ISSN: 2317-0840

Comparação de dois classificadores na identificação de espécies arbóreas baseada em caracteres contínuos

Giovani Festa Paludo[1†], Júlio Sílvio de Sousa Bueno Filho[2]

[1]*Doutorando no Programa de Pós Graduação em Estatística e Experimentação Agropecuária, Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Lavras (UFLA)*
[2]*Departamento de Estatística, Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Lavras (UFLA)*

Resumo: *A atribuição de um nome de espécie a um ramo vegetal é o processo denominado de identificação botânica e consiste em um problema de classificação. A inclusão de variáveis contínuas nesse processo de identificação não é recente, mas está em plena expansão. O objetivo do presente estudo foi avaliar dois classificadores que atribuem um nome de espécie a um conjunto de medidas contínuas obtidas em uma folha. Foram coletadas 352 folhas de 5 espécies da família botânica Myrtaceae e foram mensuradas 5 variáveis manualmente: largura máxima do limbo, largura do pecíolo, comprimento máximo da folha, do limbo e do pecíolo. Dois classificadores foram utilizados: florestas aleatórias (FA) e a análise discriminante linear (ADL). O conjunto de dados foi separado em treino (70%) e teste (30%), e depois realizadas 2000 interações para cada umas das 31 possíveis combinações das variáveis. Os modelos e classificadores foram comparados utilizando a taxa média de acertos no conjunto teste nas 2000 interações. Quando consideradas todas as variáveis o classificador ADL acertou 98,2% enquanto que o FA acertou 96,8% das classificações. A variável isolada com maior taxa média de acertos foi o comprimento máximo do pecíolo e quando combinada duas variáveis, o comprimento máximo do pecíolo e a largura máxima do limbo foram as que mais acertaram as classificações. Na maioria das combinações de variáveis o classificador ADL apresentou maiores taxas médias de acertos. Resultados que mostram o potencial que este tipo de modelo tem para contribuir como auxílio ao processo de identificação botânica.*

Palavras-chave: *Problema taxonômico; aprendizado supervisionado; morfometria; taxonomia numérica; normal multivariada.*

Comparison of two classifiers in the identification of tree species based in continuous characters

Abstract: *The species name assignment to a vegetable branch is the process called botanical identification and consists in a classification problem. The inclusion of continuous variables in this problem is not new, but it's a growing topic. The objective of this paper was evaluate two classifiers that assign a specific name to a set of continuous measures of a leaf. There were collected 352 leaves from 5 species of Myrtaceae botanical family. There were measured 5 variables manually: maximum blade width, petiole width, leaf, blade and petiole maximum length. There were utilized two classifiers linear discriminant analysis (LDA) and random forests (RF). The data set was divided into train (70%) and test (30%) and 2000 iterations were conducted to each of 31 possible combination of variables. The models and classifiers were compared by the mean of the successful classification rate in the test set obtained in the 2000 iterations. As a result, considering all variable combinations, the LDA accuracy was 98.2% while the RF classified 96.8% correctly. The best isolated variable was the petiole maximum length and the best combination of two variables was the petiole maximum length and blade maximum width. LDA had a better performance than RF in the greater part of the variables combinations. These findings show the potential that this approach has to contribute as an aid to the botanical identification process.*

Keywords: *Taxonomic problem; supervised learning; morphometrics; numeric taxonomy; multivariate normal.*

[†]Autor correspondente: gfpaludo@gmail.com

Manuscrito recebido em: 29/07/2024
Manuscrito revisado em: 29/09/2024
Manuscrito aceito em: 30/09/2024

Introdução

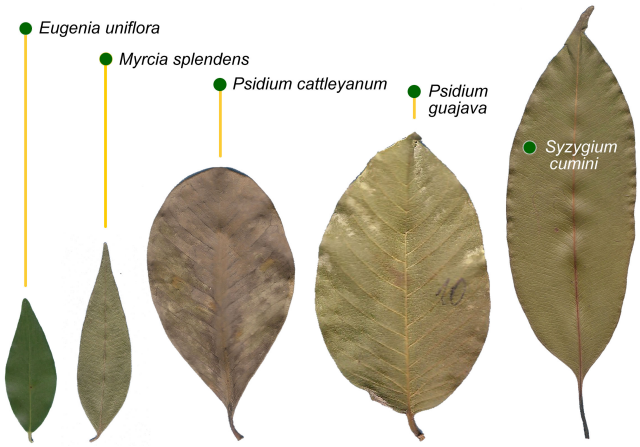
A atribuição de um nome de espécie a um ramo vegetal é o processo que recebe o nome de identificação e consiste em um problema de classificação. Problemas de classificação podem ser trabalhados com técnicas de aprendizado de máquina (*e.g.* MURPHY, 2022) ou também com métodos estatísticos clássicos como a análise discriminante linear, análise discriminante quadrática, entre outras (MARKS; DUNN, 1974; JOHNSON; WICHERN, 2014). A função discriminante linear foi sugerida na década de 30 por Fisher (1936) para ser utilizada no problema de identificação de espécies. Porém, foi apenas recentemente que as aplicações desse tipo de metodologia no problema de identificação estão se tornando efetivamente viáveis para auxiliar na identificação sendo que há um aumento expressivo no número de publicações e aplicações (NASIR *et al.*, 2014, LYSKO *et al.*, 2022; THANIKKAL; DUBEY; THOMAS, 2023; LUNA-BONILLA *et al.*, 2024). Acreditamos que isso se deve especialmente ao aumento da capacidade de processamento e armazenamento, à contínua expansão de métodos computacionais, à melhoria das técnicas de processamento de imagens, à popularização de câmeras de alta resolução com os smartphones, entre outros. Este é um tópico de grande interesse e há diversas áreas envolvidas no desenvolvimento dessas metodologias: desde a área de medidas (ELEN; AVUÇLU, 2021), estatística (MATTHEWS *et al.*, 2018), ciências de plantas (OSO; JAYEOLA, 2021), sistemas de informação (COPE *et al.*, 2012), computação na agricultura (YIGIT *et al.*, 2019), ecologia (BARRÉ *et al.*, 2017), entre outras. Portanto, estudos que levantem conjunto de variáveis e que avaliem o quanto cada variável pode contribuir na classificação são importantes, bem como os estudos que identifiquem as melhores técnicas para este fim.

O objetivo do presente estudo foi avaliar dois classificadores que atribuem uma população (*i.e.*, espécie) a um indivíduo (*i.e.*, folha) com base em medidas de tamanho da folha.

Material e Métodos

Foram coletadas, respectivamente, 72, 66, 72, 67 e 75 folhas de 5 indivíduos arbóreos, cujos indivíduos pertencem a uma espécie diferente da família botânica Myrtaceae: *Eugenia uniflora* L. (pitangueira), *Myrcia splendens* (Sw.) DC. (guamirim), *Psidium cattleyanum* Sabine (araçazeiro), *P. guajava* L. (goiabeira), *Syzygium cumini* (L.) Skeels (jambolão) (Figura 1). Para cada folha, foram observadas 5 variáveis de tamanho: comprimento máximo da folha (f_{cm}), comprimento máximo do limbo (l_{cm}), largura máxima do limbo (l_{lm}), comprimento máximo do pecíolo (p_{cm}) e largura do pecíolo logo antes do início do limbo (p_{la}).

Figure 1: Illustration of a leaf from each of the 5 species belonging to the Myrtaceae botanical family used in this study.



Source: from the authors (2024).

Cabe ressaltar que a soma dos valores das variáveis p_{cm} e l_{cm} de uma folha tendem a serem iguais ao valor f_{cm} quando o início do limbo acontece no mesmo ponto do lado direito e do lado esquerdo da folha. Quando a folha não é simétrica em relação ao começo do limbo, a soma da variável p_{cm} e l_{cm} ultrapassa o valor do f_{cm}, uma vez que são observados sempre o comprimento máximo de cada variável.

Com exceção do *S. cumini*, as espécies são nativas do Brasil. As observações foram feitas em folhas de cada um dos 5 indivíduos arbóreos, localizados no câmpus na Universidade Federal de Lavras (UFLA), município de Lavras, estado de Minas Gerais, Brasil. Ainda, *M. splendens* situa-se na borda de uma porção de mata nativa, enquanto que *S. cumini* está próxima a área de mata. As outras três espécies estão na arborização urbana do câmpus.

Na classificação, considera-se cada folha como um indivíduo e deseja-se classificar a folha em uma das populações $\pi_1, ..., \pi_m$, isto é, uma espécie botânica. A $j-$ésima espécie tem um conjunto de $1, ..., N_j$ folhas, $j = 1, ..., m$. E em cada folha são observadas $1, ..., p$ variáveis. O vetor de observações da $i-$ésima folha a qual não se conhece a população é denotada por $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^\top$. O vetor de observações de uma folha que veio da população π_j é denotada por $\boldsymbol{x}_i^{(j)} = (x_{i1}^{(j)}, ..., x_{ip}^{(j)})^\top$, $i = 1, ..., N_j$, $j = 1, ..., m$. Já $\boldsymbol{X}_{N_j p}^{(j)}$ é o conjunto de folhas que veio da população π_j, isto é,

$$\boldsymbol{X}_{N_j p}^{(j)} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_j 1} & x_{N_j 2} & \cdots & x_{N_j p} \end{pmatrix}.$$

Assim $\boldsymbol{X}_{N_1 p}^{(1)}, ..., \boldsymbol{X}_{N_m p}^{(m)}$ correspondem ao conjunto de observações das folhas em cada uma das $\pi_1, ..., \pi_m$ populações.

O primeiro classificador utilizado foi a análise discriminante linear e foi calculada com base na função `lda()` do pacote `MASS` (VENABLES; RIPLEY, 2002) nas configurações padrão. Já o classificador de florestas aleatórias (FA) baseou-se no algoritmo implementando no pacote `randomForest` do R (LIAW; WIENER, 2002) nas configurações padrão. Ambos pacotes são do software R (R CORE TEAM, 2023).

Tanto para FA quanto para ADL foram testadas todas as 31 possíveis combinações das 5 variáveis. As comparações entre FA e ADL e entre as 31 combinações de variáveis foi realizada por meio da taxa média de acertos, cujas folhas foram divididas em conjunto treino (70%) e conjunto teste (30%). O conjunto treino foi utilizado para treino dos modelos e o conjunto teste foi utilizado para obter as taxas de acertos. A taxa de acertos é 0 quando erra todas as classificações e 100% quando acerta todas as classificações. A cada separação aleatória do conjunto de dados entre treino e teste, ajustou-se o modelo de ADL e FA, e calculou-se uma taxa de acertos no conjunto teste. Esse procedimento de separação aleatória foi realizado 2000 vezes e, portanto, a taxa de acertos apresentada no artigo é a média da taxas de acertos obtidas no conjunto teste das 2000 separações aleatórias do conjunto de dados em treino e teste. Este procedimento é necessário para reduzir a variabilidade na taxa de acertos do conjunto teste, uma vez que a taxa de acertos oscila dependendo de quais observações que foram selecionadas para treino ou teste.

Resultados e Discussão

O conjunto de dados está descrito em termos de média e desvio padrão amostral (Tabela 1), sendo que as folhas com maior média de comprimento máximo do limbo (l_{cm}) e média de largura máxima do limbo (l_{lm}) foram *S. cumini* e *P. guajava*, respectivamente.

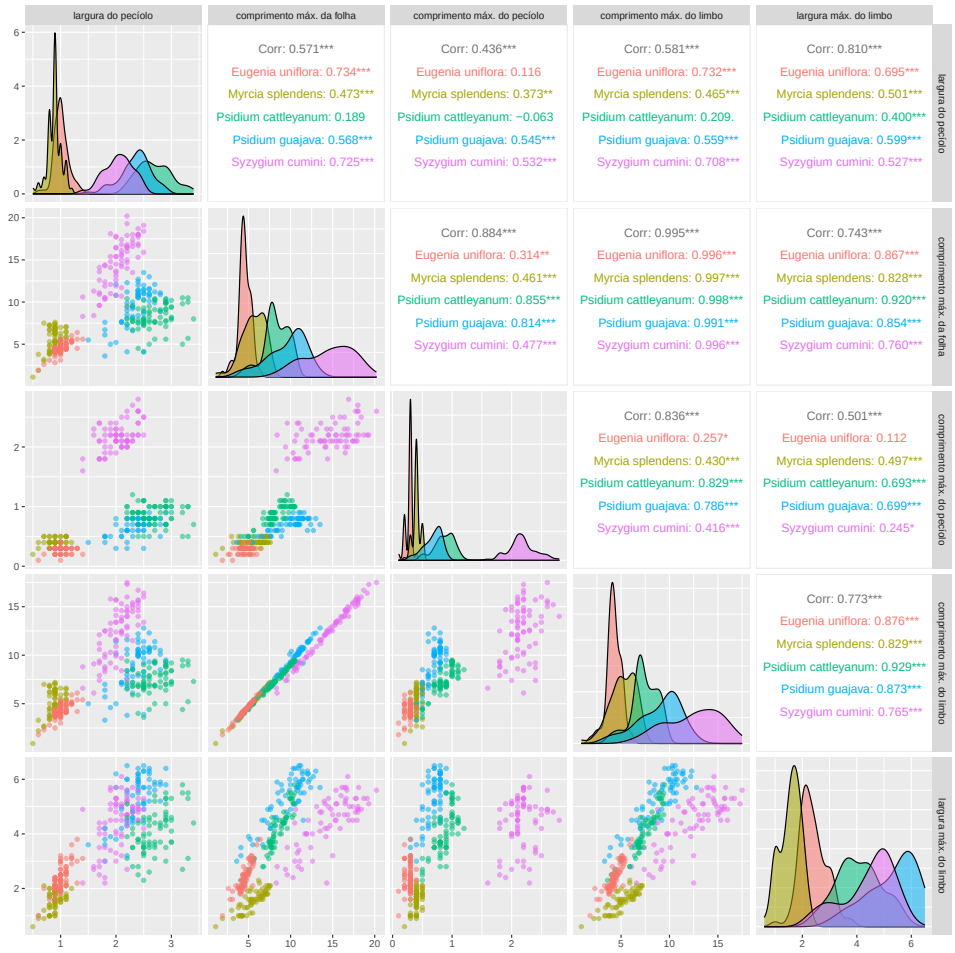
Table 1: Mean (standard deviation) in centimeters of the variable values obtained for each species.

Espécie	p_{la}	f_{cm}	p_{cm}	l_{cm}	l_{lm}
Eugenia uniflora	0,10 (0,01)	4,5 (0,8)	0,3 (0,1)	4,3 (0,8)	2,4 (0,5)
Myrcia splendens	0,09 (0,01)	5,6 (1,4)	0,4 (0,1)	5,2 (1,3)	1,6 (0,4)
Psidium cattleyanum	0,27 (0,03)	8,3 (1,5)	0,9 (0,2)	7,4 (1,4)	4,1 (0,8)
Psidium guajava	0,24 (0,03)	9,6 (2,3)	0,7 (0,2)	9,0 (2,2)	5,3 (0,9)
Syzygium cumini	0,20 (0,03)	14,6 (2,9)	2,2 (0,2)	12,5 (2,8)	4,4 (1,0)
Global	0,18 (0,08)	8,6 (4,1)	0,9 (0,7)	7,7 (3,5)	3,6 (1,5)

Source: from the authors (2024).

São apresentadas as formas da distribuição, dispersões e valores de correlações das 5 variáveis das espécies de Myrtaceae (Figura 2). Quando avaliamos visualmente as funções de densidade estimadas por kernel, foi possível observar no gráfico com as densidades da variável p_{la}, uma separação onde *M. splendens* e *E. uniflora* ficaram à esquerda e as demais espécies à direita. Também foi possível observar uma separação nas funções de densidade estimada por kernel da variável p_{cm} entre *S. cumini* que ficou separada à direita e as demais espécies que ficaram à esquerda do gráfico. Ainda, as variáveis com maior grau de correlação foram comprimento máximo do limbo (l_{cm}) e comprimento máximo da folha (f_{cm}) (Figura 2). Isto se deve ao fato que a soma do p_{cm} com l_{cm} é aproximadamente o mesmo valor do f_{cm}.

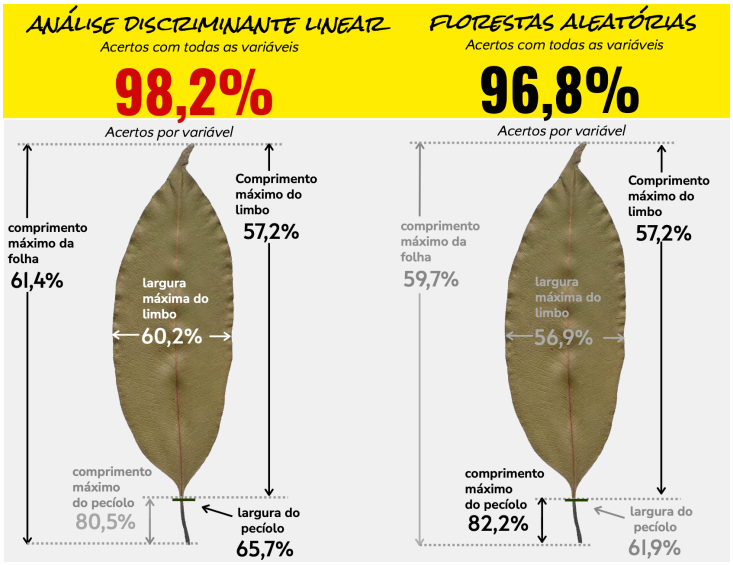
Figure 2: Scatter plots, correlations and density functions estimated by kernel to represent the density function of each of the 5 variables of the 5 plant species.



Source: from the authors (2024).

Das 31 combinações de variáveis testadas em cada classificador, em 21 delas a ADL apresentou maiores taxas médias de acertos e em 8 delas o FA apresentou maiores taxas médias de acertos, sendo que em 2 combinações ADL e FA tiveram a mesma taxa de acertos (Tabela 2). A média de acertos para as 31 combinações de variáveis foi de 87.2% para a ADL e 85.7% para o RF. A taxa global de acertos foi maior para a ADL (Figura 3), isto é, ao se utilizar todas as variáveis o classificador ADL acertou 98,2% enquanto que o FA acertou 96,8%.

Figure 3: Average accuracy rates of LDA and RF classifiers with all variables and individual variables.



Source: from the authors (2024).

Legend: Average accuracy rate with all variables for the linear discriminant analysis and random forest classifiers (yellow background), where the best-performing classifier is highlighted in red. Average accuracy rate for each variable evaluated individually (gray background), where the classifier with the highest average accuracy rate for each variable is highlighted in black.

Quando pretende-se escolher o melhor classificador para cada variável isoladamente, no intuito de combinar classificadores, o classificador FA obteve a maior taxa média de acertos no comprimento máximo do peciolo (p_{cm}) e empatou com o classificador ADL no comprimento máximo do limbo (l_{cm}) (Figura 3). Já o classificador ADL obteve a maior taxa média de acertos para o comprimento máximo da folha (f_{cm}), largura do pecíolo (p_{la}) e largura máxima do limbo (l_{lm}).

Este trabalho identificou uma ordem de importâncias nas variáveis, de modo que a variável comprimento do pecíolo foi a que mais acertou corretamente a espécie quando avaliada isoladamente, seguida pela largura do pecíolo, comprimento máximo da folha, largura máxima do limbo e comprimento máximo do limbo. Este resultado pode auxiliar novos estudos que visem criar sistemas de classificação de espécies baseados em variáveis como construído por Du *et al.* (2006). Ainda, tal resultado ressalta a importância do uso de características do pecíolo como variáveis para auxiliar no processo de identificação, pois em estudos com esse fim, é comum a remoção do pecíolo das folhas para a utilização apenas de características do limbo foliar (NASIR *et al.*, 2014; KUMAR *et al.*, 2012; OSO; JAYEOLA, 2021). Desta forma o pecíolo pode ser um importante descritor a ser considerado em sistemas que auxiliem o processo de identificação de espécies por meio de variáveis contínuas.

Considerações finais

O classificador análise discriminante linear superou o desempenho do classificador florestas aleatórias no conjunto de dados observado, resultado que indica o potencial que este classificador tem para contribuir com problemas de identificação botânica.

Entre as variáveis que se destacaram positivamente para classificar este conjunto de dados estão o comprimento e a largura da pecíolo, indicando que características do pecíolo podem ser importantes para serem consideradas em estudos futuros.

Estudos futuros poderiam aumentar o número de observações por espécies e verificar a variabilidade das variáveis de tamanho, para verificar se essas características mantém-se informativas na medida que observam-se mais folhas da mesma espécie, especialmente de indivíduos diferentes e de diferentes locais.

Agradecimentos

Nós somos gratos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro e pela bolsa de Doutorado. Somos gratos à UFLA e ao Instituto de Ciências Exatas e Tecnológicas da UFLA pela auxílio. Somos gratos às contribuições de Amanda Merian Freitas Mendes, Carlos Pereira da Silva, Alex Monito Nhancololo, Gean Pereira Damaceno, Valdeline de Paula Mequelino Ferreira, Ali William Canaza Cayo, Diana del Rocío Rebaza Fernández, Igor de Carvalho Aguiar Rodrigues e Julianne Maria Galindo Bezerra. Também somos gratos às contribuições dos revisores.

Referências Bibliográficas

BARRÉ, P. *et al.* Leafnet: A computer vision system for automatic plant species identification. *Ecological Informatics*, v. 40, p. 50–56, 2017.

COPE, J. S. *et al.* Plant species identification using digital morphometrics: A review. *Expert Systems with Applications*, v. 39, n. 8, p. 7562–7573, 2012.

DU, J.-X. *et al.* Computer-aided plant species identification (capsi) based on leaf shape matching technique. *Transactions of the Institute of Measurement and Control*, v. 28, n. 3, p. 275–285, 2006.

ELEN, A.; AVUÇLU, E. Automatic detection of petiole border in plant leaves. *Measurement and Control*, v. 54, n. 3-4, p. 446–456, 2021.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

JOHNSON, R.; WICHERN, D. *Applied multivariate statistical analysis*. Pearson, 2014.

KUMAR, N. *et al.* Leafsnap: A computer vision system for automatic plant species identification. In: SPRINGER. Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, *Proceedings*, Part II 12. [S.l.], 2012. p. 502–516.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: ¡https://CRAN.R-project.org/doc/Rnews/¿.

LUNA-BONILLA, O. Á. D. *et al.* Leaf morphometric analysis and potential distribution modelling contribute to taxonomic differentiation in the quercus microphylla complex. *Journal of Plant Research*, v. 137, n. 1, p. 3–19, 2024.

LYSKO, A. *et al.* Comparison of discriminant methods and deep learning analysis in plant taxonomy: a case study of elatine. *Nature*, v. 12, n. 1, p. 20450, 2022.

MARKS, S.; DUNN, O. J. Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association*, v. 69, n. 346, p. 555–559, 1974.

MATTHEWS, G. J. *et al.* A comparison of machine learning techniques for taxonomic classification of teeth from the family bovidae. *Journal of Applied Statistics*, v. 45, n. 15, p. 2773–2787, 2018.

MURPHY, K. P. *Probabilistic Machine Learning:* An Introduction. [S.l.]: MIT press, 2022.

NASIR, A. F. A. *et al.* Automatic identification of ficus deltoidea jack (moraceae) varieties based on leaf. *Modern Applied Science*, v. 8, n. 5, p. 121, 2014.

OSO, O. A.; JAYEOLA, A. A. Digital morphometrics: Application of morpholeaf in shape visualization and species delimitation, using cucurbitaceae leaves as a model. *Applications in Plant Sciences*, v. 9, n. 9-10, p. e11448, 2021.

R Core Team. *R*: A Language and Environment for Statistical Computing. Vienna, Austria, 2023. Disponível em: ¡https://www.R-project.org/¿.

THANIKKAL, J. G.; DUBEY, A. K.; THOMAS, M. T. A novel edge detection method for medicinal plant's leaf features extraction. *International Journal of System Assurance Engineering and Management*, v. 14, n. 1, p. 448–458, 2023.

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S.* Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: ¡https://www.stats.ox.ac.uk/pub/MASS4/¿.

YIGIT, E. *et al.* A study on visual features of leaves in plant identification using artificial intelligence techniques. *Computers and Electronics in Agriculture*, v. 156, p. 369–377, 2019.

Apêndice

Table 2: Average accuracy rate of LDA and RF classifiers for the 31 combinations of 5 predictor variables in the identification of Myrtaceae species.

Modelo	p_{la}	f_{cm}	p_{cm}	l_{cm}	l_{lm}	acertos ADL (%)	acertos FA (%)
1	1	1	1	1	1	98,2	96,8
2	0	1	1	1	1	94,7	94,8
3	1	0	1	1	1	98,2	96,8
4	0	0	1	1	1	94,7	95,0
5	1	1	0	1	1	96,9	92,5
6	0	1	0	1	1	92,9	87,5
7	1	0	0	1	1	93,8	92,5
8	0	0	0	1	1	81,6	85,1
9	1	1	1	0	1	98,1	97,3
10	0	1	1	0	1	94,5	95,0
11	1	0	1	0	1	97,5	96,2
12	0	0	1	0	1	93,6	93,2
13	1	1	0	0	1	95,0	93,6
14	0	1	0	0	1	84,9	86,9
15	1	0	0	0	1	78,5	79,3
16	0	0	0	0	1	60,2	56,9
17	1	1	1	1	0	94,6	94,1
18	0	1	1	1	0	88,6	88,5
19	1	0	1	1	0	95,1	94,1
20	0	0	1	1	0	89,2	89,2
21	1	1	0	1	0	91,6	83,9
22	0	1	0	1	0	85,9	72,8
23	1	0	0	1	0	83,5	81,3
24	0	0	0	1	0	57,2	57,2
25	1	1	1	0	0	95,2	94,0
26	0	1	1	0	0	89,0	89,1
27	1	0	1	0	0	87,4	87,3
28	0	0	1	0	0	80,5	82,2
29	1	1	0	0	0	84,6	83,5
30	0	1	0	0	0	61,4	59,7
31	1	0	0	0	0	65,7	61,9

Source: from the authors (2024).

Legend: Average accuracy rate of the linear discriminant analysis (LDA) classifier and the random forest (RF) classifier for all 31 possible combinations of the 5 predictor variables in the identification of 5 Myrtaceae species, where: the number 0 indicates when the variable was not included, and 1 indicates when it was included in the model; p_{la} is the petiole width before the beginning of the leaf blade; f_{cm} is the maximum leaf length; p_{cm} is the maximum petiole length; l_{cm} is the maximum leaf blade length; and l_{lm} is the maximum leaf blade width.