistical review, 69(3), 385-398. DOI: 10.1111/j.1751-5823.2001.tb00465.x

Habermann, E., Dias de Oliveira, E. A., Contin, D. R., Delvecchio, G., Viciedo, D. O., de Moraes, M. A., Prado, R. M., Costa, K. A. P., Braga, M. R., & Martinez, C. A. (2019). Warming and water deficit impact leaf photosynthesis and decrease forage quality and digestibility of a C4 tropical grass. Physiologia Plantarum, 165(2), 383-402. DOI: 10.1111/ppl.12891

Haykin, S. S. (2009). Neural networks and learning machines. 3ª ed. Pearson.

Iqbal, M. A., Abdul, H., Muzammil, H. S., Imtiaz, H., Tanveer, A., Saira, I., & Anser, A. (2019). A meta-analysis of the impact of foliar feeding of micronutrients on productivity and revenue generation of forage crops. Planta Daninha, 37. DOI: 10.1590/S0100-83582019370100046

Jank, L., Santos, M. F., do Valle, C. B., Barrios, S. C., & Simeão, R. M. (2021). Forage Genetic Resources in Brazil. In: XXIV International Grassland Congress/XI International Rangeland Congress (Sustainable Use of Grassland and Rangeland Resources for Improved Livelihoods), 2021.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, 338-345.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26, 159-190. DOI: 10.1007/s10462-007-9052-3

Lobato, J. F. P., Freitas, A. K., Devincenzi, T., Cardoso, L. L., Tarouco, J. U., Vieira, R. M., Dillenburg, D. R., & Castro, I. (2014). Brazilian beef produced on pastures: Sustainable and healthy. Meat science, 98(3), 336-345. DOI: 10.1016/j.meatsci.2014.06.022

Luiz Piati, G., Ferreira de Lima, S., Lustosa Sobrinho, R., dos Santos, O. F., Vendruscolo, E. P., Jacinto de Oliveira, J., do Nascimento de Araújo, T. A., Mubarak Alwutayd, K., Finatto, T., & AbdElgawad, H. (2023). Biostimulants in Corn Cultivation as a Means to Alleviate the Impacts of Irregular Water Regimes Induced by Climate Change. Plants 12, 2569. https://doi.org/10.3390/plants12132569

Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. D. (2021). Machine learning in agriculture domain: A state-of-art survey. Artificial Intelligence in the Life Sciences, 1, 100010. DOI: 10.1016/j.ailsci.2021.100010

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830.

Priya, R., Ramesh, D., Khosla, E. (2018). Crop prediction on the region belts of India: a Naïve Bayes MapReduce precision agricultural model. In 2018 international conference on advances in computing, communications and informatics (ICACCI) (pp. 99-104). DOI: 10.1109/ICACCI.2018.8554948

Reddy, E. M. K., Gurrala, A., Hasitha, V. B., & Kumar, K. V. R. (2022). Introduction to Naive Bayes and a Review on Its Subtypes with Applications. Bayesian Reason. Gaussian Process. Mach. Learn. Appl, 1-14.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence. (3)22, 41-46.

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. IEEE Access, 9, 4843-4873. DOI: 10.1109/ACCESS.2020.3048415

Silva, R. D. O., Barioni, L. G., & Moran, D. (2021). Fire, deforestation, and livestock: When the smoke clears. Land Use Policy, 100, 104949. DOI: 10.1016/j.landusepol.2020.104949

Simeão, R. M., Resende, M. D., Alves, R. S., Pessoa-Filho, M., Azevedo, A. L. S., Jones, C. S., & Machado, J. C. (2021). Genomic selection in tropical forage grasses: Current status and future applications. Frontiers in Plant Science, 12, 665195. DOI: 10.3389/fpls.2021.665195

Silva, R. O., Barioni, L. G., Hall, J. A. J., Matsuura, M. F., Albertini, T. Z., Fernandes, F. A., & Moran, D. (2016). Increasing beef production could lower greenhouse gas emissions in Brazil if decoupled from deforestation. Nat. Clim. Change 6, 493–497. doi: 10.1038/nclimate2916

Steiner, F., Zuffo, A. M., Silva, K. C., Lima, I. M. O., & Ardon, H. J. V. (2020). Cotton response to nitrogen fertilization in the integrated crop-livestock system. Scientia Agraria Paranaensis, 19(3), 211-220. DOI: 10.18188/sap.v19i3.24349

Shreya, S., Sushmitha, S. R., Vaanathe, L. R., & Madhumathi, R. (2022). Agro World: A Naive Bayes based System for Providing Agriculture as a Service. In 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1871-1875). DOI: 10.1109/ICICCS53718.2022.9788290

Theodoridis, S., & Koutroumbas, K. (2006). Pattern recognition. 2ª ed. Elsevier.

Tomlinson, I. (2013). Doubling food production to feed the 9 billion: A critical perspective on a key discourse of food security in the UK. J. Rural Stud. 29, 81–90. DOI: 10.1016/j.jrurstud.2011.09.001

Unpingco, J. (2016). Python for probability, statistics, and machine learning (Vol. 1). Cham, Switzerland: Springer International Publishing.

Zhang, H. (2004). The optimality of naive Bayes. American Association for Artificial Intelligence, 1(2), 3.

Zuffo, A. M., Steiner, F., Aguilera, J. G., Ratke, R. F., Barrozo, L. M., Mezzomo, R., Santos, A. S. d., Gonzales, H. H. S., Cubillas, P. A., & Ancca, S. M. (2022). Selected indices to identify water-stress-tolerant tropical forage grasses. Plants, 11(18), 2444. DOI: 10.3390/plants11182444

Węglarczyk, S. (2018). Kernel density estimation and its application. In ITM Web of Conferences: XLVIII Seminar of Applied Mathematics, 23, 00037. DOI: 10.1051/itmconf/20182300037

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), 2277-2293. DOI: 10.1007/s00500-020-05297-6

Yudhana, A., Sulistyo, D., & Mufandi, I. (2021). GIS-based and Naïve Bayes for nitrogen soil mapping in Lendah, Indonesia. Sensing and Bio-Sensing Research, 33, 100435. DOI: 10.1016/j.sbsr.2021.100435

6. Additional Information

6.1 Acknowledgments

We thank the State University of Mato Grosso do Sul (UEMS), Cassilândia Unit, and the researchers Zuffo et al. (2022) for making the database available.

6.2 Funding

We do not receive resources from any financing fund.

6.3 Conflicts of Interest

There are no conflicts of interest.

6.4 Complementary Material

The Python scripts used to develop the work presented, as well as the data, are available on github via the link https://github.com/brunobro/classification-of-tropical-forage-grass-using-na-ve-bayes-and- kernel-density-estimation