ISSN 2007-9737

Machine Learning-Based Classification of Habanero Pepper Yield using Mixed Metabolomic and Phenotypic Profile Features

Gerardo Acevedo-Sánchez[1], Jorge Pacheco-Senard[1], Moisés Ramírez-Meraz[2],
Reinaldo Méndez-Aguilar[2], Elvia Becerra-Martínez[3], Antonio Alarcón-Paredes[1],
Cornelio Yáñez-Márquez[1,*]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

[2] INIFAP-Campo Experimental Las Huastecas,
México

[3] Instituto Politécnico Nacional,
Centro de Nanociencias y Micro y Nanotecnologías,
México

{gacevedos2024, aalarcon, jpachecos2024, cyanez}@cic.ipn.mx,
{ramirez.moises, mendez.reinaldo}@inifap.gob.mx, elmartinezb@ipn.mx

Abstract. This study addresses the intelligent classification of habanero pepper yield through a machine learning model based on the algorithmic pairing of IBk (Instance-Based *k*) and the HEOM (Heterogeneous Euclidean-Overlap Metric), designed to handle mixed-type data, integrating both numerical (metabolomic and morphological) and categorical (phenotypic) features. The dataset included 165 instances associated with 58 features, combining 51 metabolites (sugars, amino acids, organic acids, bioactive compounds), four qualitative descriptors (race, cultivar, color, description), and three quantitative descriptors (fruit size). The target variable was binary, defining high yield (>25 tons/ha) and low yield (<14 tons/ha) exhibiting a moderate class imbalance (*IR* = 1.75). Leave-One-Out Cross-Validation (LOOCV) was employed to ensure a robust and deterministic validation process. The IBk/HEOM algorithm achieved perfect classification (100% accuracy) with 58 features for $k \leq 25$, demonstrating the high discriminatory power of the selected biomarkers. Starting from $k = 26$, a progressive increase in False Positives (Type I errors) was observed, which is typically associated with decision boundary overlap and bias towards the majority class. Feature relevance analysis identified eight critical attributes (race, cultivar, fruit width, succinic acid, ferulic acid, ascorbic acid, guanosine, and NAD) that, by themselves, maintained optimal predictive performance up to $k = 31$, providing a direct path for parsimonious model optimization and a reduction in field and laboratory costs. This work validates the utility of integrating mixed data from metabolomic biomarkers and phenotypic features. The robust HEOM-based framework natively handles data heterogeneity, eliminating the need for pre-processing transformations. This offers an inherently interpretable predictive tool ideal for decision-making in agricultural and biochemical research.

Keywords. Mixed data, yield prediction, metabolomic biomarkers, morphological-productive profiling.

1 Introduction

Modern agriculture faces the persistent challenge of enhancing crop productivity and resilience under variable agroecological conditions. In this context, establishing models, methods, or processes for yield prediction has become a priority for making effective and efficient agricultural decisions [1,2].

Traditionally, agricultural predictive models for estimating yield, or other features, have been based on homogeneous numerical data series such as climate (i.e., temperature, precipitation), vegetation indices (e.g., NDVI) obtained through

doi: 10.13053/CyS-29-4-6114

remote sensing techniques, edaphic characteristics, or simple quantitative agronomic features [3,5]. These approaches are limited to the exclusive use of this lower-complexity dataset structure, containing only numerical data, which can omit qualitative components involved in the relationships between plant genotype, environmental aspects, and productivity [4,6]. Consequently, models based solely on qualitative or categorical features lack precision, interpretability, and efficacy required for implementation in agricultural contexts with complex behavior.

The recent rise of plant biotechnology, particularly through metabolomics and high-throughput phenotyping, has generated vast amounts of information that demand more precise analytical methodologies and tools. This has spurred the exploration of more advanced approaches that integrate different data types (i.e., "multi-modal," "multi-omics," "multi-source") to improve the prediction of complex agroecological scenarios [5-6]. For instance, in winter wheat, deep learning models combining environmental, phenological, and soil data have surpassed the predictive accuracy of conventional linear models [7-9]. In maize, the combination of genetic (molecular polymorphisms) and environmental data has enabled the capture of genotype × environment interactions, resulting in higher predictive performance compared to models using these data types individually [5]. Likewise, studies on sweet peppers (*Capsicum annuum*) have employed phenotypic agronomic traits alongside convolutional neural networks (CNNs) and genomic models (GBLUP) with promising results for yield prediction [10].

Within this framework, incorporating metabolomic biomarkers alongside traditional phenotypic characteristics is a current trend, though its implementation is often unbalanced.

Recent evidence shows that metabolomic profiles can capture the internal activity of metabolic pathways, the plant's physiological status, stress responses, and the accumulation of compounds related to growth, resistance, and productive development [11]. This information, when analyzed with appropriate techniques and processes, allows for an understanding of multifactorial yield processes that are not evident in a disjointed analysis [12]. In hybrid rice, it has been demonstrated that the incorporation of metabolomic data improves yield predictability compared to the exclusive use of phenotypic or genomic features [12]. Another study on fruit flavor selection combined fruit chemistry (metabolomics) with sensory data to train ML models, showing that chemical data (e.g., metabolites, sugars, acids) provide relevant explanatory power for current challenges that require holistic and interdisciplinary solutions [13-15].

In this sense, machine learning (ML), intelligent computing (IC), and other related areas emerge as complementary technological supports to elucidate predictive relationships among multiple and multidimensional features such as genomic, molecular, and agricultural traits. However, a particular challenge in this field lies in the mixed nature of the features (qualitative and quantitative) and the limited capacity of most algorithms to process it. Most ML algorithms and other disciplines linked to AI, including classical variants (e.g., *k*-Nearest Neighbors), are designed for continuous numerical features and assume data homogeneity [16-19]. Nevertheless, when these algorithms confront mixed data processing, they require preprocessing mechanisms such as transformations or conversions (i.e., one-hot encoding, target encoding, etc.), which can distort or alter the inherent relationships within the data community by introducing potential spurious effects and, consequently, diminishing the predictive power of the algorithms [20].

Consequently, there is a critical need for algorithms specifically designed to handle this data heterogeneity or complexity. The Instance-Based *k* (IBk) algorithm, implemented alongside the Heterogeneous Euclidean-Overlap Metric (HEOM), is one of the pairings that has emerged to address this problem in agricultural domains [21]. The HEOM metric calculates the dissimilarity between two instances by combining the normalized Euclidean distance (for numerical attributes) and the overlap metric (for categorical attributes), without requiring data transformations that could introduce bias [22]. This allows each neighbor in the algorithm to contribute to the final classification proportionally to its relevant similarity or dissimilarity.

Consequently, the objective of this study is to classify habanero pepper yield using a machine learning model trained with the algorithmic pairing IBk/HEOM. The training and testing phases consider mixed features from metabolomic biomarkers (51 metabolites) and phenotypic traits (seven attributes), obtained from 11 cultivars of *Capsicum chinense*. This approach addresses the identified need to integrate different levels of biological information to improve prediction, to establish a model that does not rely solely on external numerical features, and to test algorithms capable of natively handling mixed data using standard computational resources. The purpose, beyond seeking higher predictive accuracy, is to improve the interpretability of yield-associated features, thereby providing an analytical tool for decision-making [2].

2 Materials and Methods

This section details the methodological setup for evaluating the IBk/HEOM model, covering dataset features, validation strategy, classifier selection, performance measures, computational resources, and implementation specifics (Fig. 1).

2.1. Dataset

Data Collection. The dataset used in this study was generated from previously reported research [23], in which the metabolomic profile of 11 cultivars of *Capsicum chinense* was analyzed using purely quantitative techniques. For the purposes of the present study, the original dataset from Ramírez-Meraz *et al*., containing 51 numerical attributes, was augmented with morphometric and productive features (Fig. 1).

Dimensionality. The dataset's dimensionality, or number of attributes, was 58. This consisted of a mixture of categorical and numerical features describing morphological features, primary and secondary metabolites, and other bioactive compounds from habanero pepper fruit samples.

Regarding categorical features, four descriptors related to plant origin and type were integrated: race (4 categories: Habanero, Antillano, Jolokia, and Hybrid), cultivar (11 categories: Jaguar, HRA 1-1, etc.), fruit color (5 categories: Orange, Red, Yellow, Dark brown, Orange-Yellow), and description (3 categories: OPV, IL, Com).

Concerning numerical features, these included 3 morphometric measurements (fruit weight, length, and average width) and the metabolomic profile comprising 51 biomarkers associated with metabolites extracted from *C. chinense* fruit samples. The primary metabolites were divided into sugars (fructose, galactose, glucose, mannose, sucrose), alcohols (myo-inositol), amino acids (20 compounds such as alanine, arginine, etc.), and organic acids (acetic, citric, formic, fumaric, lactic, malic, malonic, pyruvic, quinic, succinic). The secondary metabolites and bioactive compounds included phenolic acids (chlorogenic, ferulic, gallic, vanillic), hydroxylated acids (2-hydroxybutyric, 3-hydroxyisobutyric), vitamins (ascorbic acid), nucleotides (adenosine, cytidine, guanosine, uridine), and other specialized metabolites (choline, ethanolamine, NAD, NADP, O-phosphocholine, trigonelline) (Fig. 1).

Cardinality. The total number of instances (samples) in the dataset was 165. The primary distribution of instances was 15 per/variety and per/plant material type, comprising 15 from open pollination, 75 from inbred lines, and 75 from commercial hybrids (Fig. 1).

Class Label*.* The target feature for this work was yield, which framed the classification approach as a binary problem based on well-established metabolomic and productive biomarkers. Class A (high yield, >25 tons/ha), comprising n = 105 instances, was predominantly associated with orange and yellow Habanero varieties. In contrast, class B (low yield, <14 tons/ha), with 60 instances, was concentrated in dark-colored cultivars and specific inbred lines. This distribution suggests a strong association between organoleptic, genetic background, and yield potential features (Fig. 1).

Class Balance*.* The analysis of class balance using the imbalance ratio (*IR*) revealed a moderate imbalance with a value of *IR* = 1.75, calculated from the 105:60 ratio between the majority and minority classes [24]. While not critical, this level of imbalance requires consideration in the selection of *ad-hoc* performance measures and validation methods, favoring approaches such as stratified

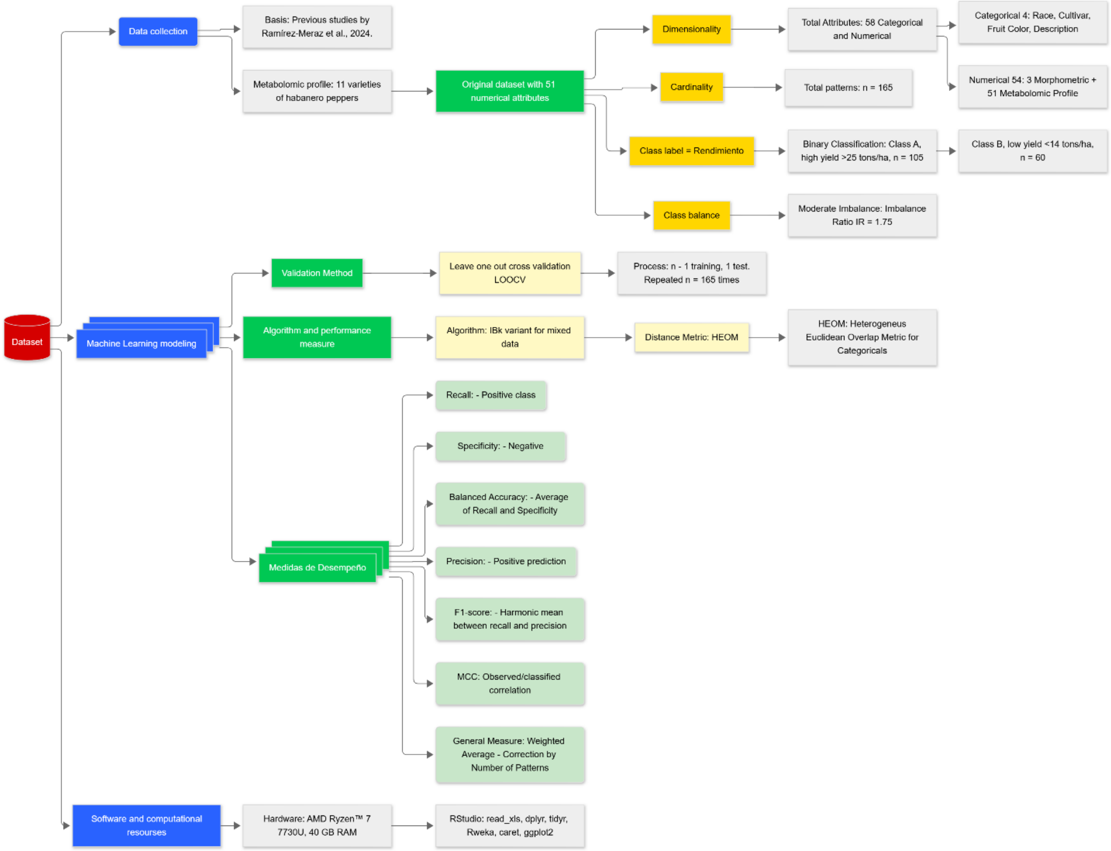
Dataset
Data collection
Basis: Previous studies by Ramírez-Meraz et al., 2024.
Metabolomic profile: 11 varieties of habanero peppers
Original dataset with 51 numerical attributes
Dimensionality
Total Attributes: 58 Categorical and Numerical
Categorical 4: Race, Cultivar, Fruit Color, Description
Numerical 54: 3 Morphometric + 51 Metabolomic Profile
Cardinality
Total patterns: n = 165
Class label = Rendimiento
Binary Classification: Class A, high yield >25 tons/ha, n = 105
Class B, low yield <14 tons/ha, n = 60
Class balance
Moderate Imbalance: Imbalance Ratio IR = 1.75
Machine Learning modeling
Validation Method
Leave one out cross validation LOOCV
Process: n - 1 training, 1 test. Repeated n = 165 times
Algorithm and performance measure
Algorithm: IBk variant for mixed data
Distance Metric: HEOM
HEOM: Heterogeneus Euclidean Overlap Metric for Categoricals
Medidas de Desempeño
Recall: - Positive class
Specificity: - Negative
Balanced Accuracy: - Average of Recall and Specificity
Precision: - Positive prediction
F1-score: - Harmonic mean between recall and precision
MCC: Observed/classified correlation
General Measure: Weighted Average - Correction by Number of Patterns
Software and computational resourses
Hardware: AMD Ryzen™ 7 7730U, 40 GB RAM
RStudio: read_xls, dplyr, tidyr, Rweka, caret, ggplot2

Fig. 1. Methodological workflow diagram showing the phases of dataset construction (*n* = 165, 58 mixed features, and class balance), the machine learning modeling with LOOCV, the IBk/HEOM algorithm, performance measures, and software and computational resources

cross-validation and measures robust to unequal distributions.

2.2. Machine Learning Algorithm Modeling

Validation Method for Training and Testing. The Leave-One-Out Cross-Validation (LOOCV) method was selected for validation, considered one of the most rigorous methods for evaluating machine learning algorithms, particularly for moderately sized datasets [25]. In this approach, each instance in the dataset is iteratively used as an individual test set, while the remaining *n* - 1 instances constitute the training set. This process is exhaustively repeated for each of the 165 available instances, ensuring that every observation is evaluated independently and that the model is trained with the maximum amount of data possible in each iteration. The main advantage of LOOCV lies in its ability to provide a nearly unbiased estimate of the generalization error, minimizing the variance associated with random data splits and offering an extremely robust evaluation of the algorithm's predictive performance.

Mixed-Data IBk Algorithm and HEOM Metric. The implemented IBk (Instance-Based k) algorithm corresponds to a specialized variant of the classic *k*-Nearest Neighbors (*k*-NN), designed specifically to handle the inherent complexity of the mixed data characterizing this dataset [22]. The uniqueness of the IBk approach lies in the use of HEOM (Heterogeneous Euclidean-Overlap Metric), a heterogeneous distance function that allows for the effective calculation of similarities between instances that simultaneously contain continuous numerical attributes and nominal categorical features [Eq. 1]. For numerical attributes, HEOM applies a normalized Euclidean distance, while for categorical features it implements the overlap metric, which assigns a distance of zero when categorical values match and one when they differ [Eq. 2 and 3]. This hybrid capability is fundamental in the context of the analyzed dataset, where quantitative metabolite measurements coexist with qualitative varietal descriptors and phenotypic features. It allows the algorithm to capture the complex similarity relationships between samples without requiring pre-processing transformations that could distort the inherent structure of the data:

$$HEOM(x,y) = \sqrt{\sum_{a=1}^{m} d_a(x_a,y_a)^2}, \tag{1}$$

$$d_a(x_a,y_a) = \begin{cases} 1, & if\ x\ or\ y\ is\ missing\ value \\ overlap_a(x_a,y_a), & if\ a\ is\ categorical\ attribute \\ rn_{diff_a}(x_a,y_a), & if\ a\ is\ numerical\ attribute \end{cases}, \tag{2}$$

$$overlap_a(x_a,y_a) = \begin{cases} 0, if\ x = y \\ 1, other\ case \end{cases}, rn_{diff_a}(x_a,y_a) = \frac{|x-y|}{range_a}, \tag{3}$$

2.3. Performance Measures

For the analysis of the IBk algorithm's performance, evaluation measures derived from a binary confusion matrix were selected, given the specific classification problem addressed in this study [26-27].

Recall. It focuses on the positive class defined as the ratio of cases correctly classified as positives and those misclassified as negatives [eq. 4]:

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

Specificity. It focuses on the negative class as it is defined as the ratio of cases correctly classified as negatives and those misclassified as positives [eq. 5]:

$$Specificity = \frac{TN}{TN + FP}, \tag{5}$$

Balanced Accuracy. Sensitivity and specificity must be calculated beforehand. This measure is suitable for imbalanced datasets since it minimizes the effect of the majority class through the implementation of an average of 'accuracy by class' [eq. 6]:

$$BA = \frac{Sensitivity + Specificity}{2}, \tag{6}$$

Precision. It is a variation of sensitivity because it focuses on the positive class, as it is defined as the ratio of TP with respect to all instances classified as positive, including the false positives [eq. 7]:

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

F1-Score. It is calculated as the harmonic average of precision and recall. The main limitation is that precision and recall must be greater than 0 for it to be calculated [eq. 8]:

$$F1 - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}. \tag{8}$$

MCC. It determines the correlation between observed and predicted by the algorithm. It is like accuracy in that it excludes the effects of specific matches and misses [eq. 9]:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}, \tag{9}$$

Cohen's Kappa. Measures the agreement between the observed classifications and those predicted, correcting for the agreement that would be expected by chance alone. Unlike performance, this measure is robust to imbalances in class distributions [eq. 10]:

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)}, \tag{10}$$

where:

$$p_o = \frac{TP + TN}{Total},$$

$$p_e = \frac{(TP+FN)(TP+FP) + (FP+TN)(FN+TN)}{Total^2}. \tag{11}$$

Overall Measures. The weighted-average (*W*) was selected since it corrects the average by the number of instances (w_i) per class c_k which is ideal for unbalanced cases [eq. 12]:

$$W = \sum_{i=1}^{n} w_i \cdot MD_{ck}, \tag{12}$$

where $w_i = \frac{Samples\ in\ c_k}{Total\ samples}$.

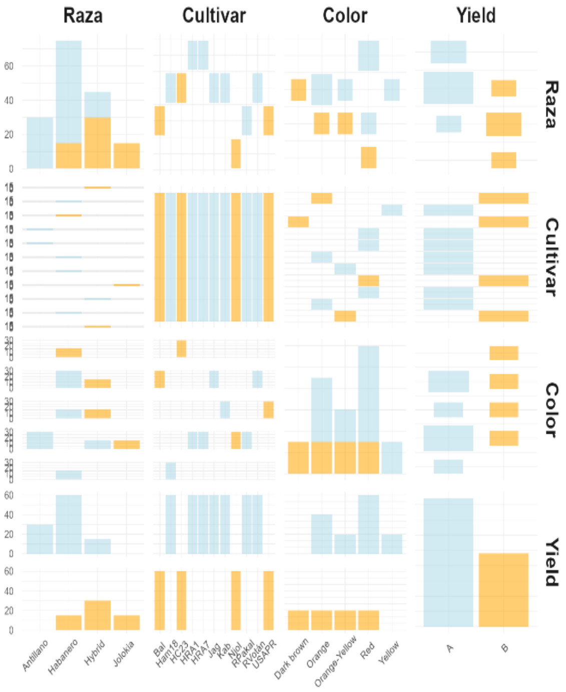

Fig. 2. Comparative profile of four productive categories of habanero pepper plants associated with two yield levels (High = blue and Low = orange)

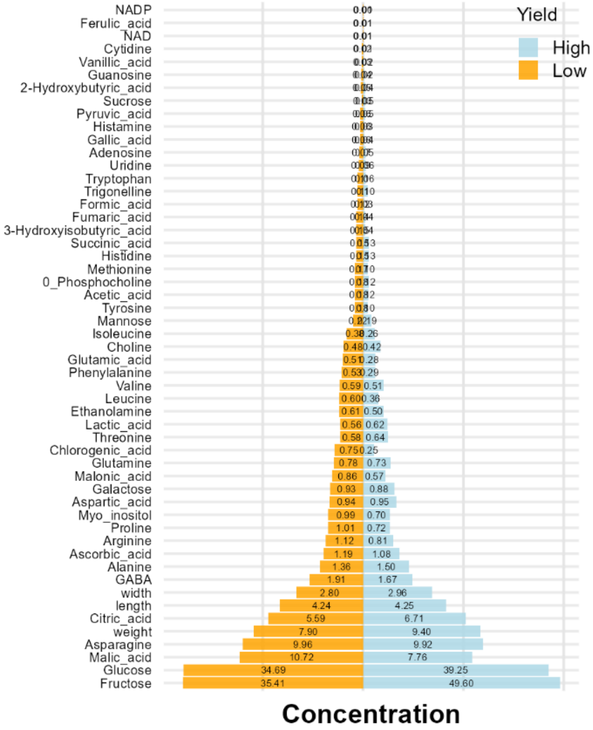

Fig. 3. Comparative profile of concentrations for 51 metabolites and three morphological features associated with two yield levels (High = blue and Low = orange) in habanero pepper samples

2.4. Features Importance

The relevance of predictor features or feature importance was quantified using a systematic permutation method. Initially, a baseline predictive performance (165 correct classifications) was established using the IBk/HEOM algorithm with LOOCV.

For each predictor feature (a total of 58), a permuted dataset was generated by randomizing its values, thereby severing its statistical relationship with the target feature.

Following each permutation, the number of correct classifications was recalculated. The relevance of each feature was then quantified as the reduction in predictive performance, that is, the decrease in the number of correct classifications between the baseline model and the model trained on the permuted data.

2.5. Analysis Software and Computational Resources

The analyses were performed using an 8-core, 16-thread AMD Ryzen™ 7 7730U processor with integrated AMD Radeon™ GPUs and 40 GB of RAM (8 GB soldered DDR4-3200 and 32 GB SO-DIMM DDR4-3200).

The analyses were performed in RStudio using the following general functions:

1) Data processing and management with read_xls, dplyr, and tidyr;

2) Training and testing of IBk model with Rweka;

3) Performance measures with caret; and

4) Graphics were generated with ggplot2 package [28].

3 Results and Discussion

3.1. Descriptive Analysis of Habanero Pepper Yield

A total of 165 instances associated with metabolomics, and productive profile of 58 mixed features evaluated in habanero pepper samples were analyzed. Regarding categorical features, the distribution of each attribute with respect to yield revealed a high discriminatory capacity in the race feature. Specifically, "Habanero" was associated with Class A (high yield), while "Jolokia" and "Hybrid" were associated with Class B (low yield). The Cultivar feature also showed a high degree of distinction, with specific cultivars (e.g., 'Jaguar', 'HRA 1-1', etc.) strongly skewed towards one of the classes. For Color, "Orange" and "Yellow" were linked to Class A, whereas "Dark brown" and "Orange-Yellow" were linked to Class B.

Furthermore, the distributions in the combinations of Race × Cultivar, Race × Color, and Cultivar × Color were not uniform, indicating a strong biological-productive dependency among these features [22,29-30]. Although multicollinearity is a concern in parametric models (e.g., logistic regression), in a distance-based algorithm like IBk/HEOM, this dependency is handled implicitly through the overlap metric [22]. The use of HEOM allows for 0 to be assigned if two instances share a category (e.g., both are "Habanero") and 1 if they differ, enabling the algorithm to capture dissimilarity through the

combination of attributes, thereby reinforcing the suitability of the selected algorithm and metric. Overall, it was confirmed that the qualitative features are not merely descriptors but act as key metabolomic and phenotypic biomarkers [23] (Fig. 2).

Regarding the numerical features, the metabolomic and phenotypic profile revealed biochemical heterogeneity associated with differences in productive yield. Metabolites with contrasting concentrations, such as fructose, glucose, malic acid, citric acid, and asparagine, were primarily associated with the high-yield class and thus represent predictors with high potential for separability in the classification process [31-33]. Conversely, features with symmetrical distributions (e.g., NADP or Ferulic Acid) demonstrated low or negligible individual capacity to distinguish between yield classes. This suggests their predictive value may depend on complex, non-linear interactions that a univariate analysis fails to capture (Fig. 3).

The presence of this discriminatory profile validates the measurement and integration of metabolomic and morphometric features into machine learning models [22]. The differential concentration of various metabolomic and morphological attributes between the yield classes suggests that the allocation of metabolic resources is a key factor in predicting yield type [34].

Furthermore, considering that the selected classification model uses IBk/HEOM, differences in magnitude (e.g., Fructose ∼35−50 vs. NADP ∼0.00) are normalized by the algorithm's criterion (Eq. 3) during preprocessing. This ensures that features with large concentration ranges do not dominate the distance calculation and the classification of the nearest neighbors (Fig. 3).

Consequently, integrating metabolomic profiles for a predictive objective, as in the present study, not only optimizes the discrimination between yield classes but also facilitates the identification of metabolic biomarkers with high explanatory value, thereby reinforcing the interpretability and applicability for decision-making.

3.2. Typification of Errors and Correct Classifications in IBk/HEOM

The results from the IBk algorithm with $k \leq 25$ showed an almost perfect separation between high (A) and low (B) yield, achieving 105 correct classifications for class A and 60 for class B. This suggests high sensitivity and specificity of the metabolomic and productive biomarkers for correctly classifying yield (Fig. 4). This perfect performance at low *k*-values is characteristic of overfitting, where the model learns the training data too closely, including its noise. In this state, the classifier memorizes local samples and performs without bias on data with high instance density, but its ability to generalize may decrease in the presence of noise or unseen data [35]. However, this potential effect was mitigated by using a robust validation method like LOOCV,

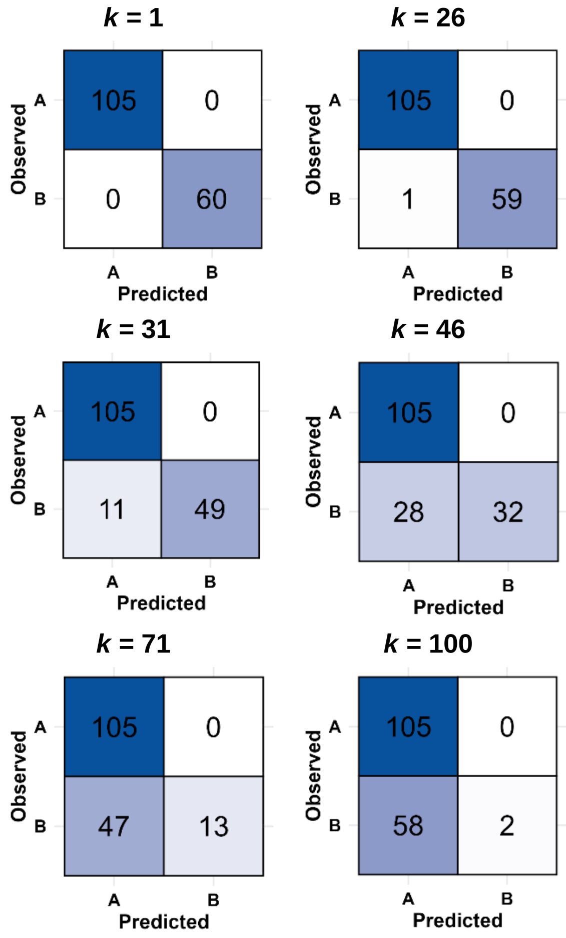

Fig. 4. Confusion matrices obtained by applying the IBk (Instance-Based *k*) classifier with *k* values of 1, 26, 31, 46, 71, and 100

which is more resilient due to its deterministic nature [25, 36].

As the value of *k* increased, the confusion matrices showed a progressive rise in Type I errors (false positives or FP). In the context of this study, these represent habanero chili fruit samples that are incorrectly classified as high yield (class A) when they belong to the low yield class (B). Specifically, this type of error was first observed at *k* = 26, with the first FP, and reached 96.7% (*n* = 58) Type I errors at *k* = 100. This instance is typical of the bias towards the majority class (Class A, *n* = 105), as with large *k*-values, the neighborhood decision is subjected to a vote that includes points from both classes. This dilutes the decision boundary and subjects the classifier to a bias towards class A. This implies that although the classifier becomes more stable, it loses effectiveness in distinguishing between proximate or overlapping classes [16, 37].

Therefore, the above evidence demonstrates that excessively large *k*-values nullified the classification capacity of the IBk algorithm, resulting in an almost constant classifier. Interestingly, Class A (high yield) was never compromised, regardless of the k-value, suggesting that the metabolomic and productive profile of the samples provides a perfect separability boundary for high yield. However, despite having a considerable margin to handle overlap (up to 25 neighbors), the probability of incurring a Type I prediction error increased significantly for *k* > 26 (Fig. 4).

3.3. IBk Performance Evaluation

The comprehensive analysis of the IBk classifier's performance, evaluated based on the increasing *k*-neighbors parameter with LOOCV, reliably revealed the fundamental dynamics of the bias-variance trade-off within the specific context of habanero pepper yield prediction. The mixed data (quantitative and categorical) extracted from the evaluation of the 165 instances generally provided solid support.

For low *k*-values (from 1 to 25), the model exhibited perfect performance, with 165 correct classifications out of the total 165 instances (value = 1) across all performance measures (Recall, Specificity, Precision, F1-score, Balanced Accuracy, MCC, and Kappa). This validates the previous results regarding the total discriminatory capacity of the employed algorithmic pair (Fig. 4 and 5). In the context of LOOCV as a validation method, this sustained perfection for high yield indicated the presence of a set of highly discriminatory metabolomic, morphological, and productive biomarkers in terms of their association with yield [38]. As observed in the exploratory analysis, the effect of critical levels of sugars like fructose and sucrose, amino acids like arginine, organic acids like malic and citric, or bioactive compounds like ascorbic acid and chlorogenic acid, in combination with genetic descriptors like race and cultivar, contributed infallibly to capturing

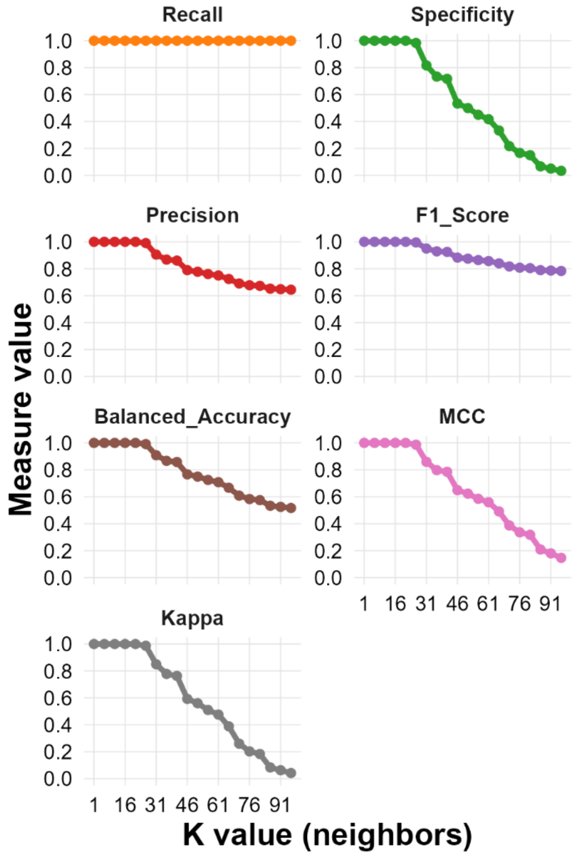

Fig. 5. Performance measures obtained by applying the IBk (Instance-Based *k*) classifier with *k* values ranging from 1 to 100. The behavior of Recall, Specificity, Precision, F1-score, Balanced Accuracy, MCC, and Kappa are shown

the heterogeneity of the profile in Class A [39-40] (Fig. 5).

During the comparative analysis, the critical point began at *k* = 26, where the first error was recorded, and Specificity began its drastic decline from 0.983 to 0.03 at *k* = 96. This drop, which signifies an increase in Type I error (FP in the confusion matrix), accelerated notably from *k* = 31.

At this neighborhood level, the number of correct classifications dropped to 154 (93.3%), consequently affecting Specificity (0.817), Precision (0.905), and other metrics, except Recall (1.0). The performance evaluation using other robust metrics like MCC (0.860) and Kappa (0.850) confirmed a significant loss in overall classification capability [41-42]. Broadly speaking, this behavior quantified the moment when the smoothing of the decision boundary, by relying on a wider neighborhood, began to systematically favor the profile of the majority class or the more consistently identifiable class (high yield), thereby diluting the more complex instances of low yield (Fig. 5).

Overall, the performance degradation continued progressively and monotonically. For *k* = 51, the number of correct classifications was 135 (81.8%), and Specificity fell notably (0.5), affecting Balanced Accuracy (0.75) and Precision (0.778). At this *k*-value, the MCC (0.624) and Kappa (0.560) reflected a moderate correlation between predicted and observed values. At the extreme end of the iterations, with *k* = 96, despite a perfect Recall, the model achieved only 107 correct classifications (65.5%), with a Specificity of just 0.03, meaning it correctly identified only 3% of the instances associated with low yield. The Precision was 0.644, indicating that at least one out of every three cases predicted as high yield was incorrect.

The MCC (0.180) and Kappa (0.063) confirmed this trend, indicating predictive power barely better than random chance. (Balanced Accuracy = 0.516) (Fig. 5).

In general, in the IBk implementation on the metabolomic, morphological, and productive profile dataset associated with habanero pepper yield, the transition from a model that faithfully captures yield variability (*k* < 26) to one with high bias (*k* > 31) is unambiguously quantified. The results demonstrate that the most effective neighborhood range, which maximizes the balance between classifying high yield (Recall = 1.0) and identifying low yield (Specificity > 0.98), ideal for an imbalance ratio of *IR* = 1.75, is found within *k* = 1 to 21. However, considering the principle of parsimony and the pursuit of a more generalizable model, the value of *k* = 26 (with 164 correct classifications, 1 error, Balanced Accuracy = 0.99, MCC = 0.987, and Kappa = 0.987) represents the margin with the greatest capacity to handle class overlap.

Furthermore, the quantitative data validates that for *k* > 26, the increase in bias leads to a progressive and measurable underfitting, where the model begins to make systematic Type I errors by failing to capture the heterogeneity of low yield [16,43]. This effect appears to be associated with complex combinations in metabolomic profile, morphological traits, and productive history.

Therefore, calibrating the k parameter within the identified range (1 - 26) is crucial for developing an intelligent classification model that is both accurate and robust, capable of leveraging the heterogeneity of the integrated metabolomic and productive profile for agricultural selection and decision-making [44].

3.4 Critical Features in Classification

The analysis of the optimal *k*-value for the IBk/HEOM algorithm's performance showed that *k* ≤ 25 maintains perfect predictive levels across all analyzed performance measures. Consequently, *k* = 26 marked the onset of predictive inefficacy (1 FP), attributable to the complexity of the metabolomic, morphological, and productive profile associated with indiscernible instances resulting from class overlap and the dataset's imbalanced condition (*IR* = 1.75).

In order to identify the critical features for classifying habanero pepper yield, a relevance analysis was conducted by sequentially eliminating or permuting each attribute at *k* = 26 to measure the reduction in predictive capacity. This process revealed that the first FP was associated with the interaction of eight features identified as the most influential or discriminative predictors: race, cultivar, fruit width, succinic acid, ferulic acid, ascorbic acid, guanosine, and NAD [45]. The remaining 50 features reduced predictive capacity by less than 0.06 (Fig. 6).

Validation of the IBk/HEOM algorithm trained with the most influential features (8 features with a predictive reduction value > 0.01, Fig. 6) demonstrated that this minimal subset effectively captured the heterogeneity of habanero pepper yield. In contrast, the model trained and validated with the 50 apparently less relevant features showed higher error (7 FPs), performing worse even than the original model with all 58 features, which had only 1 FP (Fig. 4 and 6).

The performance measures for the 8-feature model exhibited a more stochastic effect from *k* > 31 onwards. This was particularly evident in the Recall metric, which, in the *k*-range of 31 to 56, showed a reduction in the correct prediction of instances from *both* yield classes without a defined bias towards the majority class (high yield), unlike the model trained with the full feature set.

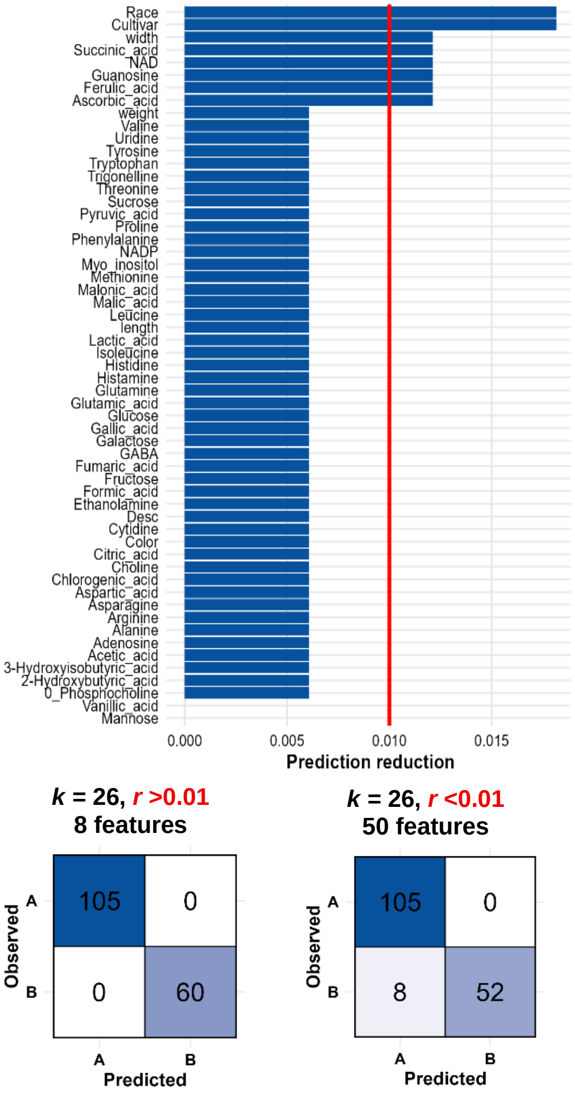

Fig. 6. Predictive reduction (r) of the IBk/HEOM model upon sequential elimination or permutation of features at *k* = 26. Prediction reduction quantifies the decrease in the number of correct classifications when each feature is individually excluded. The confusion matrices show a comparison of IBk/HEOM performance with 8 versus 50 features

Notably, this loss in classification efficacy reversed from *k* = 61 onwards, with the model once again achieving near-perfect classification (0.99) (Fig. 7).

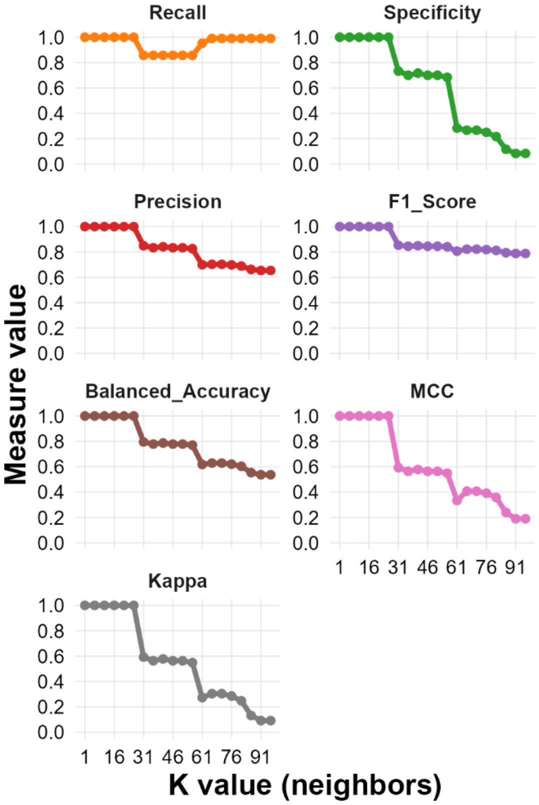

Fig. 7. Performance measures obtained by applying the IBk (Instance-Based *k*) classifier using 8 high-relevance features and *k*-values ranging from 1 to 100. The behavior of Recall, Specificity, Precision, F1-score, Balanced Accuracy, MCC, and Kappa are shown

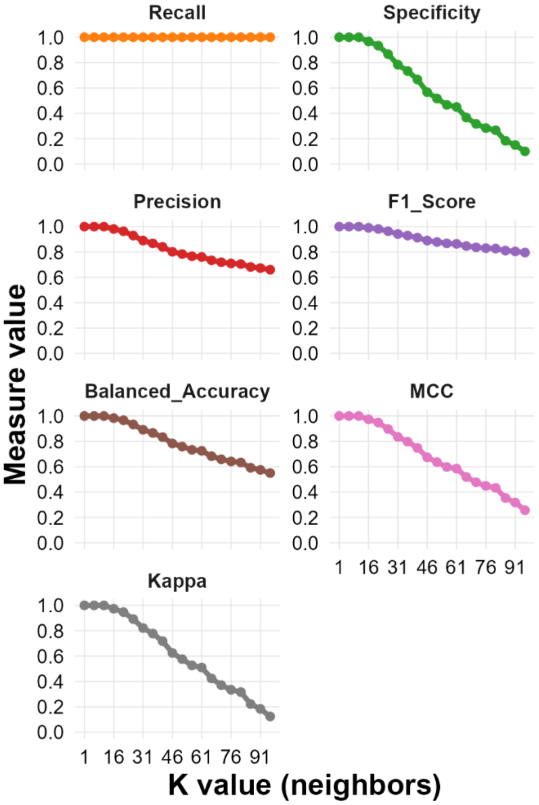

Fig. 8. Performance measures obtained by applying the IBk (Instance-Based *k*) classifier using 50 moderate-to-low relevance features and k values ranging from 1 to 100. The behavior of Recall, Specificity, Precision, F1-score, Balanced Accuracy, MCC, and Kappa are shown

The remaining performance measures showed significantly more pronounced decrements at intervals of approximately 30 neighbors.

However, in the margin for handling class overlap, this model (with 8 features) maintained perfect performance until *k* = 31, increasing its robust classification capacity by six neighbors compared to the original model trained with 58 features, and by twenty neighbors compared to the model using the 50 low-relevance features (Fig. 7).

In contrast, the model with 50 features was noticeably affected by class overlap, beginning a progressive decline in predictive capability from *k* > 11, where the effectiveness for discerning instances of both classes decreased markedly.

This finding validates the hypothesis regarding the complexity of the metabolomic, morphological, and productive biomarker profile for discerning between high and low yield (Fig. 8).

Therefore, it was evident that the IBk/HEOM model is highly effective even in a scenario of feature absence.

The 8 most relevant features can simplify the interpretability of habanero pepper yield without introducing significant bias, at least up to a neighborhood of *k* = 31.

However, even with the simultaneous absence of these critical features, the model responds effectively by managing class overlap, although within a smaller neighborhood range (*k* < 11).

4 Limitations and Future Work

Despite the promising results, this study has certain limitations that must be considered. Firstly, the moderate dataset size (n = 165) and its inherent class imbalance may affect the model's generalizability, particularly at high k-values where the bias toward the majority class is accentuated. In practical terms, this potential Type I error (false high yield) could lead to estimates that are optimistically high yet fundamentally incorrect. Furthermore, to avoid exclusive reliance on the IBk/HEOM approach, future work should include comparisons with other algorithms designed for mixed data, such as those based on decision trees with heterogeneous metrics or ensemble methods adapted for this type of data structure. It is also vital to consider the static nature of the data, which is specific to a particular moment and productive condition. This may overlook temporal or agroecological dynamics that could influence metabolomic and morphological profiles and, consequently, the final yield.

As future work, we propose expanding the dataset with samples from multiple agricultural cycles and contrasting agroecological and productive conditions. This will allow for an evaluation of the model's spatiotemporal stability, thereby helping to prevent model "memorization" and instead foster an optimal learning phase that captures scenario heterogeneity. Similarly, we suggest exploring the use of alternative, robust algorithms capable of handling mixed-data complexity. This would enable a systematic comparison of performance and robustness to select the most effective and efficient model for the scenarios of interest.

5 Conclusions

The implementation of the IBk/HEOM algorithm proved highly effective in classifying the yield of habanero peppers using mixed features of metabolomic, morphological, and phenotypic origin. The results confirm that within a k-range of 1 to 25, the model achieves perfect performance, reflecting the existence of highly discriminatory biomarkers and the suitability of the approach for capturing complex relationships between qualitative and quantitative features. This implies that a small number of neighbors can be chosen to reduce computational cost without compromising predictive efficacy. The transition observed from k = 26 onwards, marked by an increase in false positives, demonstrated the algorithm's sensitivity to neighborhood size, as well as the influence of class overlap and imbalance.

The identification of eight critical features suggests that a smaller feature set can possess predictive power comparable to the full set. This has important implications for optimizing future models and reducing the costs associated with field data collection and processing for metabolomic biomarker profiling. Furthermore, it provides a framework for maintaining predictive efficacy even in the absence of certain features that may be missing for various logistical, operational, or methodological reasons.

Additionally, HEOM's capacity to handle mixed data effectively, without the need for encoding or normalization that could distort the original data structure, highlights the advantages in agricultural contexts where integrating multiple information sources is essential, thereby enhancing the algorithm's interpretability and usability.

Collectively, this study sets a methodological precedent for intelligent classification in economically important production chains, providing a robust predictive model and an interpretative framework that facilitates the identification of key features associated with yield.

Acknowledgments

This work was supported by a scholarship from Mexico's Secretariat of Science, Humanities, Technology and Innovation (SECIHTI). The authors additionally acknowledge the Centro de Investigación en Computación of the Instituto Politécnico Nacional (CIC-IPN) for the institutional support throughout this research.

References

1. **Alemu, A., Åstrand, J., Montesinos-López, O.A., Isidro Y Sánchez, J., Fernández-Gónzalez, J., Tadesse, W., Vetukuri, R.R.,**

Carlsson, A.S., Ceplitis, A., Crossa, J., Ortiz, R., Chawade, A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. Mol Plant. 1;17(4), pp. 552–578. DOI: 10.1016/j.molp. 2024.03.007.

2. **Slimani, H., El Mhamdi, J., Jilbab, A. (2025).** UAV-based Systems for Advanced Crop Growth Monitoring with Deep Learning Framework in Complex Agriculture. Computación y Sistemas, 29(2), pp. 687–700. DOI: 10.13053/CyS-29-2-4785.

3. **García-Amaro, E., Cervantes-Canales, J., García-Lamont, F., Lara-Viveros, F. M., Ruiz-Castilla, J. S., Espejel Cabrera, J. (2024).** Use of Computer Vision Techniques for Recognition of Diseases and Pests in Tomato Plants. Computación y Sistemas, 28(2), pp. 709–723. DOI: 10.13053/CyS-28-2-3927.

4. **Jiménez-Galina, A. M., Maldonado-Macías, A. A., Olmos-Sanchez, K. M., Hernández, I., Estrada-Saldaña, F., Vázquez-Gálvez, F. A. (2024).** Framework for Heterogeneous Data Management: An Application Case in a NoSQL Environment from a Climatological Center. Computación y Sistemas, 28(1), pp. 167–178. DOI: 10.13053/CyS-28-1-4474.

5. **2- Wu, C., Peng, Y., Li, Y., Liu, X., Gao, X., Feng, Y., Zhou, Y. (2024).** Multi-omics assists genomic prediction of maize yield with machine learning approaches. Theoretical and Applied Genetics, 137, pp. 773–787. DOI: 10.1007/s11032-024-01454-z.

6. **Togninalli, M., Wang, X., Kucera, T., Shrestha, S., Juliana, P., Mondal, S., Pinto, F., Govindan, V., Crespo-Herrera, L., Huerta-Espino, J., Singh, R. P., Borgwardt, K., Poland, J. (2023).** Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics. Bioinformatics, 39(6), btad336. DOI: 10.1093/bioinformatics/btad336.

7. **Sun, Z., Li, Q., Jiang, D., Jin, S., Song, Y., Zhai, Z. (2022).** Simultaneous prediction of wheat yield and grain protein content using multitask deep learning from time-series proximal sensing. Plant Phenomics, 2022, Article 9757948. DOI: 10.34133/2022/9757948.

8. **Li, Z., Zhang, B., Wang, Q., Dong, H., Chen, J., Liu, X. (2024).** Enhancing winter wheat prediction with genomics, phenomics and environmental data. BMC Genomics, 25, Article 544. DOI: 10.1186/s12864-024-10438-4.

9. **Multi-Omics Assists Genomic Prediction of Maize Yield with Machine Learning Approaches. (2023).** Frontiers in Plant Science, (sección de agronomía). DOI: 10.1007/s11032-024-01454-z.

10. **Hamid H., Maleki, R., Darvishzadeh, N., Azad, N. (2025).** Sweet pepper yield modeling via deep learning and selection of superior genotypes using GBLUP and MGIDI. Scientific Reports. DOI: 10.1038/s41598-025-99779-y.

11. **Huang, W., Zhang, Y., Liu, J., Zhu, Z., et al. (2021).** The metabolomic landscape of rice heterosis highlights pathway biomarkers for predicting complex phenotypes. Plant Physiology, 187(2), pp. 1011–1025. DOI: 10.1093/plphys/kiab273.

12. **Wang, Q., Xu, J., Wang, K., Li, Y., Li, S., Liang, H., Wang, Y. (2022).** Integrative analyses of metabolome and genome-wide transcriptome reveal the regulatory network governing flavor formation in kiwifruit (Actinidia chinensis). New Phytologist, 233(5), pp. 2077–2094. DOI: 10.1111/nph.17618.

13. **Guijarro-Real, C., Adalid-Martínez, A. M., Pires, C. K., Ribes-Moya, A. M., Fita, A., Rodríguez-Burruezo, A., Adalid-Martínez, A. M. (2023).** The effect of the varietal type, ripening stage, and growing conditions on the content and profile of sugars and capsaicinoids in Capsicum peppers. Plants, 12(2), pp. 231. DOI: 10.3390/plants12020231.

14. **Hu, Y., Cai, Y., Wang, H., Xiong, Y., Zhang, X., Wei, L., Qiao, Z. (2022).** Systematic study of the sensory quality, metabolomics, and microbial community of fresh-cut watermelon provides new clues for its quality control and preservation. Foods, 11(21), pp. 3423. DOI: 10.3390/foods11213423.

15. **Pineda-Barneto, C., et al. (2022).** Metabolomic selection for enhanced fruit

flavor. PNAS, 119(7). DOI: 10.1073/pnas. 2115865119.

16. **Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., Khraisat, A. (2024).** Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data, 11, pp. 113. DOI: 10.1186/s40537-024-00973-y.

17. **Wu, C., Hargreaves, C. A. (2021).** Topological Machine Learning for Mixed Numeric and Categorical Data. World Scientific. DOI: 10.1142/S0218213021500251.

18. **Tuerhong, G., Wushouer, M., Zhang, D. (2021).** An Improved K Nearest Neighbor Classifier for High-Dimensional and Mixture Data. Journal of Physics: Conference Series, 1813(1), 012026. DOI: 10.1088/1742-6596/18 13/1/012026.

19. **Gárate-Escamilla A. K., Ortiz-Bayliss J. C., Terashima-Marín H. (2025).** Machine Learning, Missing Values, and Algorithm Selectors: The Untold Story. Computación y Sistemas, 29(1), pp. 311–323. DOI: 10.13053/ CyS-29-1-5508.

20. **Santos, M. S., Rivero, D., de Sá, J. E. (2022).** The impact of heterogeneous distance functions on K-Nearest Neighbors imputation of heterogeneous datasets. Information Sciences, 611, pp. 630–647. DOI: 10.1016/j. engappai.2022.104791.

21. **Acevedo-Sánchez, G., Alarcón-Paredes, A. Yáñez-Márquez, C. (2025).** Effect of agriculture-related dataset complexity on classical machine learning and deep learning classifiers performance. Computers and Electronics in Agriculture, 239, pp. 110941. DOI: 10.1016/j.compag.2025.110941.

22. **Aha, D.W., Kibler, D., (1991).** Instance-based learning algorithms. Mach. Learn. 6, pp. 37–66. DOI: 10.1007/BF00153759.

23. **Ramírez-Meraz, M., Méndez-Aguilar, R., Zepeda-Vallejo, L. G., Hernández-Guerrero, C. J., Hidalgo-Martínez, D., Becerra-Martínez, E., et al. (2024).** Exploring the chemical diversity of Capsicum chinense cultivars using NMR-based metabolomics and machine learning methods. Food Research International, 178, 113796. DOI: 10.1016/ j.foodres.2023.113796.

24. **Velázquez-Rodríguez, JL, Villuendas-Rey, Y, Camacho-Nieto, O, Yáñez-Márquez, C. (2020).** A novel and simple mathematical transform improves the performance of Lernmatrix in instance classification. Mathematics 8(5), pp. 732. DOI: 10.3390/ math8050732.

25. **Adin, A., Teixeira, E.K., Lenzi, A., Liu, Z., Martínez-Minaya, J., Rue, H. (2024).** Automatic cross-validation in structured models: Is it time to leave out leave-one-out? Spat. Stat. 62, 100843. DOI: 10.1016/j.spasta. 2024.100843.

26. **Vanacore, A., Pellegrino, M. S., Ciardiello, A. (2024).** Fair evaluation of classifier predictive performance based on binary confusion matrix. Computational Statistics, 39, 363–383. DOI: 10.1007/s00180-022-01301-9

27. **Canbek, G., Taskaya Temizel, T., Sagiroglu, S. (2021).** BenchMetrics: a systematic benchmarking method for binary classification performance measures. Neural Computing and Applications, 33, 14241-14264. https://doi.org/10.1007/s00521-021-06103-6

28. **RStudio Team (2024).** RStudio: Integrated Development Environment for R Posit Software (2024). https://posit.com/download/ rstudio/

29. **Zamljen, T., Medič, A., Veberič, R., Hudina, M., Jakopič, J., Slatnar, A. (2022).** Metabolic Variation among Fruits of Different Chili Cultivars (Capsicum spp.) Using HPLC/MS. Plants, 11(1), pp. 101. DOI: 10.3390/plants 11010101.

30. **Camposeco-Montejo, N., Flores-Naveda, A., Ruiz-Torres, N., Álvarez-Vázquez, P., Niño-Medina, G., Ruelas-Chacón, X., Torres-Tapia, M. A., Rodríguez-Salinas, P., Villanueva-Coronado, V., García-López, J. I. (2021).** Agronomic Performance, Capsaicinoids, Polyphenols and Antioxidant Capacity in Genotypes of Habanero Pepper Grown in the Southeast of Coahuila, Mexico.

Horticulturae, 7(10), pp. 372. DOI: 10.3390/horticulturae7100372.

31. **Li, S., Zhou, Y., Tang, L., Lu, J., Feng, Q., Wang, S., Jing, H., Liu, Z., Zhu, X. (2021).** The metabolomic landscape of rice heterosis highlights pathway biomarkers for predicting complex phenotypes. Plant Physiology, 187(2), pp. 1011–1025. DOI: 10.1093/plphys/kiab273.

32. **Yang, R., Li, Y., Zhang, Y., et al. (2021).** Widely Targeted Metabolomics Analysis Reveals Key Quality-Related Metabolites in Kernels of Sweet Corn. International Journal of Genomics, 2654546. DOI: 10.1155/2021/2654546.

33. **Razzaq A, Sadia B, Raza A, Khalid Hameed M, Saleem F. (2019).** Metabolomics: A Way Forward for Crop Improvement. Metabolites. 14;9(12), pp. 303. DOI: 10.3390/metabo9120303.

34. **Prade, V.M., Sun, N., Shen, J., Feuchtinger, A., Kunzke, T., Buck, A., Schraml, P., Moch, H., Schwamborn, K., Autenrieth, M., Gschwend, J.E., Erlmeier, F., Hartmann, A. and Walch, A. (2022).** The synergism of spatial metabolomics and morphometry improves machine learning-based renal tumour subtype classification. Clin. Transl. Med., 12, pp. e666. DOI: 10.1002/ctm2.666.

35. **National Center for Biotechnology Information. (2023).** Best practices for DNA methylation data analysis. In Montesinos López OA, Montesinos López A, Crossa J. (Ed.), Bioinformatics in epigenetic research. National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK583970/.

36. **Lumumba, V. W., Kiprotich, D., Mpaine, M. L., Makena, N. G., Kavita, M. D. (2024).** Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models. American Journal of Theoretical and Applied Statistics, 13(5), 127-137. DOI: 10.11648/j.ajtas.20241305.13.

37. **Chacón, A. M. P., Ramírez, I. S., Márquez, F. P. G. (2023).** K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection. Sustainable Futures, 100132. DOI: 10.1016/j.sftr.2023.100132.

38. **Fernández, O., Urrutia, M., Bernillon, S., Giauffret, C., Tardieu, F., Le Gouis, J., Langlade, N., Charcosset, A., Moing, A., Gibon, Y. (2016).** Fortune telling: metabolic markers of plant performance. Metabolomics, 12(10), pp. 158. DOI: 10.1007/s11306-016-1099-1.

39. **Dan, Z., Chen, Y., Zhao, W., Wang, Q., Huang, W., et al. (2020).** Metabolome-based prediction of yield heterosis contributes to the breeding of elite rice. Life Science Alliance, 3(1), e201900551. DOI: 10.26508/lsa.201900551.

40. **Prade, V. M., Sun, N., Shen, J., Feuchtinger, A., Kunzke, T., Buck, A., Schraml, P., Moch, H., Schwamborn, K., Autenrieth, M., Gschwend, J. E., Erlmeier, F., Hartmann, A., Walch, A. (2022).** The synergism of spatial metabolomics and morphometry improves machine learning-based renal tumour subtype classification. Clinical and Translational Medicine, 12(2), e666. DOI: 10.1002/ctm2.666.

41. **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, pp. 6. DOI: 10.1186/s12864-019-6413-7.

42. **Cohen, J. (1960).** A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), pp. 37–46. DOI: 10.1177/001316446002000104.

43. **Zhang, Y., Xu, T., Cheng, D., Li, J., Liu, L., Xu, Z., Feng, Z. (2025).** Data-driven learning optimal K values for K-nearest neighbor matching in causal inference. Data Mining and

Knowledge Discovery. DOI: 10.1007/s10618-025-01107-5.

44. **Cao, Y., Li, X., Song, H., Abdullah, M., Manzoor, M. A. (2024).** Editorial: Multi-omics and computational biology in horticultural plants: from genotype to phenotype, volume II. Frontiers in Plant Science, 15, 1368909. DOI: 10.3389/fpls.2024.1368909.

45. **Mi X, Zou B, Zou F, Hu J. (2021).** Permutation-based identification of important biomarkers for complex diseases via machine learning models. Nat Commun. 21;12(1), pp. 3008. DOI: 10.1038/s41467-021-22756-2.

Article received on 07/05/2025; accepted on 11/08/2025.
**Corresponding author is Cornelio Yáñez-Márquez.*