RESEARCH ARTICLE

# Discriminating Planted Capsicum Spp. Varieties via Machine Learning and Multivariate Data Reduction

**Matheus Costa Pereira**[1,*], **Tiago Martins de Azevedo**[1], **Caio Tertuliano Ribeiro**[1], **Ana Izabella Freire**[1], **Matheus Brendon Francisco**[1], **João Luiz Junho Pereira**[1] and **Anderson Paulo de Paiva**[1]

[1] Industrial and Management Engineering Institute, Federal University of Itajubá (UNIFEI), Itajubá, 37500-903, Brazil

## Abstract

**The classification of *Capsicum spp.* varieties is often hindered by their morphological similarities, making accurate identification a challenging task. To address this issue, this study applies a hybrid computational approach that combines data dimensionality reduction techniques using Principal Component Analysis and Factor Analysis with various supervised Machine Learning algorithms. The dataset, which is unprecedented in the literature and was collected under controlled agricultural conditions, enables a robust evaluation of models including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, Decision Tree, and Gradient Boosting. Model performance was assessed using Leave-One-Out and K-Fold cross-validation methods. Additionally, the SHapley Additive exPlanations method was applied to assess the importance of the features in species classification, providing greater interpretability that reinforces the relevance of morphological and agronomic descriptors in differentiating pepper varieties. The results show that all models achieved high performance metrics, including accuracy, F1-score, precision, and recall, consistently above 0.89, validating the effectiveness of the proposed approach. These findings highlight the potential of integrated Machine Learning frameworks for species classification in agriculture, contributing to practical applications and advancing intelligent analysis of biological data.**

**Keywords**: species prediction, pepper, machine learning, classification algorithms, factor analysis.

## 1 Introduction

The Capsicum genus, encompassing a wide array of peppers and bell peppers, is a botanical group within the Solanaceae family. Native to Central and South America, these plants have been extensively studied due to their economic, culinary, and medicinal value [1, 2]. Among its domesticated species, two hold significant cultural and commercial importance: Capsicum frutescens and Capsicum chinense. These species are renowned for their distinctive morphological, chemical, and genetic traits, making them a focus for research in plant biology and food sciences [3].

Within these species, C. frutescens includes varieties

**Submitted:** 12 August 2025
**Accepted:** 19 October 2025
**Published:** 15 November 2025

**Vol.** 1, **No.** 3, 2025.
10.62762/TMI.2025.385133

***Corresponding author:**
✉ Matheus Costa Pereira
matheusc_pereira@hotmail.com

**Citation**
Pereira, M. S., de Azevedo, T. M., Ribeiro, C. T., Freire, A. I., Francisco, M. B., Pereira, J. L. J., & de Paiva, A. P. (2025). Discriminating Planted Capsicum Spp. Varieties via Machine Learning and Multivariate Data Reduction. *ICCK Transactions on Machine Intelligence*, 1(3), 166–185.

such as Cayenne, Tabasco, and Malagueta, widely appreciated for their pungency and adaptability to diverse climates. Similarly, C. chinense comprises cultivars like Bhut Jolokia, Habanero, and Biquinho, known for their intense heat and unique flavor profiles. The rich diversity within these groups presents challenges and opportunities for classification, particularly as genetic, morphological, and biochemical factors intertwine to define the identity of each variety.

The accurate classification of pepper species and varieties is a complex task due to overlapping phenotypic characteristics and genetic variability. A robust classification framework is essential for optimizing agricultural practices, ensuring biodiversity conservation, and facilitating targeted breeding programs. Recent advances in Machine Learning (ML) have shown potential in addressing such challenges by leveraging high-dimensional genetic data. ML algorithms enable researchers to extract meaningful patterns from vast datasets, providing insights into subtle distinctions among species and varieties. These methods allow for efficient and precise classification even in the presence of complex genetic relationships.

There are few studies that apply ML in tabular, genetic, and agronomic data to classify peppers, despite disruptive success in other agricultural applications [4, 5]. One of them was done by Ramírez-Meraz et al. [6] who used ML to classify Capsicum annuum cv. Jalapeno after collecting data on this species. The authors evaluated carbohydrate contents, amino acids, organic acids, among other things. They were able to distinguish 10 new breeds of those species with this technique. The use of genetic data makes differentiation possible even when the appearance of the species is very close (which is a difficulty for the training of models that use images).

Another author was Durmuş and Atasoy [7]. The authors applied multivariate ML methods to investigate organic compound content of different pepper spices. The authors evaluated traits such as terpenoids, acids, glucose, fructose, and used Random Forest to classify the species. They came up with excellent results, showing once again the effectiveness of using vital data from these species.

Hafsah et al. [8] carried out a Classification of cayenne pepper genotypes using physical characteristics during the growing period until harvest using ML. The authors also achieved great results with data such as plant and dichotomous heights, age at flowering, age at harvest, fruit stalk length, chili dimensions, individual

**Table 1.** Studies applied to the classification of peppers.

| Author | Objective | Algorithms used | Results |
|---|---|---|---|
| Meena et al. [9] | Identifying the Geographical Origin of Red Pepper Powder. | KNN, DT, RF and SVM. | SVM had the best result with 97.22% accuracy in the predictions. |
| Abubeker et al. [10] | Classification of pepper "Kantheri mulaku". | YOLO V5. | The method obtained an accuracy of 90%. |
| Houetohossou et al. [11] | Evaluate deep learning performance to classify diseases in peppers when data is unbalanced. | GoogleNet, VGG16, and ResNet50. | GoogleNet showed greater result stability than VGG16 and ResNet50, especially under various levels of imbalance. |
| Djoulde et al. [12] | Classify pepper seeds. | MLP, DT, LDA, NB, SVM and KNN. | SVM had an accuracy of 87%. |
| Jeong et al. [13] | Discrimination of the geographical origin of chili peppers using laser ablation-inductively coupled plasma mass spectrometry, X-ray fluorescence, and near-infrared spectroscopy. | Explainable extreme gradient boosting. | XGBoost had an accuracy of 97.5%. |
| Karadağ et al. [14] | Detection of pepper fusarium disease based on spectral reflectance. | ANN, NB and KNN. | KNN method with 99% of classification performance. |
| Bhagat et al. [15] | Bell pepper leaf disease classification. | RF. | RF achieved 99.75% accuracy. |

chili weight, total number of chilies per plant, and overall yield productivity.

Other studies are highlighted in Table 1. Algorithms such as K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), Multilayer perceptron (MLP), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Artificial Neural Networks (ANN) were used. From the studies presented, it is evident that constructing a pepper classification problem using genetic and agronomic data, rather than images, offers significant advantages. Genetic and health data offer accurate and quantitative information about intrinsic characteristics of peppers, such as disease resistance, nutritional profile, and tolerance to harsh weather conditions, which are difficult to capture through imaging. In addition, this data allows for a more robust and objective analysis, avoiding subjectivity and limitations associated with image processing, such as lighting, angle, and capture quality. This approach also enables more efficient integration with agronomic and genomic studies, promoting informed decisions for genetic improvement and sustainable cultivation, while reducing the need for intensive computational resources often associated with image analysis.

The present study focuses on the classification of two key Capsicum species (C. frutescens and C. chinense) encompassing a variety of cultivars within each group. To the best of our knowledge, these species have not been analyzed from a ML perspective previously and the dataset was acquired after planting them under controlled conditions at a local experimental farm in the southern state of Minas Gerais, Brazil.

By utilizing high-dimensional genetic data, we also aim to develop a predictive framework capable of distinguishing between these species and their respective varieties. To achieve this, we employ a combination of state-of-the-art ML algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), and Gradient Boosting (GB), following a new pre-processing step of dimensionality reduction to enhance computational efficiency and model performance. Once again, our work innovates by applying and comparing different algorithms to this new problem, since, as is well established in the literature, there is no single algorithm that is universally best for all cases [16].

The findings of this study have the potential to contribute significantly to the field of agricultural sciences and biodiversity research. By establishing a robust classification methodology, we aim to provide a reliable tool for researchers and practitioners, facilitating the identification and categorization of pepper varieties. This not only enhances the understanding of genetic relationships within the Capsicum genus but also lays the groundwork for future applications in crop improvement, resource management, and the preservation of genetic diversity.

## 2 Theoretical Background

### 2.1 Importance of *Capsicum spp.* Breeding

When it comes to the genus Capsicum, certain characteristics play a crucial role in the quality of marketed products. For example, the shape, color, and weight of pepper fruits [17] are some of the key traits that meet consumer expectations. It is important to highlight that the market is undergoing significant transformations and gaining relevance, especially due to the diversity of applications [18]. Peppers not only serve as raw material in the production of condiments, spices, and preserves worldwide but are also part of the fresh vegetable market in Brazil. They can be transformed into sauces, jellies, paprikas, and even sold as ornamental plants [19].

Additionally, peppers have significant socioeconomic importance due to their ability to generate employment and income, especially for small producers [18]. Their cultivation is present in almost all regions of Brazil and serves as an excellent example of integration between small farmers and the agro-industry [20].

Genetic improvement plays a crucial role in the sustainable development of modern agriculture. It enables the creation of plants that are more tolerant, productive, and adapted to local climatic and environmental conditions. This practice is fundamental to ensuring food security and efficiency in agricultural production. Over the years, breeders have employed selection techniques to obtain varieties through crossbreeding, resulting in plants that are more productive and adapted to the local environment. This process is essential for enhancing agricultural sustainability and food security [21, 22].

Breeding programs for *Capsicum spp.* have played a significant role by releasing various new varieties. As a result, the possibility of expanding production and occupying different market niches has increased [18]. Breeding aims to meet the essential pillars of modern agriculture and contributes to the development of

more productive plant varieties. This is achieved through the selection of desirable traits, such as high yield, greater resistance to biotic and abiotic stresses, and improved food quality [23].

### 2.2 Dimensionality Reduction

#### 2.2.1 *Principal Component Analysis*

Problems involving complex, high-dimensional datasets are commonly addressed using Principal Component Analysis (PCA). Gaudêncio et al. [24] and Teodoro et al. [25] present this technique as an effective solution for reducing data complexity without losing essential information. It identifies comprehensive influencing factors while maintaining the basic characteristics of the data [26–29]. In other words, the Principal Components (PC) are characterized by uncorrelated representations of the original correlated variables [30]. According to Equation (1), data centering is performed as follows:

$$X_C = X - \overline{X} \tag{1}$$

where $X$ represents the original data matrix and $\overline{X}$ is the mean of $X$.

Next, as shown in Equation (2), the covariance matrix is calculated by:

$$C = \frac{1}{n-1} X_C^T X_C \tag{2}$$

where $n$ is the number of observations.

Then, according to Equation (3), eigenvalue and eigenvector decomposition is performed:

$$Cv_i = \lambda_i v_i \tag{3}$$

where $\lambda_i$ and $v_i$ are the eigenvalues and eigenvectors of the covariance matrix $C$.

Data transformation is achieved by projecting the centered data onto the eigenvectors, as described in Equation (4):

$$Z = X_C V \tag{4}$$

where $V$ is the matrix of eigenvectors.

These steps provide a transformed dataset $Z$ that captures the essential structure of the original data with reduced dimensionality.

#### 2.2.2 *Varimax Factor Analysis*

Following a similar approach, Factor Analysis (FA) with varimax rotation is used for dimensionality reduction by identifying latent factors that explain most of the variability in the original data [31]. These factors are constructed from weights among the factors, facilitating pattern interpretation. This method is widely used in environmental studies to simplify complex datasets and create factors with clusters of inter-correlated variables [32, 33]. As shown in Equation (5), the factor loadings matrix ($\mathbf{L}$) is defined as:

$$L = [l_{ij}] \tag{5}$$

where $l_{ij}$ represents the factor loading of the $i$-th variable on the $j$-th factor.

According to Equation (6), the varimax criterion $V$ is given by:

$$V = \sum_{j=1}^{m} \left[ \frac{1}{p} \sum_{i=1}^{p} \left( l_{ij}^2 - l_j^2 \right)^2 \right] \tag{6}$$

where $l_j^2$ is the sum of squared loadings for the $j$-th factor, $p$ is the number of variables, and $m$ is the number of factors.

### 2.3 Classification Models

To overcome traditional statistical approaches, which can be costly due to the quantity of data, ML methods for data classification have been developed, as presented below.

#### 2.3.1 *Logistic Regression*

LR is a technique of statistics commonly employed for classification purposes, aiming to predict the probability of an event by analyzing datasets, as presented by Townsend et al. [34] and used by Pradhan et al. [35]. In contrast to linear regression and other approaches focused on continuous variables, LR uses a logistic function to represent probabilities, transforming a set of independent variables into probabilities ranging from 0 to 1. The coefficients in LR signify the impact of the independent variables on the event's probability, providing a straightforward explanation of how each variable influences the prediction. As shown in **Equation (7)**, the LR function $\sigma(z)$ is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

where $z$ is the linear combination of the independent variables.

According to **Equation (8)**, $z$ is the linear combination of the independent variables:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{8}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients of the independent variables, and $x_1, x_2, \ldots, x_k$ are the independent variables.

**Equation (9)** represents the probability prediction:

$$\begin{aligned} &P(Y = 1 \mid x_1, x_2, \ldots, x_k) \\ &\quad = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}} \end{aligned} \tag{9}$$

where $P(Y = 1 \mid x_1, x_2, \ldots, x_k)$ is the predicted probability of the event occurring.

**Equation (10)** shows the LR model:

$$\log\left(\frac{P(Y=1 \mid x)}{1 - P(Y=1 \mid x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{10}$$

**Equation (11)** presents the likelihood function $L(\beta)$ used for parameter estimation:

$$L(\beta) = \prod_{i=1}^{n} P(y_i \mid x_i)^{y_i} \left(1 - P(y_i \mid x_i)\right)^{1-y_i} \tag{11}$$

where $y_i$ is the observed outcome for the $i$-th observation.

### 2.3.2 Support Vector Machine

The SVM model was introduced by Vapnik [36], aiming to maximize the distance between support vectors and the hyperplane to achieve optimal classification performance. It transforms data into a higher-dimensional space using a kernel function, enabling the processing of non-linearly separable data [37, 38]. This mapping makes it possible to divide instances of each class by a margin denoted by a hyperplane, thus enabling the detection of the hyperplane to distinguish possible outcomes and predict corresponding classes [39]. As shown in **Equation (12)**, the Lagrangian and duality form of the SVM optimization problem is given by:

$$\text{Max} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{12}$$

Subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{13}$$

$$0 \leq \alpha_i \leq C \quad \forall i \tag{14}$$

where $\alpha_i$ are the Lagrange multipliers, $y_i$ are the class labels, $\mathbf{x}_i$ are the input vectors, and $C$ is the regularization parameter.

**Equation (13)** shows the decision function $f(\mathbf{x})$:

$$f(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\right) \tag{15}$$

where $\mathbf{w}$ is the weight vector, and $b$ is the bias term.

**Equation (14)** presents the kernelized version of the decision function [40]:

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{16}$$

where sgn is the sign function, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function.

### 2.3.3 K-Nearest Neighbors

Used in both classification and regression scenarios, as in the case of Getin et al. [41], KNN is a non-parametric method that involves inserting a new data point, and the algorithm searches for the closest points to it in the training data. The prediction for the new point is then based on its classification. Adjusting the point or hyperparameter $K$ can lead to underfitting or overfitting [42]. The distance metric can vary, and the ideal value is determined through cross-validation on the training set to maximize model accuracy [43]. As shown in Equation (15), the distance $d(x, x_i)$ between the new sample and all dataset points is calculated as:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \tag{17}$$

where $d(x, x_i)$ is the Euclidean distance between the new sample $x$ and the $i$th sample $x_i$ in the dataset, and $x_j$ and $x_{ij}$ are the $j$th features of the new sample $x$ and the $i$th sample $x_i$, respectively.

For classification problems, the predicted label $\hat{y}$ is determined by the mode of the labels of the KNN:

$$\hat{y} = \text{mode}(y_{i_1}, y_{i_2}, \ldots, y_{i_k}) \tag{18}$$

where $\hat{y}$ is the predicted label or value for the new sample, and $y_{i_1}, y_{i_2}, \ldots, y_{i_k}$ are the labels of the KNN.

For regression problems, the predicted value $\hat{y}$ is the average of the values of the KNN:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} y_{i'} \tag{19}$$

where $K$ is the number of nearest neighbors considered.

*2.3.4 Random Forest*

The RF algorithm is widely used in ML for tasks involving classification and regression. Its methodology involves combining multiple decision trees, each trained using a random subset of the dataset, to increase the accuracy and resilience of predictive results [44]. An example of its execution is the work by Chen et al. [45], who used RF to develop prediction models for daily concentrations of PM2.5, PM10, NO2, and O3 MDA8 in Great Britain. As shown in **Equation (18)**, the RF algorithm operates as follows:

- Bootstrap Samplings: Using random sampling with replacement, several subsets of the dataset are produced;
- Feature Selection for Splitting: To find the optimal split for each decision tree, a random subset of features is chosen at each split.

For classification problems, the predicted label $\hat{y}$ is determined by majority voting among the decision trees:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_m(x)\} \tag{20}$$

where $T_i(x)$ is the prediction made by the $i$-th decision tree for the input $x$, $m$ is the total number of decision trees in the RF.

For regression problems, the predicted value $\hat{y}$ is the average of the predictions of all the decision trees:

$$\hat{y} = \frac{1}{m} \sum_{i=1}^{m} T_i(x) \tag{21}$$

*2.3.5 Decision Tree*

A DT, as used by Teodoro et al. [25] is a ML model used for classification and regression, structured with internal nodes and leaf nodes. It splits the data based on criteria such as entropy and Mean Squared Error (MSE) reduction. Decision trees are easy to interpret and do not require data normalization, but they can suffer from overfitting and instability, which can be mitigated with pruning or multiple trees, such as in RF. Each internal node represents a test based on a feature, with two or more possible outcomes, leading to another internal node or a leaf node that indicates the predicted class. A path from the root node to a leaf node forms a rule that predicts a class [46]. As shown in **Equation (20)** for classification and in **Equation (21)** for regression.

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{22}$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{23}$$

where $p_i$ is the proportion of class $i$ examples in set $S$, $y_i$ is the actual value, $\hat{y}$ the predicted value, and $n$ is the number of examples.

So, the information gain for a feature $A$ is defined as:

$$\begin{aligned} &\text{Information Gain}(S, A) \\ &= \text{Entropy}(S) - \sum_{\nu \in \text{Values}(A)} \frac{|S_\nu|}{|S|} \text{Entropy}(S_\nu) \end{aligned} \tag{24}$$

where $S_\nu$ is the subset of $S$ where feature $A$ has value $\nu$.

*2.3.6 Gradient Boosting*

Complementary to the decision tree methodology, GB is a ML method that uses decision trees to sequentially create predictive models. To build a robust ensemble model, it combines several weak predictive models [47]. The method begins with a simple model and then iteratively improves it by focusing on the errors found in previous models, so each new model corrects the residuals or errors

of the ensemble to reduce the overall prediction error [48]. This method allows GB to achieve extremely accurate levels in tasks such as regression and classification. Due to its effectiveness in handling complex relationships in data, it is popular in various fields as shown in subsequent equations. The first one is the initialization model with a constant that minimizes the loss function.

$$F_0(x) = \arg\min_y \sum_{i=1}^{n} L(y_i, y) \tag{25}$$

where $L$ is the loss function and $y_i$ are the target values.

The iterative process is carried out, calculating the residuals in the Equation (24).

$$r_i^{(m)} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right] \tag{26}$$

where $F_{m-1}(x_i)$ is the prediction of the model in the previous iteration.

Fit a new tree $h_m(x)$ to the residuals $r_i^{(m)}$:

$$h_m(x) \approx r_i^{(m)} \tag{27}$$

Update the model:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{28}$$

where $\eta$ is the learning rate, which controls the contribution of each tree.

After $m$ iterations, the final model is described by Equation (27).

$$F_m(x) = F_0(x) + \sum_{m=1}^{M} \eta h_m(x) \tag{29}$$

### 2.4 Validation Methods

The Leave-One-Out (LOO) technique is a form of cross-validation where, in each iteration, a single observation is left out of the training dataset and used as the test set. This process is repeated for each observation in the dataset, ensuring that each sample is used once as the test set. LOO allows evaluating how the model performs in each iteration, providing an unbiased estimate of its performance. Additionally, it helps in comparing algorithms and detecting overfitting, being widely used in various studies, as evidenced by Lephalala et al. [49] and by De Meester et al. [50]. LOO is given by:

$$\text{LOO} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_{-i}(x_i)) \tag{30}$$

where $n$ is the total number of observations in the dataset, $L(y_i, \hat{f}_{-i}(x_i))$ is the loss function evaluating the difference between the true value $y_i$ and the prediction $\hat{f}_{-i}(x_i)$ made by the model trained on the dataset excluding the $i$th observation, $\hat{f}_{-i}(x_i)$ represents the prediction for $x_i$ using the model trained without the $i$th observation.

Classifying data is essential, but without validation, it is impossible to guarantee the robustness, effectiveness, and generalizability of predictive models, avoiding overfitting or underfitting issues. Thus, some methods ensure this accuracy, such as LOO and K-Fold, among others. It is evident that LOO is a versatile method that allows data classification and cross-validation and applying other methods for comparison will bring better reliability. Therefore, methods like K-Fold Cross-Validation, widely used in statistics to evaluate the performance of a predictive model, help validate the model's ability to generalize to unseen data, reducing bias and variability that can occur when dividing the dataset into training and testing [51]. K-Fold is given by:

$$\text{K-Fold} = \frac{1}{K} \sum_{k=1}^{K} L_k \tag{31}$$

where $K$ is the number of folds, $L_k$ is the loss for the $k$th fold, where every fold serves as a validation once, with the remaining $K-1$ constituting the training set.

### 2.5 Evaluation Metrics

Evaluation metrics are essential in the analysis and improvement of ML models and data science, providing systematic methods to assess and validate predictive performance and the generalization ability of models. In classification problems, commonly used metrics include confusion matrix, accuracy, F1-score, precision, and recall.

The confusion matrix accounts for the correct and incorrect predictions made by the model concerning class classifications [52]. Accuracy is a metric that expresses how many of the total predictions are accurate. Recall is the percentage of correctly predicted examples to all truly positive examples, whereas

Precision shows the percentage of correct predictions to positive predictions. The F1-score is the harmonic mean between precision and recall, ideal for situations with class imbalance [53, 54]. Table 2 displays the evaluation metrics.

**Table 2.** Evaluation metrics.

| Metrics | Evaluation |
|---|---|
| Confusion Matrix | It is a 2x2 table for binary problems, where columns show the model's predicted classes and rows show the actual classes: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). |
| Accuracy | $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (32)$$ |
| Precision | $$\text{Precision} = \frac{TP}{TP + FP} \quad (33)$$ |
| Recall | $$\text{Recall} = \frac{TP}{TP + FN} \quad (34)$$ |
| F1-Score | $$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (35)$$ |

In regression problems, metrics such as the Coefficient of Determination ($R^2$), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) are used, as utilized by Vasconcelos et al. [55]. The percentage of the data's variance that the model can explain is indicated by the $R^2$. Whereas MSE computes the mean of the squared errors, MAE computes the average of the absolute errors between predicted and actual values. RMSE is the square root of MSE, providing an error measure in the same unit as the original data, and MAPE calculates the average of absolute differences between predicted and actual values in percentage terms.

## 3 Methodology

This section is divided into two parts: i) how the data was collected and ii) the application of ML algorithms for the classification of species in the genus Capsicum, focusing on the varieties *C.* frutescens and *C.* chinense.

### 3.1 Experimental Methodology

The first step was to plant the peppers. The seeds were planted in polystyrene trays with 128 cells, using commercial Bioplant® substrate. The seedlings, with fewer than six true leaves, were transplanted into five-liter pots containing the same type of substrate. Monthly fertilization was carried out with N-P-K (4-14-8). Seventeen pepper genotypes belonging to two species of the Capsicum genus were used, as presented in Table 3.

**Table 3.** Peppers species and genotypes (varieties) used in the work.

| Scientific Name | Varieties |
|---|---|
| *Capsicum frutescens* | • Malagueta<br>• Tabasco<br>• Cayenne<br>• Etna ornamental<br>• Pirâmide Ornamental |
| *Capsicum chinense* | • Bhut jolokia<br>• Cumari<br>• Habanero chocolate<br>• Habanero vermelha<br>• Habanero amarela<br>• Peito de moça<br>• Murupi<br>• Pião zinho<br>• Cheiro do norte<br>• Biquinho vermelha<br>• Biquinho amarela<br>• Arari bode amarela |

It provides a list of pepper varieties, including the species *C. frutescens* (*Cayenne, Tabasco, Peter, Malagueta, Etna Ornamental, and Pirâmide Ornamental*) *and C. chinense* (*Bhut Jolokia, Biquinho Amarela, Habanero chocolate, Habanero Amarela, Habanero Vermelha, Cumari, Peito de Moça, Murupi, Pião zinho, Cheiro do Norte, Biquinho Vermelha, Biquinho Amarela, Arari Bode Amarela*).

The variables defined for the dataset were selected as quantitative and qualitative descriptors. The quantitative descriptors include fruit weight ($x_{10}$), fresh mass of mature seedless fruit ($x_9$), collected using an analytical balance (model Bel M214Ai) with results expressed in grams, and dry mass of mature seedless fruit ($x_{14}$), measured after drying the fruits in an oven (model SSDcr) for 72 hours at 60 °C with ventilation, following the same procedure used for fresh mass data collection.

The length of the mature fruit ($x_{19}$), diameter of the mature fruit ($x_3$), pericarp thickness ($x_7$), corolla diameter ($x_{18}$), leaf size ($x_8$) was measured with a caliper (Stainless Hardened, precision of 0.01 mm) at the median portion of the fruits, with results expressed in millimeters (mm). The number of seeds per fruit

**Figure 1.** Data collection of *Capsicum spp.*

($x_{20}$) was obtained by counting the seeds. Plant height ($x_{15}$), crown diameter ($x_1$), stem diameter ($x_5$), and stem length ($x_{16}$) were measured using a measuring tape, with results expressed in centimeters (cm).

For qualitative descriptors, the evaluated characters were: fruit shape ($x_4$), branch density ($x_{11}$), leaf shape ($x_2$), corolla color ($x_{12}$), immature fruit color ($x_{17}$), intermediary fruit color ($x_6$), and mature fruit color ($x_{13}$), following the morphological-agronomic characterization according to the International Plant Genetic Resources Institute (IPGRI) (1995). Branch density was classified as sparse, intermediate, or dense. Leaf shape was categorized as deltoid, oval, or lanceolate. Corolla color was classified as white, light yellow, yellow, yellow-green, purple with white base, white with purple base, white with purple margin, or purple. Immature fruit color was recorded as white, yellow, green, orange, purple, or dark purple. Mature fruit color was described as white, lemon yellow, pale orange-yellow, orange-yellow, pale orange, orange, light red, red, dark red, purple, brown, or black. In each case, the values were represented as the average of triplicate determinations.

Descriptors were evaluated at two distinct moments: some from the appearance of the first flower, while others were measured when more than 50% of each plant had at least one mature fruit, following the protocol of IPGRI [56]. A summarized photographic record of the entire data collection process for the case study is illustrated in Figure 1, which includes 85 observations: 25 from C. frutescens and 60 from C. chinense. Each genotype was represented by 5 observations.

Figure 2 illustrates the 17 fruits of the two studied species, C. chinense and C. frutescens, where: (a) *Malagueta*, (b) *Tabasco*, (c) *Cayenne*, (d) *Etna Ornamental*, (e) *Pirâmide Ornamental*, (f) *Bhut Jolokia*, (g) *Cumari*, (h) *Habanero Chocolate*, (i) *Habanero Vermelha*, (j) *Habanero Amarela*, (k) *Peito de Moça*, (l) *Murupi*, (m) *Piãozinho*, (n) *Cheiro do Norte*, (o) *Biquinho Vermelha*, (p) *Biquinho Amarela*, and (q) *Arari Bode Amarela*.

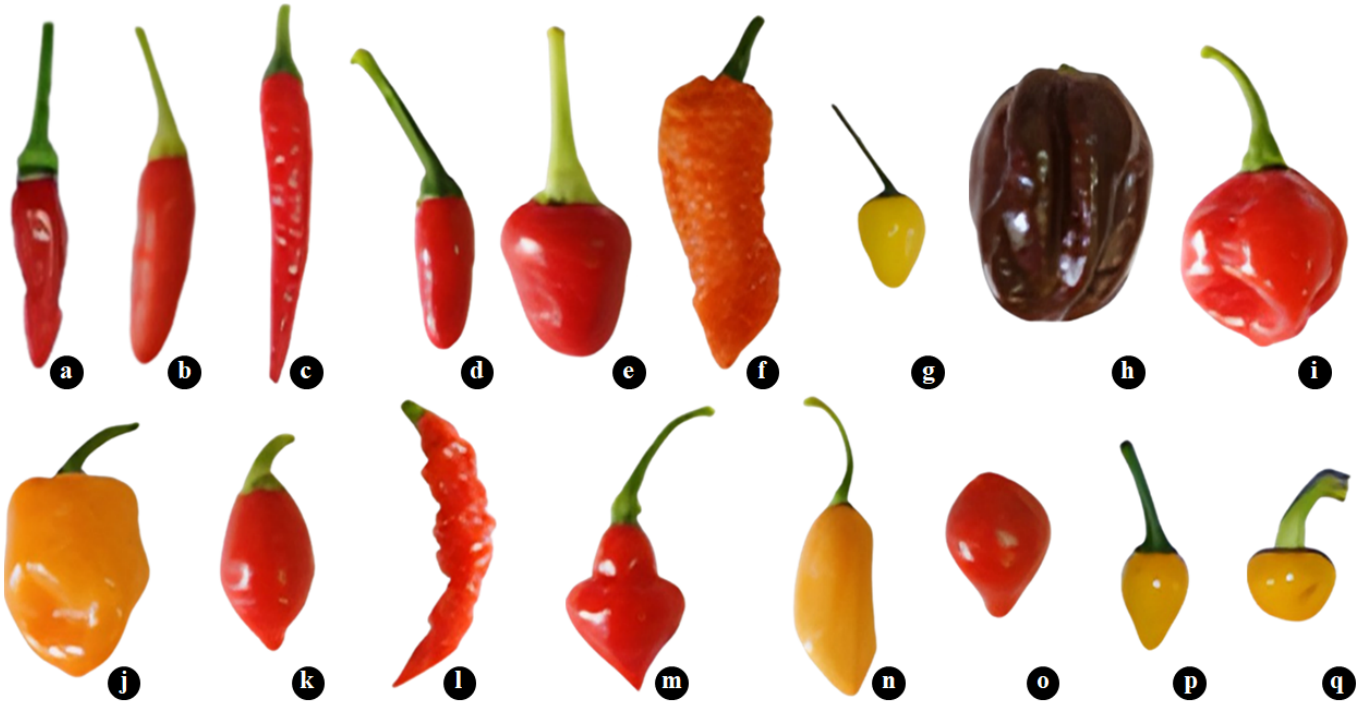


**Figure 2.** Fruits of *Capsicum spp.*

### 3.2 Numerical Methodology

The numerical methodology was designed to develop effective predictive models for classifying pepper species, with a focus on parameter adjustment and mitigating potential overfitting, which is common when working with small datasets. The process was divided into key stages, such as data preprocessing, dimensionality reduction, model selection and training, and result validation. This entire process was carried out using Python, utilizing various libraries and tools to implement each stage effectively.

The data were preprocessed to ensure data integrity and quality. It was necessary to encode categorical variables by transforming them into numerical formats for the application of ML models. This was done using the One-Hot-Encoder, which creates a binary column (0 or 1) for each category of the variables, allowing the information to be properly represented for the model. The independent variables (features) were separated from the dependent variable (target), ensuring a clear distinction between predictors and the prediction target. This step is crucial to avoid processing errors and ensure that the models receive clean and structured data. The data were split into training and testing sets, with 80% used for training and 20% for testing.

The target variable (species) was encoded as an integer numeric variable, with values of 0 or 1. The variables genotypes, leaf shape, branch density, corolla color, fruit shape, immature fruit color, and mature fruit color were transformed into dummy categorical variables for one hot encoding, resulting in a total of 60 columns.

PCA was performed, followed by FA with Varimax rotation to reduce data dimensionality. It's important to note that a comparison was also made between the results obtained from PCA and those from FA regarding evaluation metrics.

A variety of ML models were used, such as LR, SVM, KNN, RF, DT, and GB. These algorithms were intentionally selected to cover a spectrum ranging from simple and interpretable models (LR, DT), to ensemble-based approaches with higher computational cost and robustness (RF, GB), distance-based methods (KNN), and margin-based classifiers (SVM). This diversity allowed a more comprehensive comparison of model behaviors under the same dataset conditions. Each model was tuned using Optuna, a hyperparameter optimization tool that employs techniques like Bayesian optimization to find the best parameter configuration. The goal was to balance complexity and performance, while avoiding overfitting.

To explicitly address overfitting risk due to the small dataset size, LOO validation was applied, as it provides nearly unbiased estimates by using each sample

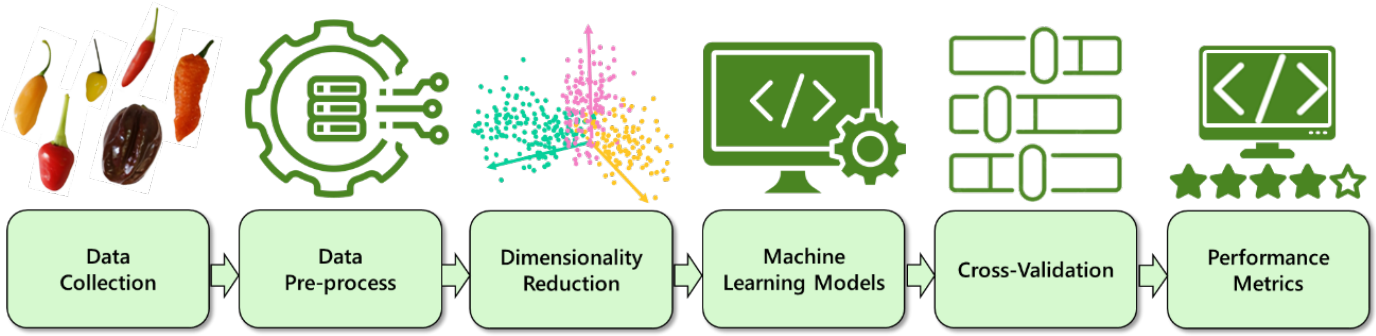


**Figure 3.** Methodology framework.

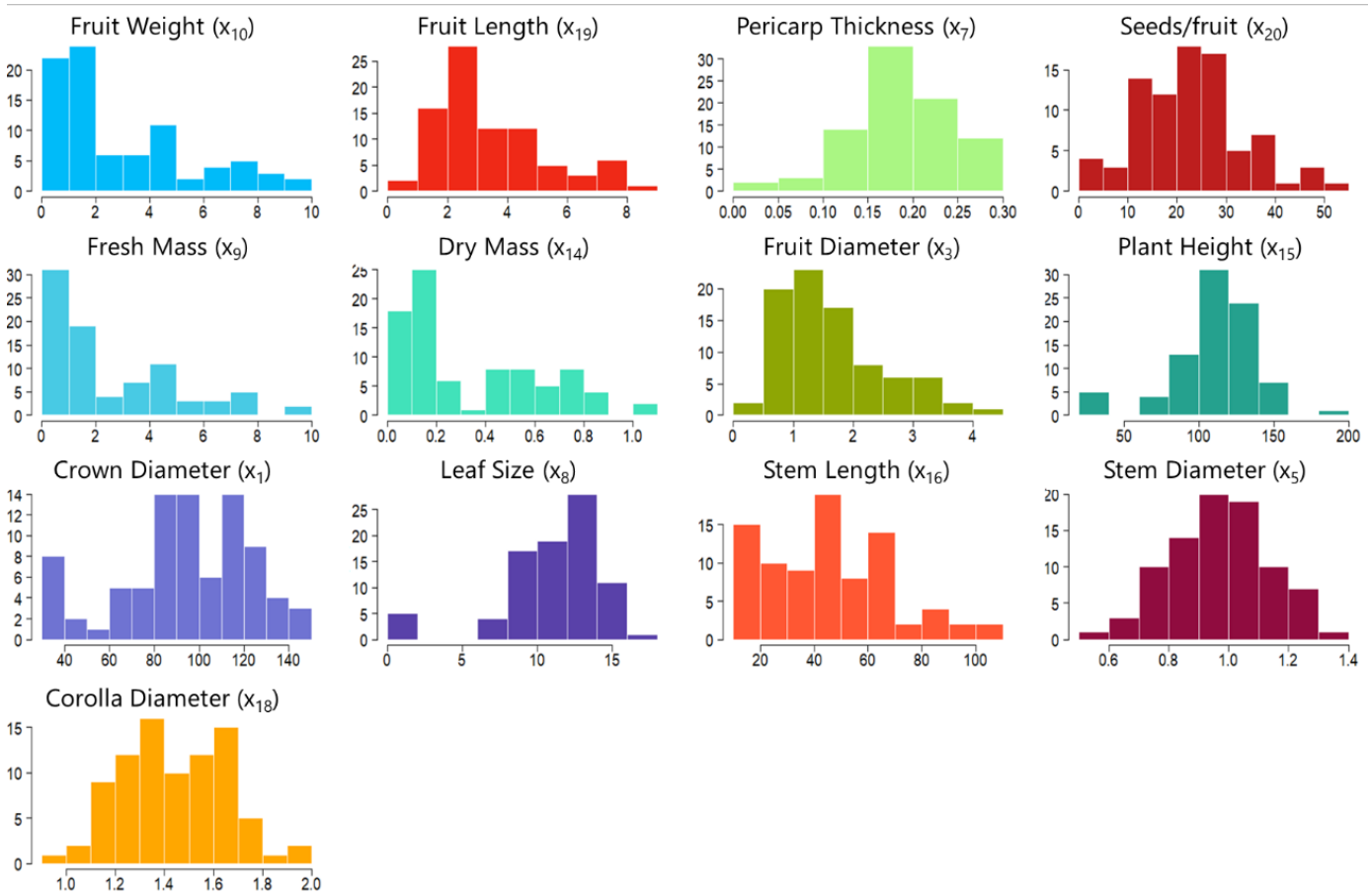


**Figure 4.** Histogram for quantitative variables used as Machine Learning inputs.

once as a test case. To strengthen the evaluation, LOO was combined with Stratified K-Fold, ensuring that class balance was preserved during validation. This methodological choice ensured greater reliability in performance estimation compared to traditional K-Fold alone.

To verify the quality of the models, performance metrics such as accuracy, precision, recall, and F1-Score were used. These metrics were calculated from the predictions obtained during the validations, allowing comparisons among model. Figure 3 summarizes the step-by-step process of the proposed methodology.

## 4 Application of Analysis, Challenges and Results

Before applying ML classification techniques, an initial exploratory data analysis was conducted. It began with descriptive statistics, followed by histograms to investigate the distribution patterns of numerical features. Subsequently, the analysis deepened to assess the suitability of the chosen algorithms. Figure 4 illustrates these histograms concerning the numerical variables.

The correlation between numeric variables was assessed, as shown in Figure 5, highlighting those

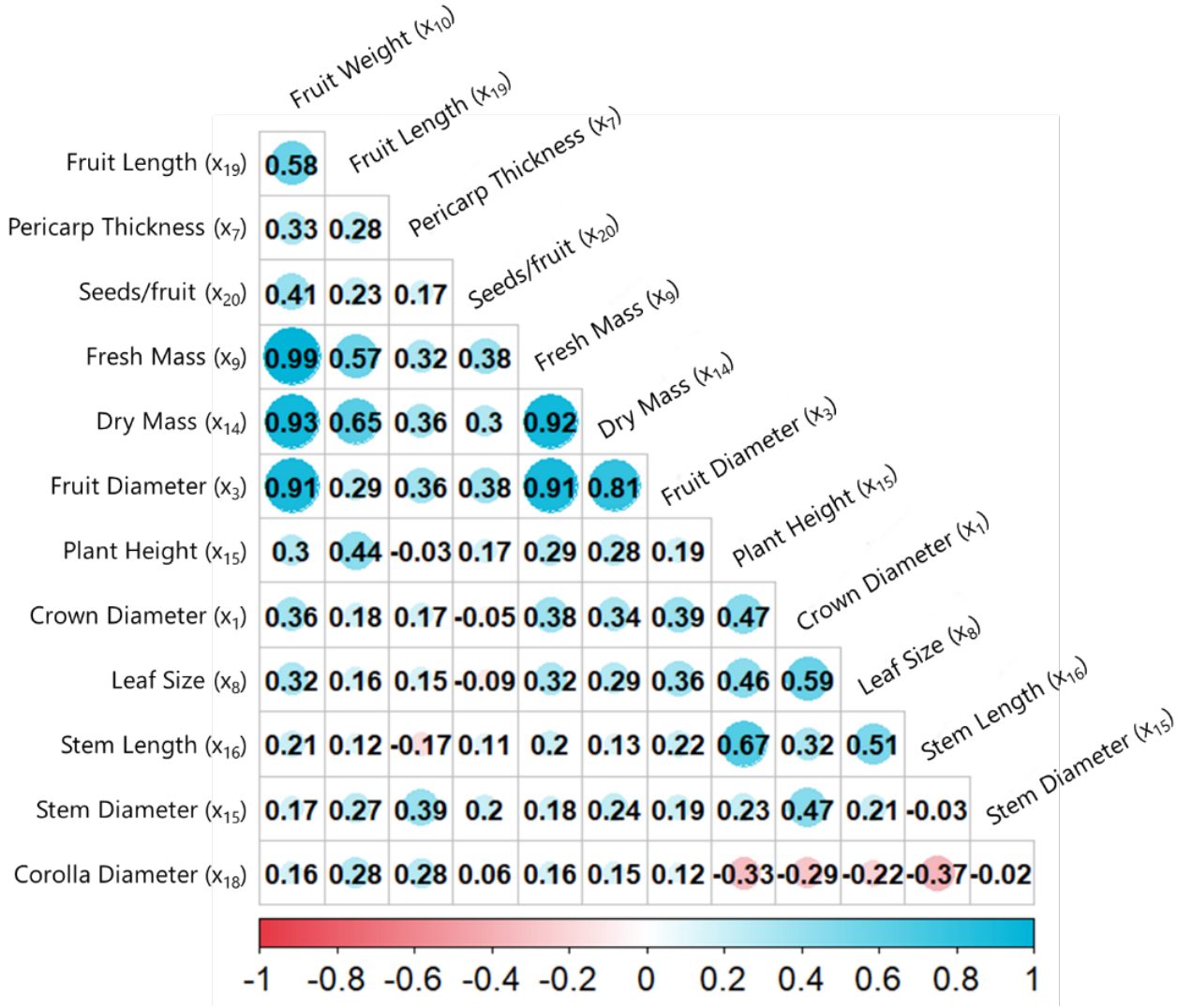


**Figure 5.** Correlation matrix for quantitative variables.

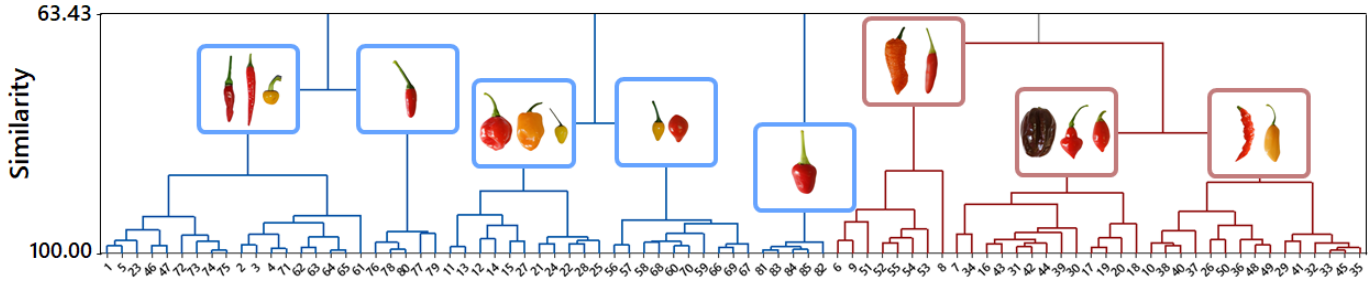


**Figure 6.** Dendrogram for the clusters of Capsicum genotypes.

with a correlation coefficient above 0.70. In addition to verifying the variables, a cluster analysis was also performed on the observations using the Ward Linkage method and Euclidean distance. Based on this, the dendrogram was used to assess species separability of the species, as illustrated in Figure 6.

Due to the high dimensionality following variable encoding, PCA was applied to the standardized variables, capturing 80% of the explained variance and selecting eigenvalues greater than 1, as suggested by Kaiser [57]. This reduced the number of variables to 11. Following PCA, FA with Varimax rotation was conducted, retaining the number of factors determined by PCA. Figure 7(a) illustrates the number of PC and the proportion of explained variance, while Figure 7(b)

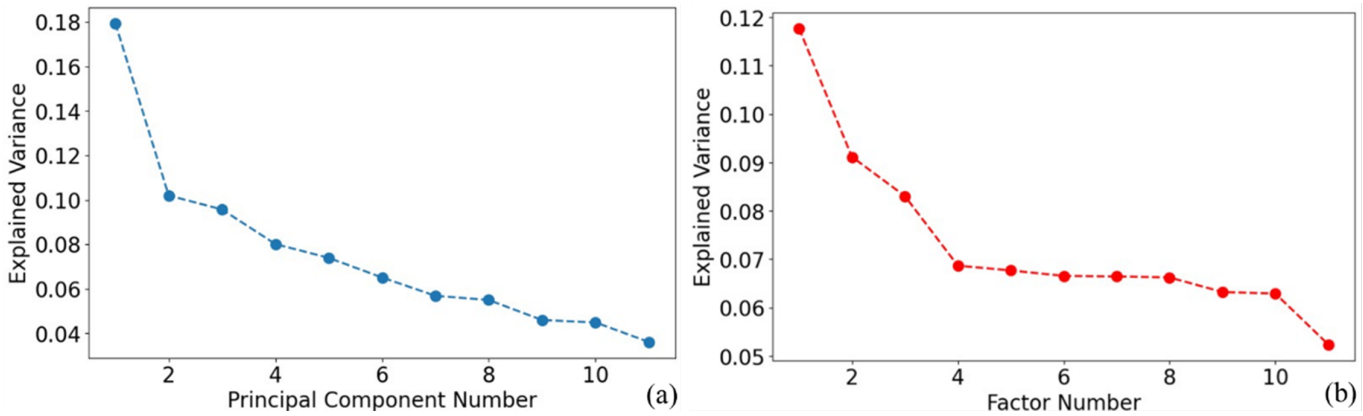


**Figure 7.** Proportion of explained variances to the principal components and factors.

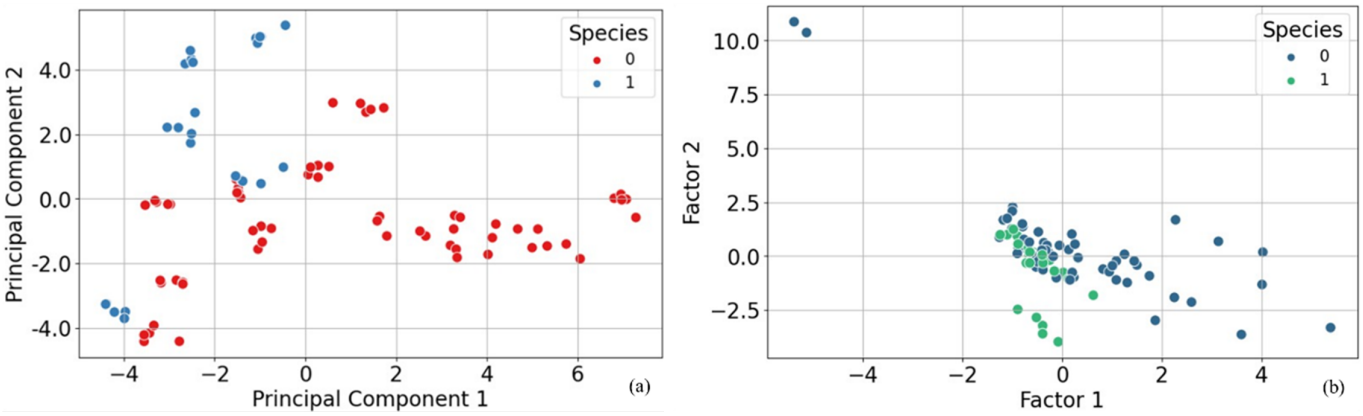


**Figure 8.** Species distribution according to the two principal components and factors.

presents the same for rotated factors. The Figure 8(a) displays the dispersion of species across the first two PC, with Figure 8(b) depicting this for rotated factors.

In the next stage, multiple models were trained and tested, including LR, SVM, KNN, RF, DT, and GB. The choice of these algorithms was based on the intention to cover both simpler and more interpretable methods, such as LR and DT, and more robust and computationally demanding ones, such as RF, SVM, and GB.

- Simpler models, like LR and DT, offer interpretability and low computational cost, serving as useful baselines;
- Ensemble-based models, like RF and GB, provide greater generalization capacity, though at a higher computational cost and with greater sensitivity to parameterization;
- Margin-based models, such as SVM, are robust for class separation in multidimensional spaces, though they demand careful parameter tuning;
- Distance-based methods, like KNN, are intuitive and easy to apply, but may be sensitive to noise and dimensional variability.

Given the small dataset size, traditional K-Fold might not provide precise estimates. Therefore, LOO validation was applied, where each observation serves as a test, being more appropriate for small datasets and helping to mitigate overfitting. For greater robustness, stratified cross-validation was also combined with LOO.

Initially, tests were carried out using the original variables, without dimensionality reduction. At this stage, results were less favorable, with overfitting occurring in some models, evidenced by discrepancies between training and test performance. PCA partially reduced this issue but inconsistencies persisted. The use of FA with Varimax rotation proved to be the most effective approach, eliminating overfitting and yielding high performance across all models.

Table 4 presents the results obtained after FA, showing that simpler models (e.g., LR) and more robust ones (e.g., SVM) achieved similarly superior performances compared to the others. This finding suggests that

the optimal model selection depends not only on algorithmic complexity but also on the adequacy of preprocessing and dimensionality reduction methods to the dataset.

**Table 4.** Results of classification models using Factor Analysis with varimax rotation.

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 0.9765 | 0.9782 | 0.9775 | 0.9767 |
| Random Forest | 0.9059 | 0.9170 | 0.9070 | 0.8999 |
| Decision Tree | 0.8941 | 0.8998 | 0.8940 | 0.8958 |
| Support Vector Machine | 0.9765 | 0.9772 | 0.9647 | 0.9762 |
| K-Nearest Neighbors | 0.9647 | 0.9646 | 0.9657 | 0.9645 |
| Gradient Boosting | 0.9176 | 0.9189 | 0.9175 | 0.9181 |

In addition to evaluating classification metrics, we also conducted an interpretability analysis to identify which variables most contributed to the model predictions. For this purpose, the SHAP (SHapley Additive exPlanations) method was applied, which quantifies the contribution of each variable to the models' decision-making process.

The results (Figure 9) show that crown diameter ($x_1$), leaf size ($x_8$), leaf shape ($x_2$), pericarp thickness ($x_7$), fruit diameter ($x_3$), fruit weight ($x_{10}$), stem diameter ($x_5$), and corolla color ($x_{12}$) were the most important variables, playing a decisive role in classifying *Capsicum* species. Traits related to fruit biomass (fresh mass ($x_9$) and dry mass ($x_{14}$)) also exhibited high relevance, reinforcing that morphological, biometric, and qualitative descriptors are fundamental for genotype differentiation.

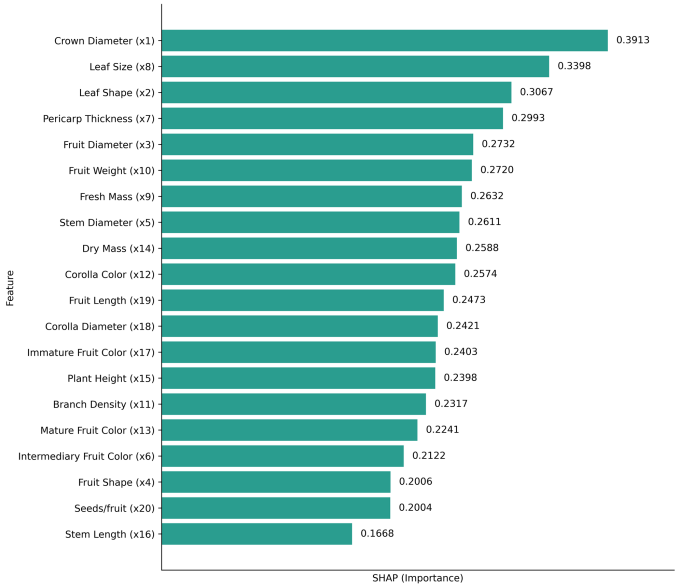


**Figure 9.** Contribution of features to model predictions according to SHAP analysis.

This analysis not only complements accuracy results but also provides biological insight, highlighting which morphological and fruit descriptors are most discriminative, with direct implications for breeding and genetic improvement programs.

## 5 Discussions

### 5.1 Conditioning Models

It is worth noting that without using Stratified K-Fold together with LOO, the performance with original variables, PCA, and FA was very poor, showing overfitting in at least five out of six models. This highlights that, in terms of ML models, not using Stratified K-Fold together with LOO imposes limitations on the evaluation of the dataset. Thus, adding mechanisms such as those presented here provides an alternative to improve the classification performance of the dual-path approach with attention mechanisms (DPACR) model by Zhang et al. [58].

In this study, the diversity of algorithms tested proved fundamental to assessing how model complexity interacts with overfitting control. Simpler models, such as Logistic Regression and Decision Tree, provided interpretability and lower computational costs, while ensemble-based models (Random Forest and Gradient Boosting) aimed at greater robustness but showed higher sensitivity to overfitting in small datasets. Support Vector Machines presented strong separation capacity, although requiring careful parameterization, whereas KNN offered intuitive classification but was affected by dataset size and dimensionality. In addition, the SHAP interpretability analysis reinforced that morphological and biometric traits, such as crown diameter, leaf size, and pericarp thickness, played a decisive role in the predictions, corroborating the biological consistency of the obtained models.

Using FA with Varimax rotation as proposed, all methods achieved results above 0.85 in accuracy, precision, recall, and F1-Score. The best results were achieved with LR and SVM, respectively. For the original variables, LR performed the best, with results like those using FA, while for PCA, DT showed the best performance, also similar to FA.

These findings indicate that overfitting control was not only achieved through validation strategies (LOO and Stratified K-Fold) but also strongly influenced by dimensionality reduction and model choice. Importantly, the results reinforce that simpler algorithms can perform on par with more robust methods when preprocessing and validation are

properly applied.

Thus, as shown in the work of Zhao et al. [59], the use of K-Fold indicated that the ML models exhibited strong generalization capabilities, with minimal variations in MAE and RMSE. This aligns with the presented work, based on the metrics of confusion matrix, accuracy, precision, and F1-score. The results indicate that the models performed well in classification tasks. Precision quantifies the ratio of TP to all positive predictions, while accuracy reflects the percentage of correct predictions relative to all predictions. The F1-Score balances precision and recall by calculating their harmonic mean. Recall, also known as sensitivity, quantifies the percentage of TP compared to all actual positives.

The initial hypothesis was to verify whether training a ML model could classify *Capsicum spp.* species using biodescriptors as an alternative to image recognition algorithms, which typically require more labeled data and greater effort. The idea was to validate this initial hypothesis by training models with tabular data instead of using images. The main challenge was the considerably small amount of data. In the end, based on the metric results, it was found that virtually all models are good options for this classification function when using FA.

### 5.2 Study Object and its Limitations

The study object presented consists of a database developed in a laboratory following planting steps. Thus, it is easy to understand that generating experimental repetitions or even replicas, when working with tabular data, is a very challenging task, as it requires labor, growth time, fertilization, and specific care to prevent diseases. Furthermore, some experiments are destructive in nature (i.e., measurements that require cutting the seedlings), which can lead to errors and often to the need for sample discards.

Another important aspect is that the dataset was obtained under controlled conditions, using the same soil type, cultivation environment, and agricultural management practices. These variables were kept fixed throughout the experiment in order to reduce external interferences and ensure consistency of the results. However, this approach limits the diversity of scenarios considered and restricts the generalization of the models to environmental conditions, soil types, and management practices that were not included in this study.

Another limitation of the study object relates to the qualitative characterization of plant-specific factors, which may vary across technicians, even when following a standard protocol. Therefore, the use of classification models based on machine learning, as presented by Li et al. [60], could facilitate this stage in future work, in addition to enabling analyses based on other classifications, such as cluster size and main stems.

### 5.3 Future Studies

Future studies will focus on genotype identification and the application of computer vision, through the development and implementation of sophisticated algorithms capable of analyzing genetic markers and patterns with high precision. This approach aims to automate phenotypic recognition, significantly reducing the time and effort required for manual analyses, as well as allowing the processing of large genomic datasets to identify subtle variations associated with desirable traits in peppers.

In addition, we intend to expand data collection to include plants cultivated under different environments, soil types, and growth stages. This expansion will make it possible to evaluate the robustness and generalization capacity of the proposed models, as well as to enable direct comparisons with the results obtained in the present study. In this way, it will be possible to verify to what extent the models developed under controlled conditions remain effective in more complex and diverse scenarios, contributing to broader practical applications in breeding programs and agricultural management.

## 6 Conclusion

This article explores the development of a ML algorithm aimed at classifying *Capsicum spp.* data, playing a crucial role in breeding programs, particularly in effective diversity management. Despite the limited sample size compared to ML methodologies using large datasets, the results show that the algorithms can accurately classify *Capsicum spp.* species. Furthermore, the study compares different approaches, such as PCA and FA with varimax rotation, providing valuable insights for advances in the field. With the main points of the study being:

- Developed a ML algorithm for classifying *Capsicum spp.* data, supporting breeding programs and diversity management. Despite

the limited sample size, the algorithms proved effective in classifying *Capsicum spp.* species;

- Detailed analysis revealed differences between using original variables, reducing dimensionality with PCA, and FA with varimax rotation;
- Overfitting was observed in some models using original variables and PCA. FA with varimax rotation proved high efficacy, though less commonly explored compared to PCA;
- The use of multiple ML models aims to compare the performance of each, as each model has different assumptions, generalization capabilities, and sensitivities to specific patterns in the data. This approach allows for the verification of result consistency and a deeper understanding of the dataset's characteristics;
- All six models performed well with FA, achieving results above 0.89 in all evaluated metrics, with LR and SVM standing out with results above 0.96;
- Cross-validation with LOO and Stratified K-Fold reinforced the findings; without these techniques, overfitting appeared in three tests;
- Positive outcomes were obtained using tabular data, which typically yield inferior results compared to computer vision techniques;
- Models achieved high performance, although none reached a perfect score of 1.00 in any metric;
- The study emphasizes the importance of advanced innovation systems to optimize agricultural processes, improve crop management, and provide valuable tools for farmers and researchers;
- These systems not only enhance agricultural efficiency and reduce human errors but also significantly elevate crop management practices;
- The hypothesis regarding the challenge of training models with tabular data was confirmed.

## Data Availability Statement

The dataset is available upon request and can also be accessed at the following link: https://github.com/Matheuscp98/PepperCapsicum.

## Funding

This work was supported in part by the FAPEMIG under Grant BPD-01045-22; in part by the CAPES; in part by the CNPq; and in part by the NOMATI–UNIFEI, which provided access to laboratories, materials, and technical expertise.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Ethical Approval and Consent to Participate

Not applicable.

## References

[1] Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., ... & Choi, D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics, 46*(3), 270-278. [CrossRef]

[2] Menichini, F., Tundis, R., Bonesi, M., Loizzo, M. R., Conforti, F., Statti, G., ... & Menichini, F. (2009). The influence of fruit ripening on the phytochemical content and biological activity of Capsicum chinense Jacq. cv Habanero. *Food Chemistry, 114*(2), 553-560. [CrossRef]

[3] Batiha, G. E. S., Alqahtani, A., Ojo, O. A., Shaheen, H. M., Wasef, L., Elzeiny, M., ... & Hetta, H. F. (2020). Biological properties, bioactive constituents, and pharmacokinetics of some Capsicum spp. and capsaicinoids. *International journal of molecular sciences, 21*(15), 5179. [CrossRef]

[4] Waqas, M., Naseem, A., Humphries, U. W., Hlaing, P. T., Dechpichai, P., & Wangwongchai, A. (2025). Applications of machine learning and deep learning in agriculture: A comprehensive review. *Green Technologies and Sustainability*, 100199. [CrossRef]

[5] Botero-Valencia, J., García-Pineda, V., Valencia-Arias, A., Valencia, J., Reyes-Vera, E., Mejia-Herrera, M., & Hernández-García, R. (2025). Machine learning in sustainable agriculture: systematic review and research perspectives. *Agriculture, 15*(4), 377. [CrossRef]

[6] Ramirez-Meraz, M., Méndez-Aguilar, R., Hidalgo-Martinez, D., Villa-Ruano, N., Zepeda-Vallejo, L. G., Vallejo-Contreras, F., ... & Becerra-Martínez, E. (2020). Experimental races of Capsicum annuum cv. jalapeno: Chemical characterization and classification by 1H NMR/machine learning. *Food research international, 138*, 109763. [CrossRef]

[7] Durmuş, Y., & Atasoy, A. F. (2023). Application of multivariate machine learning methods to investigate organic compound content of different pepper spices. *Food Bioscience, 51*, 102216. [CrossRef]

[8] Hafsah, S., Surya, M. I., & Syukur, M. (2024). Classification of IPB variety of cayenne pepper genotypes using physical characteristics during the

growing period until harvest using machine learning. *Future Foods, 10*, 100500. [CrossRef]

[9] Meena, D., Chakraborty, S., & Mitra, J. (2024). Geographical origin identification of red chili powder using NIR spectroscopy combined with SIMCA and machine learning algorithms. *Food Analytical Methods, 17*(7), 1005-1023. [CrossRef]

[10] Abubeker, K. M., Akhil, S., Kumar, V. A., & Jose, B. K. (2024). Computer Vision-Assisted Real-Time Bird Eye Chili Classification Using YOLO V5 Framework. *Journal of Artificial Intelligence and Technology, 4*(3), 265-271. [CrossRef]

[11] Houetohossou, S. C. A., Hounmenou, C. G., Houndji, V. R., & Glèlè Kakaï, R. (2024, July). Empirical Performance of Deep Learning Models with Class Imbalance for Crop Disease Classification. In *International Conference on Deep Learning Theory and Applications* (pp. 118-135). Cham: Springer Nature Switzerland. [CrossRef]

[12] Djoulde, K., Ousman, B., Hamadjam, A., Bitjoka, L., & Tchiegang, C. (2024). Classification of pepper seeds by machine learning using color filter array images. *Journal of Imaging, 10*(2), 41. [CrossRef]

[13] Jeong, S., Kim, Y. K., Hur, S. H., Bang, H., Kim, H., & Chung, H. (2024). Explainable extreme gradient boosting as a machine learning tool for discrimination of the geographical origin of chili peppers using laser ablation-inductively coupled plasma mass spectrometry, X-ray fluorescence, and near-infrared spectroscopy. *Journal of Agriculture and Food Research, 18*, 101446. [CrossRef]

[14] Karadağ, K., Tenekeci, M. E., Taşaltın, R., & Bilgili, A. (2020). Detection of pepper fusarium disease using machine learning algorithms based on spectral reflectance. *Sustainable Computing: Informatics and Systems, 28*, 100299. [CrossRef]

[15] Bhagat, M., Kumar, D., & Kumar, S. (2023). Bell pepper leaf disease classification with LBP and VGG-16 based fused features and RF classifier. *International journal of information technology, 15*(1), 465-475. [CrossRef]

[16] Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. *Soft computing and industry: Recent applications*, 25-42. [CrossRef]

[17] Thul, S. T., Lal, R. K., Shasany, A. K., Darokar, M. P., Gupta, A. K., Gupta, M. M., ... & Khanuja, S. P. S. (2009). Estimation of phenotypic divergence in a collection of Capsicum species for yield-related traits. *Euphytica, 168*(2), 189-196. [CrossRef]

[18] Ribeiro, C. S., Soares, R. S., de Carvalho, S. I., Nass, L. L., Lopes, C. A., Lima, M. F., ... & Reifschneider, F. J. (2024). BRS Araçari e BRS Biguatinga: Novas cultivares de pimenta habanero da Embrapa Hortaliças. *Horticultura Brasileira, 42*, e280540. [CrossRef]

[19] Carrizo García, C., Barfuss, M. H., Sehr, E. M., Barboza, G. E., Samuel, R., Moscone, E. A., & Ehrendorfer, F. (2016). Phylogenetic relationships, diversification and expansion of chili peppers (Capsicum, Solanaceae). *Annals of botany, 118*(1), 35-51. [CrossRef]

[20] Sosa-Herrera, J. A., Alvarez-Jarquin, N., Cid-Garcia, N. M., López-Araujo, D. J., & Vallejo-Pérez, M. R. (2022). Automated health estimation of capsicum annuum L. crops by means of deep learning and RGB aerial images. *Remote Sensing, 14*(19), 4943. [CrossRef]

[21] Cruz, R. P. D., Federizzi, L. C., & Milach, S. C. K. (1998). A apomixia no melhoramento de plantas. *Ciência rural, 28*, 155-161. [CrossRef]

[22] Ren, R., Zhang, S., Sun, H., & Gao, T. (2021). Research on pepper external quality detection based on transfer learning integrated with convolutional neural network. *Sensors, 21*(16), 5305. [CrossRef]

[23] Brzozowski, L., & Mazourek, M. (2018). A sustainable agricultural future relies on the transition to organic agroecological pest management. *Sustainability, 10*(6), 2023. [CrossRef]

[24] Gaudêncio, J. H. D., de Almeida, F. A., Turrioni, J. B., da Costa Quinino, R., Balestrassi, P. P., & de Paiva, A. P. (2019). A multiobjective optimization model for machining quality in the AISI 12L14 steel turning process using fuzzy multivariate mean square error. *Precision Engineering, 56*, 303-320. [CrossRef]

[25] Teodoro, L. P. R., Silva, M. O., dos Santos, R. G., de Alcântara, J. F., Coradi, P. C., Biduski, B., ... & Teodoro, P. E. (2024). Machine learning for classification of soybean populations for industrial technological variables based on agronomic traits. *Euphytica, 220*(3), 40. [CrossRef]

[26] Shahbeig, H., & Nosrati, M. (2020). Pyrolysis of biological wastes for bioenergy production: Thermo-kinetic studies with machine-learning method and Py-GC/MS analysis. *Fuel, 269*, 117238. [CrossRef]

[27] SP, S. P., Swaminathan, G., & Joshi, V. V. (2020). Energy conservation–A novel approach of co-combustion of paint sludge and Australian lignite by principal component analysis, response surface methodology and artificial neural network modeling. *Environmental Technology & Innovation, 20*, 101061. [CrossRef]

[28] Xin, X., Pang, S., Mercader, F. M., & Torr, K. M. (2019). The effect of biomass pretreatment on catalytic pyrolysis products of pine wood by Py-GC/MS and principal component analysis. *Journal of Analytical and Applied Pyrolysis, 138*, 145-153. [CrossRef]

[29] Alqahtani, S., & Echekki, T. (2021). A data-based hybrid model for complex fuel chemistry acceleration at high temperatures. *Combustion and Flame, 223*, 142-152. [CrossRef]

[30] de Freitas Gomes, J. H., Salgado Junior, A. R., de Paiva, A. P., Ferreira, J. R., da Costa, S. C., & Balestrassi, P. P. (2012). Global Criterion Method Based on Principal Components to the Optimization of Manufacturing Processes with Multiple Responses.

*Journal of Mechanical Engineering/Strojniški Vestnik, 58*(5). [CrossRef]

[31] Naves, F. L., de Paula, T. I., Balestrassi, P. P., Braga, W. L. M., Sawhney, R. S., & de Paiva, A. P. (2017). Multivariate normal boundary intersection based on rotated factor scores: a multiobjective optimization method for methyl orange treatment. *Journal of Cleaner Production, 143*, 413-439. [CrossRef]

[32] Asimakopoulos, D. N., Bougiatioti, A., Maggos, T., Vasilakos, C., & Mihalopoulos, N. (2014). Assessment of PM2. 5 and PM1 chemical profile in a multiple-impacted Mediterranean urban area: Origin, sources and meteorological dependence. *Science of the Total Environment, 479*, 210-220. [CrossRef]

[33] Liu, M., Fan, D., Bi, N., Sun, X., & Yang, Z. (2019). Impact of water-sediment regulation on the transport of heavy metals from the Yellow River to the sea in 2015. *Science of The Total Environment, 658*, 268-279. [CrossRef]

[34] Townsend, J., Evans, B., & Tudor, T. (2016). Aerodynamic optimisation of the rear wheel fairing of the land speed record vehicle BLOODHOUND SSC. *The Aeronautical Journal, 120*(1228), 930-955. [CrossRef]

[35] Pradhan, B., & Lee, S. (2010). Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environmental Modelling & Software, 25*(6), 747-759. [CrossRef]

[36] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

[37] Ma, Y., Hou, Y., Liu, Y., & Xue, Y. (2016, March). Research of food safety risk assessment methods based on big data. In *2016 IEEE International Conference on Big Data Analysis (ICBDA)* (pp. 1-5). IEEE. [CrossRef]

[38] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications, 13*(4), 18-28. [CrossRef]

[39] Xu, X., Xiao, C., Dong, Y., Zhan, L., Bi, R., Song, M., ... & Xiong, Z. (2024). Machine learning algorithms realized soil stoichiometry prediction and its driver identification in intensive agroecosystems across a north-south transect of eastern China. *Science of the Total Environment, 906*, 167488. [CrossRef]

[40] Papandrea, P. J., Frigieri, E. P., Maia, P. R., Oliveira, L. G., & Paiva, A. P. (2020). Surface roughness diagnosis in hard turning using acoustic signals and support vector machine: A PCA-based approach. *Applied Acoustics, 159*, 107102. [CrossRef]

[41] Çetin, N., Ozaktan, H., Uzun, S., Uzun, O., & Ciftci, C. Y. (2023). Machine learning based mass prediction and discrimination of chickpea (Cicer arietinum L.) cultivars. *Euphytica, 219*(1), 20. [CrossRef]

[42] Sappl, J., Harders, M., & Rauch, W. (2023). Machine learning for quantile regression of biogas production rates in anaerobic digesters. *Science of The Total Environment, 872*, 161923. [CrossRef]

[43] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine, 4*(11), 218. [CrossRef]

[44] Martinez-Sanchez, L., See, L., Yordanov, M., Juan, P. D., Lesiv, M., & McCallum, I. (2024). Automatic classification of land cover from LUCAS in-situ landscape photos using semantic segmentation and a Random Forest model. *Environmental Modelling & Software, 172*, 105931. [CrossRef]

[45] Chen, J., Zhu, S., Wang, P., Zhang, Y., Liu, Y., & Li, W. (2024). Predicting particulate matter, nitrogen dioxide, and ozone across Great Britain with high spatiotemporal resolution based on random forest models. *Science of The Total Environment, 926*, 171831. [CrossRef]

[46] Davenport, G., Ellis, N., Ambrose, M., & Dicks, J. (2004). Using bioinformatics to analyse germplasm collections. *Euphytica, 137*(1), 39-54. [CrossRef]

[47] Alawee, W. H., Al-Haddad, L. A., Basem, A., Jasim, D. J., Majdi, H. S., & Sultan, A. J. (2024). Forecasting sustainable water production in convex tubular solar stills using gradient boosting analysis. *Desalination and Water Treatment, 318*, 100344. [CrossRef]

[48] Lee, H. P., Li, Y., Song, L., Wu, D., & Lu, N. (2024). An iterative bidirectional gradient boosting approach for CVR baseline estimation. *Applied Energy, 369*, 123456. [CrossRef]

[49] Lephalala, M., Vives, S. S., & Bisetty, K. (2024). Chaotic neural network algorithm with competitive learning integrated with partial Least Square models for the prediction of the toxicity of fragrances in sanitizers and disinfectants. *Science of The Total Environment, 942*, 173754. [CrossRef]

[50] De Meester, J., & Willems, P. (2024). Assessing the power of non-parametric data-driven approaches to analyse the impact of drought measures. *Environmental Modelling & Software, 172*, 105923. [CrossRef]

[51] Schleier, J. J., Peterson, R. K. D., Irvine, K. M., Marshall, L. M., & Preftakes, C. J. (2012). Environmental fate model for ultra-low-volume insecticide applications used for adult mosquito management. *Science of The Total Environment, 438*, 72-79. [CrossRef]

[52] Vanacore, A., Pellegrino, M. S., & Ciardiello, A. (2024). Fair evaluation of classifier predictive performance based on binary confusion matrix. *Computational Statistics, 39*(1), 363-383. [CrossRef]

[53] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. [CrossRef]

[54] Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

[55] Vasconcelos, G. A. V. B., Francisco, M. B., da Costa, L. R. A., Ribeiro Junior, R. F., & Melo, M. D. L. N. M. (2024). Prediction of surface roughness in duplex stainless steel face milling using artificial neural network. *The International Journal of Advanced Manufacturing Technology, 133*(5), 2031-2048. [CrossRef]

[56] International Plant Genetic Resources Institute, Asian Vegetable Research, Development Center, & Centro Agronómico Tropical de Investigación y Enseñanza. (1995). *Descriptors for Capsicum (Capsicum spp.)*. Bioversity International.

[57] Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200. [CrossRef]

[58] Zhang, F., Yin, J., Wu, N., Hu, X., Sun, S., & Wang, Y. (2024). A dual-path model merging CNN and RNN with attention mechanism for crop classification. *European Journal of Agronomy, 159*, 127273. [CrossRef]

[59] Zhao, G., Zhao, Q., Webber, H., Hoffmann, H., Junker, L. V., Rezaei, E. E., ... & Ewert, F. (2024). Integrating machine learning and change detection for enhanced crop disease forecasting in rice farming: A multi-regional study. *European Journal of Agronomy, 160*, 127317. [CrossRef]

[60] Li, Y., Feng, Q., Liu, C., Wang, Y., & Li, J. (2023). MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting. *European Journal of Agronomy, 146*, 126812. [CrossRef]

**Matheus Costa Pereira** PhD student in Industrial Engineering at the Federal University of Itajubá. Master's degree in Industrial Engineering from the Federal University of Itajubá (2025). Postgraduate degree in Data Analysis from Descomplica Digital College (2023). MBA in Business Intelligence from Descomplica Digital College (2023). Postgraduate degree in Occupational Safety Engineering from Descomplica Digital College (2023). Bachelor's degree in Mathematics from Cesumar University (2013). Bachelor's degree in Mechanical Engineering from Centro Universitário Una (2022). Exchange program in Business Systems Engineering at Universidad Científica del Sur, Peru (2021). Exchange program in Industrial Civil Engineering at Universidad Finis Terrae, Chile (2020). (Email: matheusc_pereira@hotmail.com)

**Tiago Martins de Azevedo** Graduated in Mechanical Engineering from the Federal University of Itajubá (2014). Completed an internship in energy utilization and has experience in fluid machinery, with emphasis on Computational Fluid Dynamics (CFD), focusing on hydraulic machines and machine design. Holds a Master's degree in Mechanical Engineering in the area of Thermal, Fluids, and Flow Machines from UNIFEI (2020), with research centered on the design, manufacturing, numerical, and experimental analysis of axial hydraulic propeller-type machines. Currently pursuing a PhD student in Industrial Engineering at UNIFEI's Institute of Industrial Engineering, and is a researcher at the National Reference Center for Small Hydroelectric Plants (CERPCH). (Email: tiago.deazevedo@yahoo.com.br)

**Caio Tertuliano Ribeiro** Graduated in Industrial Engineering from the Federal University of Itajubá (2023). Has experience in production management, operational excellence, transactional database modeling, and development of production control systems using VBA. Also skilled in predictive modeling and exploratory data analysis within data science and AI projects. Strong interest in machine learning and artificial intelligence with engineering applications. Currently pursuing a Master's degree in Industrial Engineering at UNIFEI with the research topic "Application of Design of Experiments (DOE) in Hyperparameter Optimization for Clustering Models," focusing on robust statistical techniques for refining machine learning models to achieve state-of-the-art results. (Email: caio.tertu@hotmail.com)

**Ana Izabella Freire** Holds a Bachelor's degree in Agronomy from the Federal University of Lavras – UFLA (2013). Master's degree in Genetics and Plant Breeding from UFLA (2015). PhD in Crop Science from the Federal University of Viçosa – UFV (2019). Faculty member at the Federal Institute of Southern Minas Gerais. Holds an MBA in Logistics and Supply Chain from UNINTER. Lecturer at the Assistencial Course and Intelligence and Culture Center – CACIC. Completed a postdoctoral fellowship at the Federal University of Lavras (UFLA) and is currently pursuing a second postdoctoral position at the Federal University of Itajubá (UNIFEI). (Email: anaizabellinha2014@gmail.com)

**Matheus Brendon Francisco** Bachelor's, Master's, and PhD in Mechanical Engineering from the Federal University of Itajubá (UNIFEI), with a focus on Aerospace Mechanical Engineering. Currently a professor at UNIFEI's Institute of Industrial and Management Engineering, teaching undergraduate and graduate courses. Member of the Computational Mechanics and Optimization Research Group (GEMEC) and the Manufacturing Optimization and Innovation Technology Center (NOMATI). Coordinator of FabLab (UNIFEI's Innovation Lab within its Entrepreneurship Center). Research interests include optimization methods, applied artificial intelligence, multivariate data analysis, design and analysis of experiments, additive manufacturing, and metaheuristics. (Email: matheus_brendon@unifei.edu.br)

**João Luiz Junho Pereira** Adjunct Professor at the Institute of Industrial and Management Engineering at the Federal University of Itajubá (UNIFEI), where he earned his Bachelor's, Master's, and PhD in Mechanical Engineering, focusing on Design, Materials, and Processes. He is a collaborating researcher at UNIFEI's GEMEC group and at the Computing Science Division of the Aeronautics Institute of Technology (ITA), where he completed a postdoctoral fellowship in Artificial Intelligence. He also held a postdoctoral position at the ARC Training Centre for Optimisation Technologies (OPTIMA) at the University of Melbourne (UNIMELB). Winner of the 2023 CAPES Thesis Award in Engineering III and recipient of an Honorable Mention for the ABCM-EMBRAER Award (2022). (Email: joaoluizjp@unifei.edu.br)

**Anderson Paulo de Paiva** Full Professor at the Federal University of Itajubá (UNIFEI/IEPG). CNPq Research Productivity Fellow. Mechanical Engineer with a Master's degree in Industrial Engineering (UNIFEI) and a PhD in Mechanical Engineering (UNIFEI). Works in the areas of Optimization Methods, Design and Analysis of Experiments (DOE), Multivariate Statistical Analysis, and Machine Learning. Main research line: Optimization of Manufacturing Processes. Coordinator of NOMATI/IEPG/UNIFEI (Manufacturing Optimization and Innovation Technology Center). (Email: andersonppaiva@unifei.edu.br)