



RESEARCH ARTICLE - ANTS

A Machine Learning Approach to Identify Ant Species from the Genus *Ectatomma* Smith, 1858 (Hymenoptera: Formicidae)

AMANDA A. JESUS SANTOS¹, JULIO O. SILVA², JACQUES HUBERT C. DELABIE³, GABRIELA S. C. COSTA⁴, ELTAMARA S. CONCEIÇÃO¹

1 - Universidade do Estado da Bahia, PPGMSB, Alagoinhas, Bahia, Brazil

2 - Instituto Federal da Bahia, ADS, Eunápolis, Bahia, Brazil

3 - Universidade Estadual de Santa Cruz, DCAA & CEPEC-CEPLAC, Ilhéus, Bahia, Brazil

4 - Universidade Estadual de Feira de Santana, DCHF, Feira de Santana, Bahia, Brazil

Article History

Edited by

Evandro Nascimento Silva, UEFS, Brazil

Received 28 March 2025

Initial acceptance 03 August 2025

Final acceptance 29 September 2025

Publication date 23 October 2025

Keywords

Ants; taxonomy; Ectatomminae; artificial intelligence.

Corresponding author

Amanda Araujo de Jesus Santos 
Universidade do Estado da Bahia, PPGMSB
Rodovia BR-110, KM 03, Zona Rural
CEP: 48000-000 - Alagoinhas/BA, Brasil.
E-Mail: amandappgmsb@gmail.com

Abstract

There are some gaps in the taxonomy of certain ant species, making their identification and study more challenging. Regarding some genera, it is an extensive, detailed, and laborious process. This study aims to automate the process of identifying ant species from the genus *Ectatomma*, utilizing machine learning as a tool, and verifying whether it is possible to make the process more accessible and agile. To this end, the Logistic Classifier, Stochastic Gradient Descent, Random Forest Classifier, k-nearest neighbours, Decision Tree Classifier, Support Vector Classification, and Gaussian Naive Bayes algorithms were applied. The algorithms were implemented in the standard version; therefore, no calibration or changes to internal parameters were made. The data set was divided into 70% for training and 30% for testing. The adaptation of the models to the data set was excellent. All methods applied had positive adaptation in the identification of Formicidae. Only two models did not achieve 100% accuracy, but still maintained accuracy above 80%, which is considered highly positive for ant classification. Four of the six algorithms achieved 100% accuracy, validating the efficacy of these methods for species identification. In terms of Myrmecology, the use of supervised algorithms represents a valuable tool in Taxonomy, especially for species of the genus *Ectatomma*.

Introduction

Taxonomy seeks to identify organisms, providing information on the names and distribution of species (Bicudo, 2004). Although it is a very traditional and consolidated discipline, a reflection on the ant scenario appears necessary.

There are more than 14.328 valid ant species names throughout the World (<https://antcat.org>), consulted on 24/03/2025). Therefore, for many ant genera, it is not easy to find a specialist able to help identify some taxa (Martins-Da-Silva, 2014).

Sometimes, taxonomic analysis based on morphological characteristics is insufficient to understand a group. When there is no recent revision of a species group, it will be challenging to conduct scientific studies on it (Senna & Magrin, 1999).

A good example of this is the subfamily Ectatomminae, which has undergone several changes over the years in terms of species and genus composition, based on phylogenetic relationships, resulting in significant changes to its classification (Emery, 1798; Bolton, 2003; Ouellette et al., 2006; Moreau et al., 2006).



The genus *Ectatomma* (Smith, 1858) comprises 15 registered species, of which 10 occur in Brazil and are found in various environments, ranging from forests to savannas (Baccaro et al., 2015). Their diversity is underestimated due to insufficient sampling, and some species are cryptic (Delabie et al., 2015). They are generalists, polyphagous, and feed on annelids, a range of arthropods, and even on other ants (Fernández, 1991); they also rarely feed on flower nectaries (Oliveira & Pie, 1998). Most of them nest in the ground, but they can also form colonies in damp logs (Del Valle et al., 2009).

They have different reproductive strategies, such as monogyny and polygyny, as well as microgynes (small-sized queens associated with standard-sized queens) (Baccaro et al., 2015).

It is already known that species groups can irradiate in response to genetic and environmental differences (Nijhout & Davidowitz, 2003), producing similar phenotypes, undergoing mutations, and deviating from their morphological pattern under normal conditions (Pélabon et al., 2010).

Regarding the difficulties encountered by the taxonomy of these ants, and to speed up the identification process, Artificial Intelligence and the field of Machine Learning (ML) emerge, which allows studying and providing models so that machines can suggest specific behaviors that can be used to solve taxonomic problems (Pacola, 2021; Welchen, 2019).

Machine Learning (ML) is a system that trains a predictive model to identify patterns in input data (Rocha et al., 2023). This model is used to make predictions based on unknown data, allowing algorithms to learn and improve their performance without the need for explicit programming (Chandrashekar et al., 2023; Ale Ebrahim Dehkordi et al., 2023; Estrela et al., 2023).

Regarding ant taxonomy, a few studies align with this line of research, utilizing machine learning as a tool for identification. We can mention Wang and Liang (2012), Marques (2018), Santos et al. (2023a,b), who developed lines of research applying artificial intelligence to identify ant species.

One of the subareas of ML is supervised learning, through which a database is used containing characteristics associated with labels, which represent the expected response based on the inputs provided (Marques, 2018). Thus, the system learns by comparing expected responses and predicted values, using an error measure to evaluate the differences between the provided output and the expected value.

To facilitate and accelerate the identification process, this study aimed to automate the process of identifying ant species from the genus *Ectatomma* using Machine Learning as a tool, to verify whether it is possible to make the identification process more efficient and reduce gaps in their taxonomy.

Material and Methods

The Formicidae specimens were obtained from the collection of the Zoology Laboratory at the Universidade do

Estado da Bahia (UNEB), *Campus* II, Alagoinhas, Bahia, from studies conducted in the Northern and Agreste Coast Identity Territory of Bahia (Território de Identidade Litoral Norte e Agreste da Bahia), from 2011 to 2023. Three ant species were used: *Ectatomma opaciventre* (Roger, 1861), *Ectatomma edentatum* (Roger, 1863), and *Ectatomma tuberculatum* (Olivier, 1792), 30 individuals of each, due to their abundance in the region.

The ant metrics followed the standard indicated by Silva-Freitas (2015), which suggests the parameters that should be analyzed. Three analyses were performed to differentiate a more general identification, which can be used for any ant, and two analyses using more specific parameters of the genus.

The parameters chosen for the genus in question were the following generic traits: a- head length, b- head width, c- antenna, d- interocular distance, e- eye length, f- eye width, g- dorsal mesosome, h- lateral mesosome, i- dorsal gaster, j- lateral gaster, k- dorsal petiole, l- lateral petiole, m- femur, and n- tibia, for the first data analysis. For the second analysis, the same 14 parameters were used, including four more parameters specific to the genus, namely: Weber distance (the longest rigid axis of an ant's body), presence of spines, distance between spines, and the height of the three-quarters ($\frac{3}{4}$) portion of the petiole, which consists of reading and measuring the upper part of an ant's petiole. In the third analysis, only these four group-specific parameters mentioned above were analyzed to compare the intermediate results and observe how the algorithms behave with different numbers of parameters.

Ant morphometrics were determined using a stereomicroscope with a camera attached to an HD LITE 1080P equipment, using Capture 2.3 software for photo capture and measurements. The programming language used for other analyses was Python, using the Jupyter 6.4.12 notebook.

Taxonomic keys for the genera Ectatomminae were based on Camacho et al. (2022) and for the species *Ectatomma*, according to Kugler and Brown (1982) and Bolton (1995).

The dataset was constructed from manually obtained ant measurements using a 0.5 mm scale on the stereomicroscope. Therefore, much of the machine learning preprocessing phase could be eliminated, as there were no missing data records. A normalization process was performed to enhance model accuracy and facilitate the visualization of trait correlations. Different models were then selected, tested, and a comparative analysis was performed based on the success or failure of each model on the dataset. The models included: Logistic Classifier, Stochastic Gradient Descent, Random Forest Classifier, k-Nearest Neighbors (KNN), Decision Tree Classifier, Support Vector Classification, and Gaussian Naive Bayes.

The algorithms were simulated in their standard version, meaning that no calibration or alterations were made to their internal parameters. The dataset was divided into 70% for training and 30% for testing. The data were then normalized to avoid potential biases or distortions in the results.

The models' performance in identifying ants will be evaluated using metrics such as accuracy, precision, recall, F1-score, ROC curve, scatter plots, and correlation matrices.

Results

In the first analysis, with 14 parameters, all models adapted very well to the data set. Overall, only two models did not achieve 100% accuracy. Even so, their accuracy was above 80%, which, for the scenario in question – ant classification – is a desirable value (Table 1).

Data analysis revealed that all models performed well on ant classification, with most achieving 100% accuracy. The results obtained for the decision tree and random forest models were very similar to those reported by Gnoatto (2023),

who achieved 96% accuracy in predicting customer churn in service organizations.

In this scenario, 5 out of the seven algorithms used achieved 100% accuracy. This indicates that, based on this criterion, these algorithms classified the ants with complete accuracy, without any errors.

Prediction with custom hyperparameters demonstrated satisfactory results for the Scikit-learn library methods, resulting in improvements in 68 of the 120 performances analyzed. The Random Forest model stood out as the best performer, which can be attributed to its intrinsic nature of reordering the data to optimize the final result.

Figure 1 illustrates the performance of the algorithms using the ant dataset, with model accuracy as the metric. However, for a more comprehensive assessment of the

Table 1. Training, testing, and prediction reference values for supervised algorithms used to classify ants of the genus *Ectatomma*.

Model	Training	Test	Prediction
<i>Logistic Classifier</i>	100%	100%	100%
<i>Stochastic Gradient Descent</i>	100%	93%	93%
<i>Random Forest Classifier</i>	100%	100%	100%
<i>K-Nearest Neighbors</i>	100%	93%	93%
<i>Decision Tree Classifier</i>	100%	100%	100%
<i>Support Vector Classification</i>	100%	93%	93%
<i>Gaussian Naive Bayes</i>	96.72%	100%	100%

Source: Author's collection.

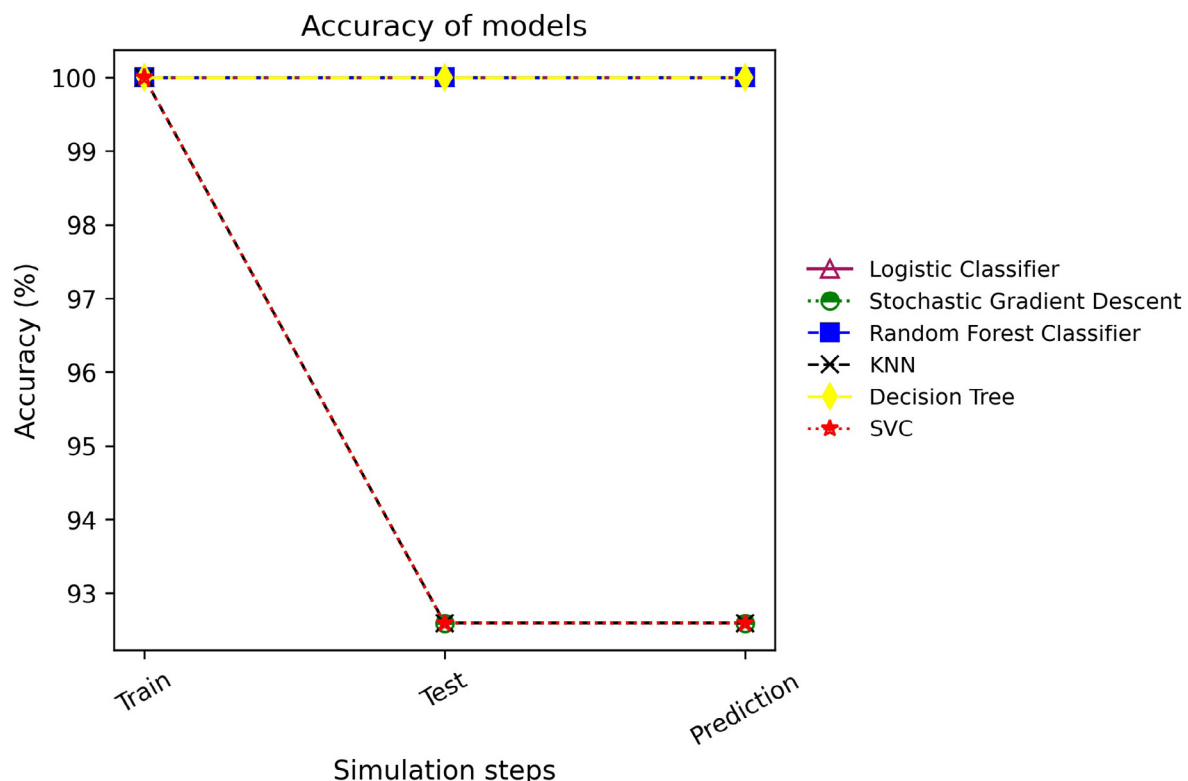


Fig 1. Simulation of the accuracy of supervised models used to classify ants from the genus *Ectatomma*.

Source: Author's collection.

efficiency and assertiveness of the classification algorithms, in addition to accuracy, Table 2 was created, which presents the main classification metrics.

Once again, the results were similar to those previously reported, with the algorithms demonstrating good performance in precision, recall, f1-score, and roc score. This reaffirms their effectiveness in classifying ants. In this case, the models were rerun to generate results applicable to all metrics simultaneously. It was found that the results differed slightly from those previously presented (Gaussian and KNN).

As supported by Chaugan (2022), regarding accuracy in this case, five out of the seven algorithms used achieved 100% accuracy, which implies that, based on this score, these algorithms did not make any errors during classification.

Despite the results presented, a study was conducted to identify the most correlated features/parameters, that is, the best features for the Machine Learning process, especially for building data visualization.

Two approaches were used to determine the optimal number of ant attributes to use as research parameters:

a) Recursive Feature Elimination (parameters) with Cross-Validation (RFECV) with linear regression;

b) Recursive Feature Elimination (RFE) with linear SVC. Both approaches suggested that the optimal number of features is seven out of the 14 attributes in the dataset. The linear SVC simulation is illustrated in Figure 2 below.

Since one of the objectives of this work was to facilitate manual taxonomy, it was determined that measuring

Table 2. Evaluation metrics for supervised algorithms, using computational metrics to measure the performance of supervised algorithms.

Model	Accuracy_score	Precision_score	Recall_score	F1_score	Roc_auc_score
<i>Logistic Classifier</i>	81.48%	82.48%	71.42%	66.66%	78.21%
<i>Stochastic Gradient Descent</i>	85.18%	85.18%	71.42%	71.42%	80.71%
<i>Random Forest Classifier</i>	100%	100%	100%	100%	100%
<i>K-Nearest Neighbors</i>	100%	100%	100%	100%	100%
<i>Decision Tree Classifier</i>	100%	100%	100%	100%	100%
<i>Support Vector Classification</i>	100%	100%	100%	100%	100%
<i>Gaussian Naive Bayes</i>	100%	100%	100%	100%	100%

Source: Author's collection.

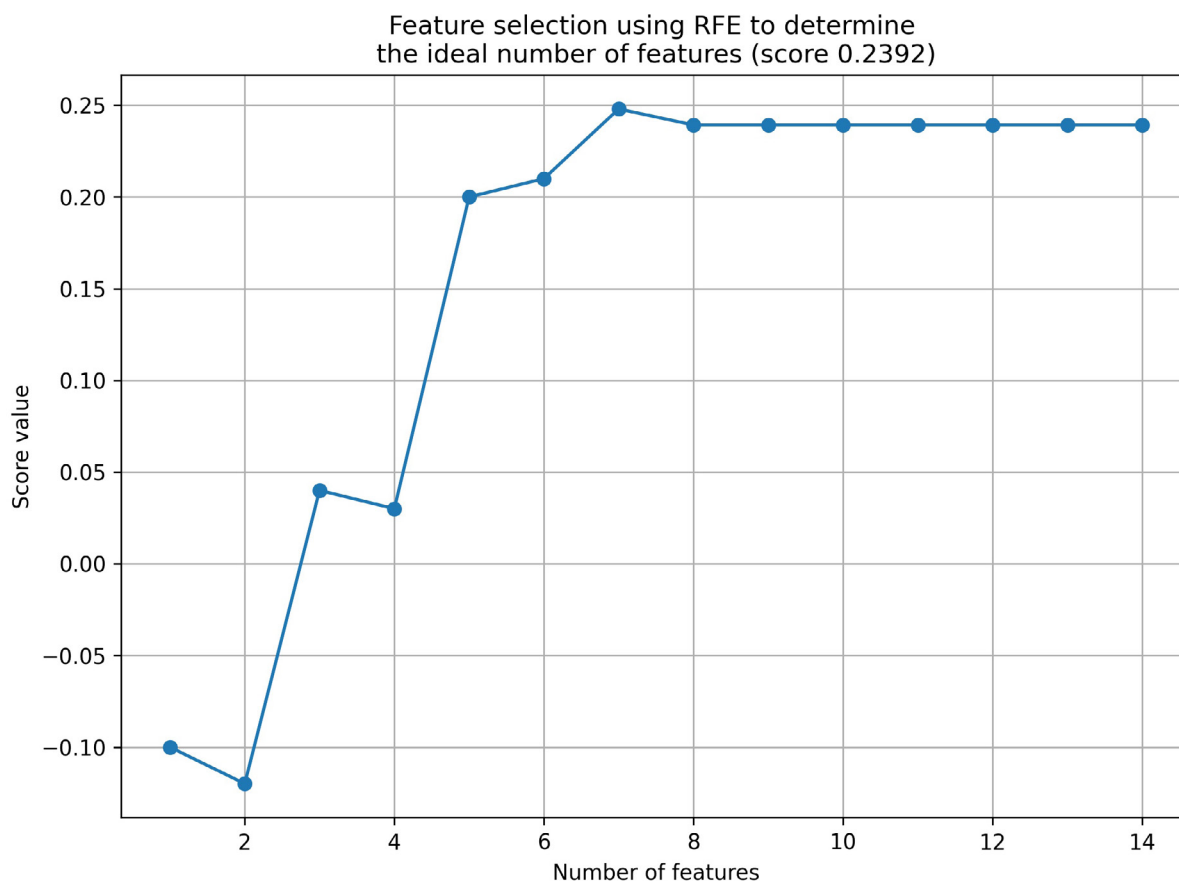


Fig 2. Feature Selection Verification, using RFE number ideal for seven features.

Source: Author's collection.

14 parameters per species was unnecessary. Ideally, seven traits should be used, as after eight traits, there is no change in algorithm performance due to trait collinearity.

Based on this information, a correlation matrix was created, and subsequently, the correlations were pivoted using the ‘unstack’ function, grouping the traits pairwise according to their respective correlations. Thereafter, the seven traits with the highest pivot values were then sorted and selected, pair by pair (Figure 3).

The detected traits were: interocular distance, head width and length, mesodorsal and mesolateral, dorsal and lateral gaster, and antenna. Note that seven attributes are still a relatively large number to include in a single visualization, making it difficult to understand the data. Therefore, to improve the analysis of these cause-and-effect relationships, Figure 4 was created with only the first two most closely related attributes.

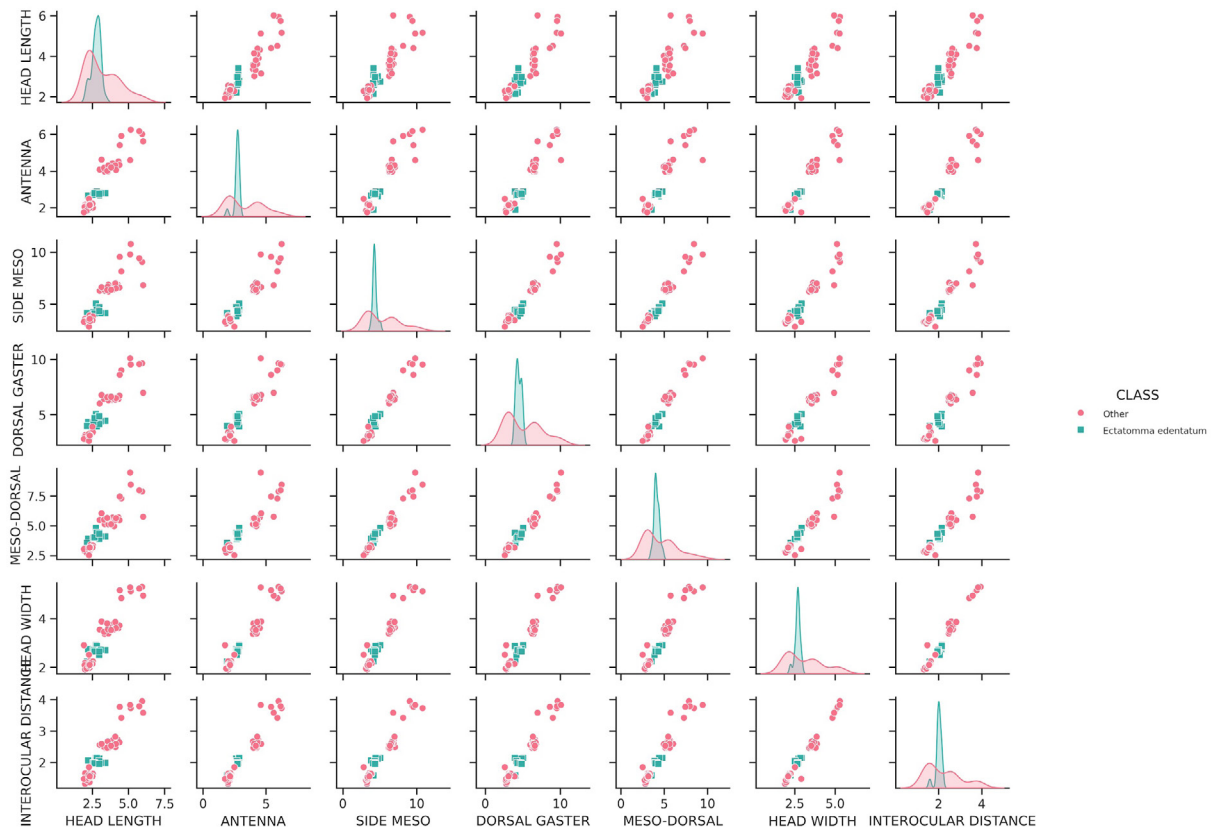


Fig 3. Correlation matrix of the seven ant traits with the highest reference value for identification using supervised algorithms

Source: Author's collection.

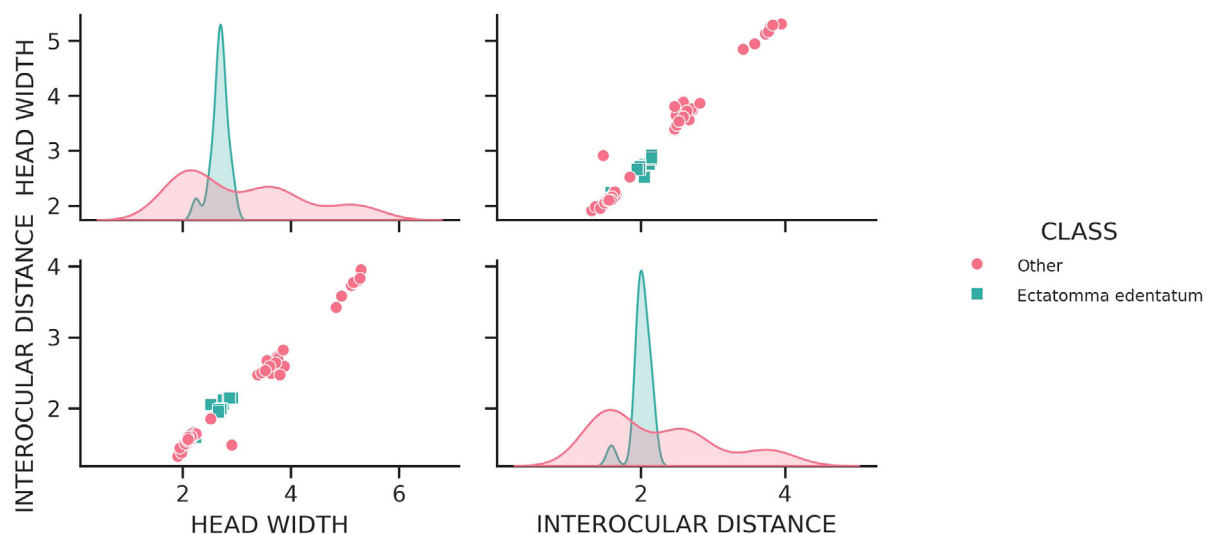


Fig 4. Correlation matrix using two traits that were most correlated in the matrix.

Source: Author's collection.

As illustrated in the figure, interocular distance and head width were the traits that showed the highest correlation. This result indicates a good clustering in the distribution of these measurements, justifying the inclusion of these variables in the group of best features.

Although the seven attributes (Table 3) are considered the best features, including only these in the database resulted in a decrease in model accuracy. This indicates that the model's quality depends on the use of the other features. In the context of taxonomy, this means that the three characteristics identified by the model as most important are not sufficient to delimit the species.

Comparing the data in Tables 1 and 3 reveals a decrease in the performance of some algorithms. However, three algorithms maintain 100% accuracy, and the others exceed 96%, values still considered satisfactory, with only

Table 3. Evaluation metrics for supervised algorithms, using only 7 of the most correlated traits.

Model	Training	Test	Prediction
<i>Logistic Classifier</i>	78.68%	66.66%	66.66%
<i>Stochastic Gradient Descent</i>	73.77%	62.96%	62.96%
<i>Random Forest Classifier</i>	100%	100%	100%
<i>K-Nearest Neighbors</i>	96.36%	96.29%	96.29%
<i>Decision Tree Classifier</i>	100%	100%	100%
<i>Support Vector Classification</i>	98.36%	96.29%	96.29%
<i>Gaussian Naive Bayes</i>	98.36%	96.29%	96.29%

Source: Author's collection.

The identification of these species achieves optimal performance when analyzing both generic and specific features, with 18 features appearing to be the ideal number for this type of identification. This indicates that automated identification, utilizing machine learning algorithms, is more efficient when analyzing a more complex dataset with additional information about the group under examination. In the context of machine learning, a system's inference capability on a dataset is directly proportional to the amount of information provided for its training. Figure 5 shows the accuracy of the models. Based on these results, other evaluation metrics are verified,

two being deemed inadequate (with performance below 70%). This change is attributed to the reduced data volume in the dataset used.

In machine learning, the quality of results improves with the amount of data provided. However, the results obtained indicate that seven features are sufficient to achieve satisfactory performance, with 96% accuracy, which is statistically close to 100%. Considering the study's objectives, a crucial aspect is the reduction of the intensive manual labor inherent in taxonomy.

Table 4 presents the results obtained by introducing four additional parameters into the analyses, totaling 18 measured features. These new parameters are specific to the species group of the genus *Ectatomma* and were added to observe the algorithms' behavior when faced with more specific data from the group.

Table 4. Accuracy of models using 18 traits.

Model	Training	Test	Prediction
<i>Logistic Classifier</i>	100%	100%	100%
<i>Stochastic Gradient Descent</i>	100%	100%	100%
<i>Random Forest Classifier</i>	100%	100%	100%
<i>K-Nearest Neighbors</i>	100%	100%	100%
<i>Decision Tree Classifier</i>	100%	93%	93%
<i>Support Vector Classification</i>	100%	100%	100%
<i>Gaussian Naive Bayes</i>	100%	100%	100%

Source: Author's collection.

allowing inferences about their results to remain consistent across other indices (Table 5).

This evaluation shows an improvement in the model's accuracy, precision, recall, and f1-score. When compared to the initial analysis, it is important to note that the behavior of each algorithm varies depending on the number of parameters. Specifically, SVC achieved 100% performance with 14 features, but showed a slight drop with 18, the reason for this behavior remaining unclear. In machine learning projects, it is essential to consider a trial-and-error approach. Adding or removing features and modifying the dataset can significantly influence the algorithm's behavior and performance.

Table 5. Model checking supervised, using other computational metrics to measure the performance of algorithms using 18 traces.

Model	Accuracy_Score	Precision_score	Recall_score	F1_score
<i>Logistic Classifier</i>	100%	100%	100%	100%
<i>Stochastic Gradient Descent</i>	100%	100%	100%	100%
<i>Random Forest Classifier</i>	100%	100%	100%	100%
<i>K-Nearest Neighbors</i>	100%	100%	100%	100%
<i>Decision Tree Classifier</i>	100%	100%	100%	100%
<i>Support Vector Classification</i>	96.29%	96.75%	96.29%	96.29%
<i>Gaussian Naive Bayes</i>	100%	100%	100%	100%

Source: Author's collection.

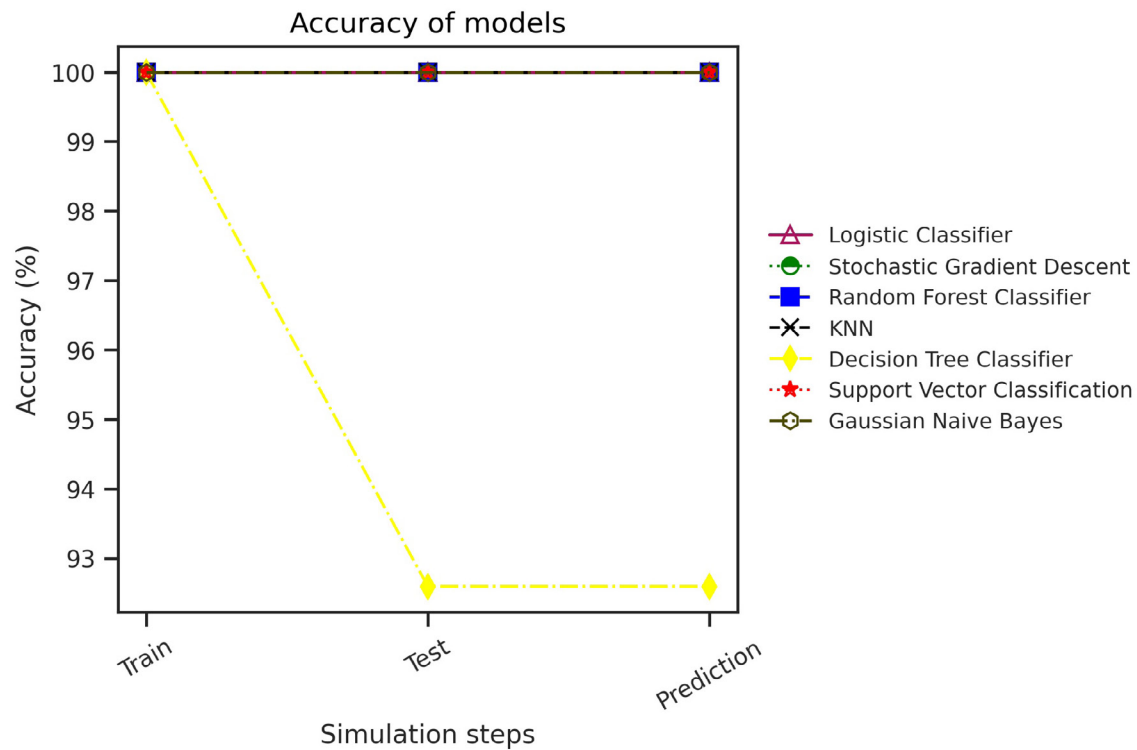


Fig 5. Accuracy of supervised models, used to classify ants of the genus *Ectatomma*, using 18 features.

Source: Author's collection.

Figure 6 reveals consistently strong and positive correlations among the 18 features analyzed. All the matrices constructed demonstrate a pattern where an increase in one variable is accompanied by an increase in all others, with few exceptions to this general trend.

Figure 7 clearly demonstrates the mathematical values that establish the correlation between the parameters. This study was limited to correlation analysis, without advancing to regression, as the objective was not to determine the significance of the factors; consequently, there was no in-depth analysis of null models. Future studies could explore regression and employ additional evaluation metrics to enhance the understanding of this phenomenon further.

Considering that a correlation is significant at 0.3 and that values close to 0 indicate little correlation, while values close to 1 indicate a more perfect relationship, it is surprising to note that Figure 7 shows that the distance parameters between the spines and the petiole³/₄ have values below 0. This demonstrates the absence of any correlation between these traits, making it impossible to prove their relationship. Taxonomically, this lack of correlation may be indicative of the traits that define the specificities of each species within the group.

Generally, values above 0.3 indicate that most parameters are highly correlated with each other. Antenna/head width; interocular distance/head length; head width/lateral and dorsal mesosome; interocular distance/dorsal gaster; eye/with femur and tibia; meso gaster/dorsal lateral; dorsal mesosome with dorsal gaster; dorsal petiole with lateral petiole; and femur with tibia were those that showed

the highest correlation with each other. The most correlated are precisely the most generic, as indicated in the feature selection graph, which shows them as the most correlated when analyzing the 14 generic parameters: antenna, head width and length, interocular distance, dorsal and lateral mesosome, dorsal and lateral gaster, eyes, femur, and tibia. A correlation close to 1 between the values indicates that they increase simultaneously, which is in agreement with the results presented in Figures 7 and 5 of the previous matrices.

Finally, to evaluate the algorithms' behavior, we analyzed their performance using only four specific parameters: the distance between spines on the ant's pronotum, the height of the petiole at three-quarters (³/₄), the Weber distance (measurement of the longest rigid axis of the ant's body), and the presence of isolated spines, separated from the general parameters. The results obtained are presented in Table 6.

Table 6. Training, testing, and prediction reference values for the supervised algorithms using four specific parameters for the *Ectatomma* genus.

Model	Training	Test	Prediction
<i>Logistic Classifier</i>	100%	100%	100%
<i>Stochastic Gradient Descent</i>	100%	100%	100%
<i>Random Forest Classifier</i>	100%	100%	100%
<i>K-Nearest Neighbors</i>	100%	100%	100%
<i>Decision Tree Classifier</i>	100%	100%	100%
<i>Support Vector Classification</i>	100%	100%	100%
<i>Gaussian Naïve Bayes</i>	100%	100%	100%

Source: Author's collection.



Fig 6. Correlation matrix of the 18 ant traits with the highest reference value for identification using supervised algorithms.

Source: Author's collection.

The performance analysis of the algorithms, using only the four genus-specific parameters under study, revealed significant benchmark values.

A general analysis of the results shows that using a smaller, more specific set of parameters for a given group is more efficient than using 14 general parameters applicable to all ants. The combination of the 14 general parameters with the four specific parameters results in performance virtually identical to that obtained using only the four specific parameters.

Finally, the final matrix (Figure 8), which utilizes only the four genus-specific traits, reveals a stronger correlation between the species *E. opaciventre* and *E. tuberculatum* in relation to the analyzed traits. This finding is consistent with the morphology of these species. When consulting their identification keys, it is observed that both share more similar characteristics, such as the presence of spines, which are absent in *E. edentatum*, for example.

Discussion

Based on the accuracy achieved in classifying ants in this study, and considering the suitability of all models for this purpose, it can be concluded that the use of supervised algorithms is an acceptable approach for identifying the studied species, as the models demonstrated 100% accuracy.

Similarly, Wang & Liang (2012) also employed supervised algorithms to develop a new automated system for identifying insect images. Using digital image progression methods and Support Vector Machines (SVMs), the model achieved 93% accuracy in tests with nine common insect orders and suborders, including ants.

The use of models such as Decision Trees, Random Forests, Linear Regression, and Logistic Regression demonstrated satisfactory results in modeling river basins. Among these, the Decision Tree model stood out for its data predictability

Correlation between ant dataset variables

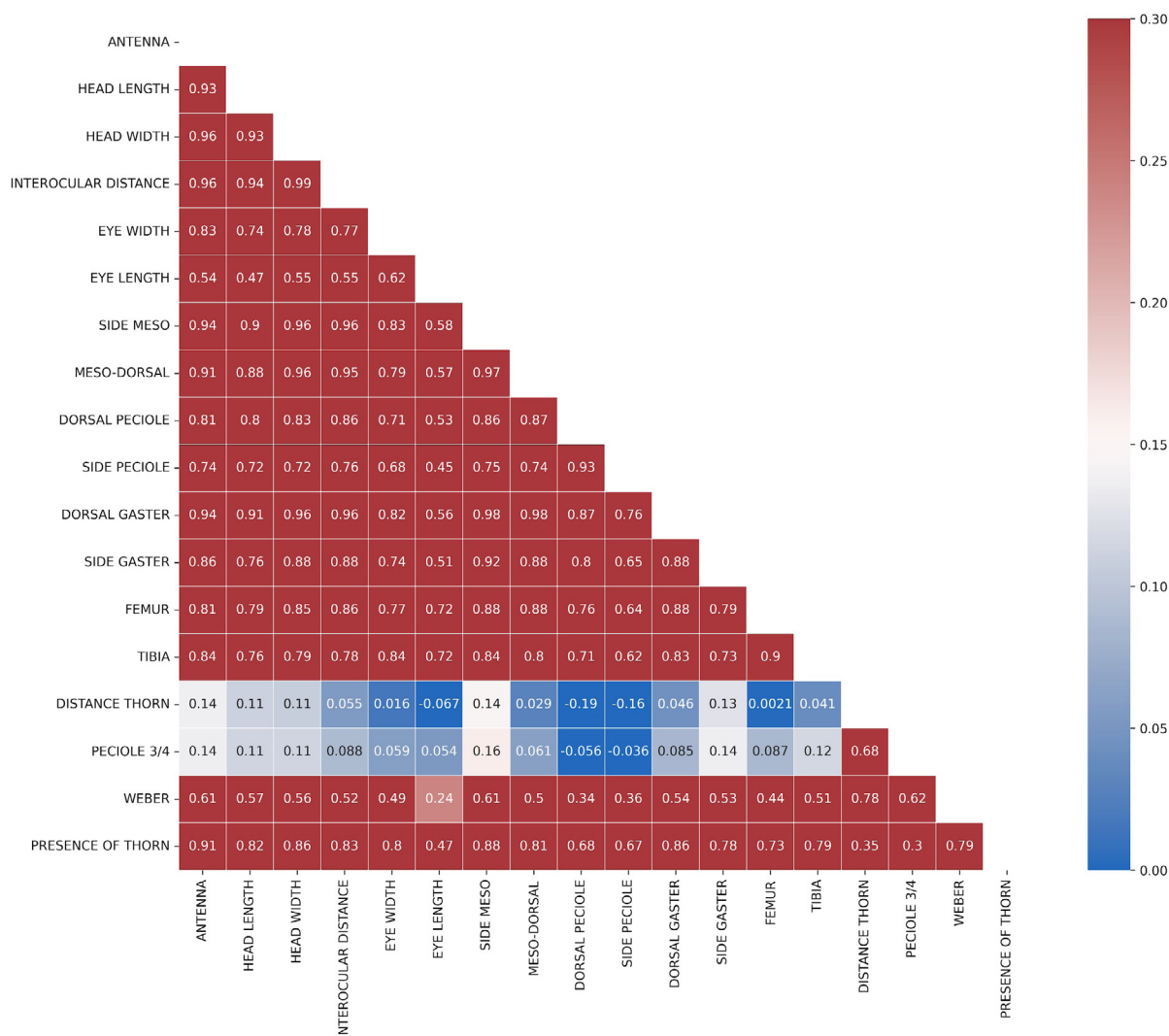
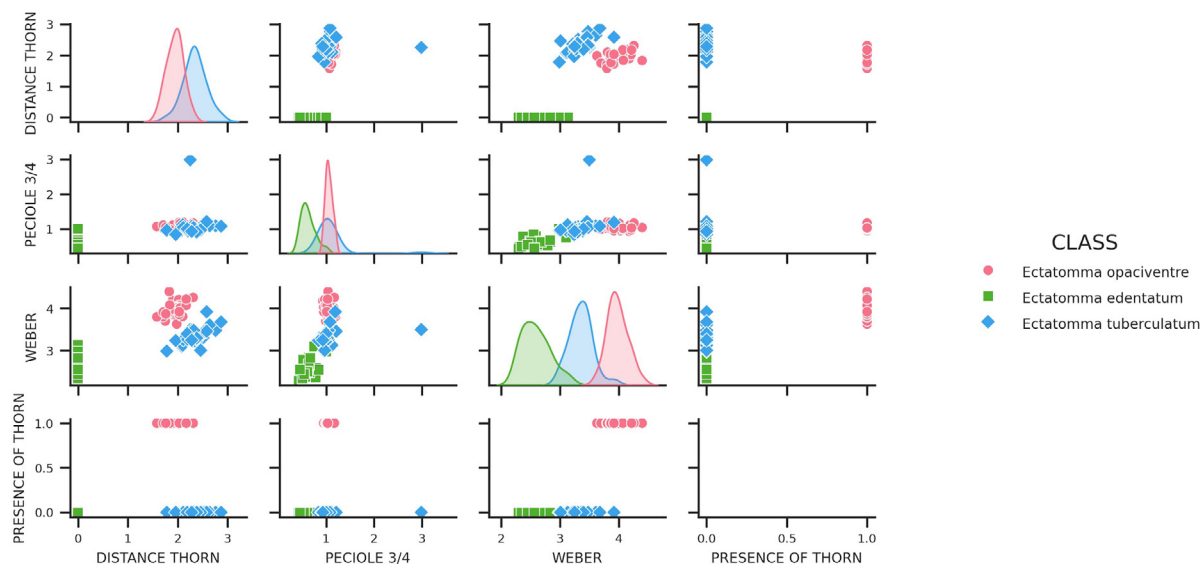


Fig 7. Correlation between variables (traits) of the ant dataset for 18 traits.

Source: Author's collection.

Fig 8. Correlation matrix using only the four specific traits of the *Ectatomma* genus that were most correlated in the matrix.

Source: Author's collection.

(Azevedo-Silva et al., 2023). Random Forest also demonstrated better overall performance in Takáo (2023), specifically in evaluating the individual probability of a confirmed diagnosis of Inborn Errors of Immunity (IEI). Both algorithms were also considered among the best-performing models for *Ectatomma* classification.

In Freitas et al. (2023), the Decision Tree and Support Vector Classification (SVM) algorithms outperformed Naive Bayes and k-nearest neighbors (KNN), with the Decision Tree presenting the best results. Similarly, in Jacinto (2023), when analyzing electrical behaviors, Random Forest demonstrated superior performance, followed by Decision Tree.

The models Random Forest, Xgboost, Support Vector Classification SVM, and a neural network were used to analyze the occurrence of yellow fever in Minas Gerais. They showed that Random Forest performed best, along with Support Vector Classification SVM (de Lourdes, 2023). The confusion matrices for the Decision Tree, KNN (k-nearest neighbors), Naive Bayes, and SVM (Support Vector Machine) algorithms (Freitas, 2023) showed overall performance below 70%, which differs from our results. The author explained that this low performance is due to the small number of samples used to explain the phenomenon, and that a larger dataset will be necessary to improve the algorithms' performance.

When comparing the performance of nine machine learning models to identify the best-performing model in credit analysis (Lopes et al., 2023), it was found that the following models achieved the following accuracies: Logistic Regression, 87.68%; KNN, 88.41%; and SVM, 91.74%. These values are close to those found in this work.

The results obtained in this research were consistent with those of other authors, corroborating the effectiveness of supervised algorithms in automating tasks and identifying patterns. This is especially true for tree-structured algorithms, which, by their nature, perform rearrangements to improve classification and identification of the database.

The results of this study confirm two main conclusions: 1) In the field of Machine Learning, the accuracy of responses is directly proportional to the size of the dataset used for training; 2) When working with ant databases, it is more effective to measure specific characteristics of the group under study, as using generic traits results in a less satisfactory response than exploring the specificities of each species.

It is essential to emphasize that there is no "right" or "wrong" method for developing group identification through supervised algorithms. It is an iterative process of trial and error, aimed at discovering the most suitable models and optimal parameters for the analysis.

Authors' Contribution

AAJS: Conceptualization, validation, investigation, data curation, writing-original draft, visualization.

JOS: Methodology, validation, formal analysis, data curation

JHCD: Writing-review & editing, supervision.

GSCC: Translation, writing-original draft, writing-review & editing.

ESC: Conceptualization, investigation, supervision, project administration.

Declaration of non-financial interests

All authors certify that they have no affiliation or involvement with any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Declaration of consent and data

The research employed several machine learning algorithms by importing ready-made models and applying them. Therefore, it can be concluded that the research uses open-source models and can be reproduced by anyone or institution with access to the data and pipeline used. All authors declare that they agree with the data written and consent to the publication of the article in the journal Sociobiology. Link the data, pipeline, and figures: https://drive.google.com/drive/folders/1ryQkFW9Jr_zYd9n7QOI-8X_BjOW_zCP9?usp=sharing.

References

- Dehkordi, M. A. E., Lechner, J. M., Ghorbani, A., Nikolic, I., Chappin, E. J. L. & Herder, P. M. (2023). Using machine learning for agent specifications in agent-based models and simulations: A critical review and guidelines. *Journal of Artificial Societies and Social Simulation*, 26: 9. <https://doi.org/10.18564/jasss.5016>
- ANTWEB. Version 8.103.2. California Academy of Science, online at <https://www.antweb.org>. Accessed 30 January 2024.
- Baccaro, F. B., Feitosa, R. M., Fernández, F., Fernandes, I. O., Izzo, T. J., Souza, J. D. & Solar, R. (2015). *Guia para os gêneros de formigas do Brasil*. Manaus: Editora INPA, 388p. <https://doi.org/10.5281/zenodo.32912>.
- Bicudo, C. E. D. M. (2004). *Taxonomia*. Biota Neotropica, 4, I-II. <https://doi.org/10.1590/S1676-06032004000100001>.
- Bolton, B. (2003). *Synopsis and classification of Formicidae*. *Memoirs of the American Entomological Institute*, 71: 1-370.
- Bolton, B. (1995). *A New General Catalogue of the Ants of the World*. Harvard University Press, Cambridge, Mass.
- Camacho, G.P., Franco, W., Branstetter, M.G., Pie, M.R., Longino, J.T., Schultz, T.R. & Feitosa, R.M. (2022). UCE phylogenomics resolves major relationships among ectaheteromorph ants (Hymenoptera: Formicidae: Ectatomminae, Heteroponerinae): a new classification for the subfamilies and the description of a new genus. *Insect Systematics and Diversity*, 6: 5. <https://doi.org/10.1093/isd/ixab026>.

- Chandrashekar, D. V., Babu, Y. S., Suneetha, K., & Kumar, T. K. (2023). Machine Learning Meets the Semantic Web. In: Patel, A. & Debnath, N.C. (Eds). Science with Semantic Technologies: Deployment and Exploration. Boca Raton, CRC Press. p. 1-12.
- de Azevedo Silva, V., de Feiras Souza, R., Oliveira, M. P., & dos Santos, F. L. (2023). Aplicação de machine learning e deep learning para modelagem de uma bacia hidrográfica. *Paranoá*, 34: 1-21.
- de Freitas, M., Piai, V., Souza, P., Ramirez, A. & Fernandes, A. (2023). Uso de Aprendizado de Máquina para Identificar o Tipo de Afasia Progressiva Primária a partir do Desempenho no Trog-2Br. *Anais do Computer on the Beach*, 14: 512-514. <https://doi.org/10.14210/cotb.v14.p512-514>.
- de Lourdes Araújo, I. (2023). Comparação da performance de algoritmos de aprendizado de máquina para análise preditiva de febre amarela no estado de Minas Gerais. Tese de doutorado. Universidade Federal de Minas Gerais. Recovered from: <http://hdl.handle.net/1843/52262>.
- Del Valle, E. E., Dolinski, C., Barreto, E. L. & Souza, R. M. (2009). Effect of cadaver coatings on emergence and infectivity of the entomopathogenic nematode *Heterorhabditis baujardi* LPP7 (Rhabditida: Heterorhabditidae) and the removal of cadavers by ants. *Biological Control*, 50: 21-24. <https://doi.org/10.1016/j.biocontrol.2009.01.007>.
- Delabie, J. H., Feitosa, R. M., Serrão, J. E., Mariano, C. D. S. F. & Majer, J. D. (Eds.). (2015). As formigas poneromorfas do Brasil. SciELO-Editus-Editora da UESC.
- Emery, C. (1798). Die Gattung *Dorylus* Fab. und die systematische Eintheilung der Formiciden. *Histoire*, 6: 18.
- Estrela, V. V., de Jesus, M. A., Intorne, A. C., Batista, K. K., Deshpande, A., Shi, F., ... & Oliveira, L. P. (2013). Medical Visual Theragnostic Systems Using Artificial Intelligence (AI) – Principles and Perspectives. In: Estrela, V. V. (Ed.) *Intelligent Healthcare Systems*. Boca Raton, CRC Press. p. 301-321.
- Fernández, F. (1991). Las hormigas cazadoras del género *Ectatomma* (Formicidae: Ponerinae) en Colombia. *Caldasia*, 551-564. Recovered from: <https://www.jstor.org/stable/23641395>.
- Gnoatto, R. & Franzen, E. (2023). Análise do desempenho de hiperparâmetros de aprendizagem de máquina aplicados na previsão da taxa de rotatividade de clientes. *Revista Destaques Acadêmicos*, 15: 3635. <https://doi.org/10.22410/issn.2176-3070.v15i4a2023.3635>
- Jacinto, G. L., Imamura, L. Y., Grellert, M., & Meinhardt, C. (2023). Explorando Predição da Caracterização Elétrica com Machine Learning. *Anais do Computer on the Beach*, 14: 194-201. <https://doi.org/10.14210/cotb.v14.p194-201>.
- Kugler, C. & Brown, W. J. (1982). Revisionary and other studies on the ant genus *Ectatomma*, including the descriptions of two new species. *Search: Agriculture* (Cornell University Agricultural Experiment Station), 24: 1-8.
- Lopes, R.D.S.A., Colombi, L.R. & Mutz, F. (2023). Comparação de Algoritmos de Aprendizado de Máquina para Predição de Pontuação de Crédito. *Anais do Computer on the Beach*, 14: 424-431. Recovered from: <https://repositorio.ifes.edu.br/handle/123456789/1755>
- Marques, A.C.R. (2018). Contribuição à abordagem de problemas de classificação por redes convolucionais profundas (Doctoral dissertation, Tese (Doutorado em Engenharia Elétrica com Ênfase em Automação) – Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas. Campinas-SP). Recovered: <https://core.ac.uk/reader/296892260>.
- Martins-da-Silva, R. C. V., Silva, A. D., Fernandes, M. M. & Margalho, L. F. (2014). Noções morfológicas e taxonômicas para identificação botânica. Embrapa Amazônia Oriental.
- Moreau, C. S., Bell, C. D., Vila, R., Archibald, S. B. & Pierce, N. E. (2006). Phylogeny of the ants: diversification in the age of angiosperms. *Science*, 312: 101-104. <https://doi.org/10.1126/science.1124891>
- Nijhout, H. F. & Davidowitz, G. (2003). Developmental perspectives on phenotypic variation, canalization, and fluctuating asymmetry. *Developmental instability: causes and consequences*, 3-13.
- Oliveira, P. S. & Pie, M. R. (1998). Interaction between ants and plants bearing extrafloral nectaries in cerrado vegetation. *Anais da Sociedade Entomológica do Brasil*, 27: 161-176. <https://doi.org/10.1590/S0301-80591998000200001>.
- Ouellette, G. D., Fisher, B. L. & Girman, D. J. (2006). Molecular systematics of basal subfamilies of ants using 28S rRNA (Hymenoptera: Formicidae). *Molecular Phylogenetics and Evolution*, 40: 359-369. <https://doi.org/10.1016/j.ympev.2006.03.017>.
- Pacola, V. (2021). Inteligência artificial na engenharia de software. Monografia de Especialização. Pontifícia Universidade Católica, São Paulo, SP.
- Pelabon, C., Hansen, T. F., Carter, A. J. & Houle, D. (2010). Evolution of variation and variability under fluctuating, stabilizing, and disruptive selection. *Evolution*, 64: 1912-1925. <https://doi.org/10.1111/j.1558-5646.2010.00979.x>.
- Rocha, M. B., Silveira, B. P. & Pilger, D. (2023). Aprendizado de máquina nos serviços farmacêuticos: uma revisão integrativa. *Clinical and Biomedical Research*, 43: 75-82. <https://doi.org/10.22491/2357-9730.124096>.
- Santos, A. A. R., Silva, J. O., Fontoura, J. R. A., Lima, D. M., Delabie, J. H. C., Conceição, E. S. (2023a). Machine learning's using of classifying algorithm on identifying *Ectatomma* genus ant's species. XXVI Simpósio de Mirmecologia at Manaus, Amazonas, Brazil.

Santos, A. A. R., Silva. J. O., Lima. D. M., Araujo. V. V., Delabie. J. H. C., Conceição. E. S. (2023b). Automated Identification of *Ectatomma edentatum* (Hymenoptera: Formicidae) using Supervised Algorithms. *Journal of Bioengineering, Technologies and Health*, 6: 1-8. <https://doi.org/10.34178/jbth.v6iSuppl2.347>.

Senna, P. A. C. & Magrin, A. G. E. (1999). A importância da “boa” identificação dos organismos fitoplanctônicos para os estudos ecológicos. In: Pompêo, M.L.M. (Ed.) *Perspectivas da Limnologia no Brasil*. Gráfica e Editora União, São Luís, p. 131-146.

Silva-Freitas, J. M. Mariano, C. F. & Delabie, J.H.C. (2015). *Morfometria de Formicidae*. (Doctoral dissertation, Universidade Estadual de Santa Cruz).

Takáo, M. M. V. (2023). *Inteligência artificial em alergologia e imunologia: desenvolvimento de modelos de predição de risco para erros inatos da imunidade* (Doctoral dissertation, [sn]).

Wang, J., Lin, C., Ji, L. & Liang, A. (2012). A new automatic identification system of insect images at the order level. *Knowledge-Based Systems*, 33: 102-110. <https://doi.org/10.1016/j.knosys.2012.03.014>.

Welchen, Vandoir. *Uso de inteligência artificial em apoio à decisão clínica: o caso do Hospital de Câncer Mãe de Deus com a ferramenta cognitiva Watson for oncology*. 2019. *Dissertação de Mestrado*. UCS. Recovered from: <https://repositorio.ucs.br/xmlui/handle/11338/4755>.

