

## Machine learning–assisted early selection of S1 segregants in strawberry, *Fragaria × ananassa* Duch

Selección temprana de segregantes S1  
en fresa (*Fragaria × ananassa* Duch) asistida por aprendizaje automático

Liz Katherine Lagos-Santander <sup>1</sup>, David Esteban Duarte-Alvarado <sup>1</sup>, Tulio César Lagos-Burbano <sup>1</sup>.

<sup>1</sup>. Universidad de Nariño, Pasto, Nariño, Colombia. [lkagoss@unal.edu.co](mailto:lkagoss@unal.edu.co); [deduarte@unal.edu.co](mailto:deduarte@unal.edu.co); [tclagosb@udenar.edu.co](mailto:tclagosb@udenar.edu.co)

**Cite:** Lagos-Santander, L. K., Duarte-Alvarado, D. E., & Lagos-Burbano, T. C. (2025). Machine learning–assisted early selection of S1 segregants in strawberry, *Fragaria × ananassa* Duch. *Revista de Ciencias Agrícolas*, 43(2), e2304. <https://doi.org/10.22267/rcia.2026432.304>

### ABSTRACT

Strawberry (*Fragaria × ananassa*) breeding requires analytical tools that integrate productivity and fruit quality in highly variable segregating populations. This study evaluated the integration of mixed models and machine learning algorithms to support early selection of S1 genotypes with high yield potential and fruit quality. A total of 320 genotypes were evaluated under an augmented block design for phenological, vegetative, productive, and fruit quality traits. Broad-sense heritabilities and predicted genetic values were estimated using mixed models and integrated with predictions from XGBoost and CatBoost. Moderate to high heritabilities were observed for structural and productive traits ( $H^2 = 0.52\text{--}0.86$ ). Positive genetic correlations were found between fruit number and yield, whereas soluble solids content was negatively associated with yield components, indicating trade-offs between productivity and fruit quality. XGBoost showed high predictive accuracy for commercial yield ( $R^2 = 0.908$ ), intermediate performance for fruit weight ( $R^2 = 0.461$ ), and poor performance for soluble solids content as a continuous variable ( $R^2 = -0.096$ ). Recoding soluble solids content into categorical classes improved interpretability, although class discrimination remained limited. Selection indices based on best linear unbiased predictors (BLUPs), machine learning predictions, and their combination produced consistent rankings and identified genotypes with balanced agronomic and quality performance. Integrating mixed models and machine learning into multivariate selection indices provides a useful strategy for improving early selection efficiency in strawberry breeding programs.

**Keywords:** CatBoost; fruit quality; heritability; machine learning algorithms; selection index; XGBoost

### RESUMEN

El mejoramiento genético de la fresa (*Fragaria × ananassa*) requiere herramientas analíticas que integren la productividad y la calidad del fruto en poblaciones segregantes altamente variables. Este estudio evaluó la integración de modelos mixtos y algoritmos de aprendizaje automático para apoyar la selección temprana de genotipos S1 con alto potencial de rendimiento y calidad de fruto. Se evaluaron 320 genotipos bajo un diseño de bloques aumentados para caracteres fenológicos, vegetativos, productivos y de calidad del fruto. Las heredabilidades en sentido amplio y los valores genéticos predichos se estimaron mediante modelos mixtos y se integraron con predicciones de XGBoost y CatBoost. Se observaron heredabilidades moderadas a altas para caracteres estructurales y productivos ( $H^2 = 0,52\text{--}0,86$ ). Se encontraron correlaciones genéticas positivas entre el número de frutos y el rendimiento, mientras que el contenido de sólidos solubles se asoció negativamente con los componentes del rendimiento, indicando compromisos entre productividad y calidad del fruto. XGBoost mostró alta precisión predictiva para el rendimiento comercial ( $R^2 = 0,908$ ), desempeño intermedio para el peso del fruto ( $R^2 = 0,461$ ) y bajo desempeño para el contenido de sólidos solubles como variable continua ( $R^2 = -0,096$ ). La recodificación del contenido de sólidos solubles en clases categóricas mejoró la interpretabilidad, aunque la discriminación entre clases continuó siendo limitada. Los índices de selección basados en los mejores predictores lineales insegados (BLUP), las predicciones de aprendizaje automático y su combinación produjeron rankings consistentes e identificaron genotipos con desempeño agronómico y de calidad equilibrado. La integración de modelos mixtos y aprendizaje automático en índices de selección multivariados proporciona una estrategia útil para mejorar la eficiencia de la selección temprana en programas de mejoramiento genético de fresa.

**Palabras clave:** algoritmos de aprendizaje automático; calidad del fruto; CatBoost; heredabilidad; índice de selección; XGBoost

## INTRODUCTION

Strawberry (*Fragaria × ananassa* Duch.) has consolidated its strategic importance within global horticultural production systems, driven by high profitability, increasing integration into technologically advanced value chains, and sustained demand for fruits with superior quality standards. This productive dynamism is reflected in the active participation of multiple stakeholders, ranging from specialized nurseries and clonal propagation units to collection, processing, and export platforms, which have positioned strawberry as a highly competitive crop worldwide (Gutiérrez-Esquivel & Ruíz-Rivas, 2025). In this context, countries such as the United States, Mexico, Turkey, Spain, and Egypt maintain consolidated leadership, supported by intensive production systems under protected environments, precision agronomic management, and substantial advances in breeding programs (Acuña Caíta & Fischer Gebauer, 2024).

In Colombia, national strawberry production between 2020 and 2024 ranged from 99,900 to 110,402 tons, cultivated on 2,520 to 3,307 ha, with yields fluctuating between 33.40 and 39.14 t ha<sup>-1</sup> (AGRONET, 2025). Despite increases of 6.07% in planted area and 3.28% in total production, average yield declined by 2.85%, indicating that the expansion of cultivation has not been accompanied by proportional improvements in physiological efficiency or agronomic performance. This trend underscores the need to optimize varietal selection processes and crop management, particularly in regions facing edaphoclimatic constraints or emerging biotic pressures. In the Colombian context, predictive early selection strategies could help national breeding programs reduce field space, evaluation time, and operational costs, while accelerating the identification of genotypes with high yield and fruit quality. In addition to its economic relevance, strawberry is valued for fruit quality attributes related to sweetness, firmness, postharvest performance, and functional compounds, increasing the demand for cultivars that combine high yield, superior fruit quality, and adaptation to environmental stresses (Newerli-Guz *et al.*, 2023; Altieri *et al.*, 2025).

Within this framework, machine learning (ML) tools have emerged as a robust methodological alternative for expanding the predictive capacity of selection models, particularly in segregating populations where phenotypic variability is high and the heritability of certain traits is moderate or environmentally dependent. The integration of algorithms such as XGBoost, Random Forest, CatBoost, and deep neural networks enables the modeling of nonlinear relationships, the capture of complex interactions among variables, and the simultaneous use of multiple sources of phenotypic, environmental, or potentially genomic information, thereby improving the accuracy of early selection (Tong & Nikoloski, 2021; Jones *et al.*, 2023; M'hamdi *et al.*, 2024; Şimşek, 2024). Although machine-learning algorithms such as Random Forest, XGBoost, and CatBoost have been explored in fruit and vegetable crops, their application in early selection within genetic improvement programs remains limited. Recent studies have demonstrated the potential of these methods for modeling fruit-related traits, including postharvest quality assessment, nondestructive fruit quality prediction, and fruit size prediction (Chen *et al.*, 2024; Lin *et al.*, 2024; Takahashi *et al.*, 2025). These methodologies have shown clear advantages over traditional linear approaches, particularly when data structures exhibit high dimensionality, strong internal correlations, or non-trivial patterns of variability.

Strawberry S1 populations display high levels of segregation for critical traits such as yield, soluble solids content (°Brix), firmness, fruit size, and nutritional quality components. This phenotypic heterogeneity complicates the timely identification of

superior plants and increases the uncertainty associated with the selection process, which has historically depended on prolonged agronomic cycles and substantial experimental resources (Chandler *et al.*, 2012; Zeist & Resende, 2019; Sturzeanu *et al.*, 2025). Consequently, strategies are required that enable the anticipation of the performance segregants, the reduction of unexplained variability, and the prioritization of promising individuals based on more stable and reproducible predictions.

The use of machine learning models in segregating strawberry populations, particularly those based on boosting and ensemble learning, provides an opportunity to capture complex patterns in the expression of quantitative traits and to generate more robust predictions for decision-making at early stages of the selection cycle (Brasileiro *et al.*, 2015; Khatibi & Ali, 2024). These computational approaches also enable the construction of multi-objective indices that integrate normalized predictions of key variables such as fruit weight, soluble solids content, and yield, thereby facilitating the prioritization of genotypes with balanced performance and commercial potential.

It was hypothesized that the integration of mixed models and machine learning algorithms strengthens the selection of superior S1 genotypes by capturing non-linear relationships between phenotypic traits associated with fruit yield and quality. Under this scenario, the present study aimed to evaluate the capacity of machine learning to support and optimize the early selection of S1 segregants in *Fragaria × ananassa*. Specifically, an analytical workflow based on algorithms such as XGBoost and CatBoost was developed and validated by integrating available phenotypic information with predictions from nonlinear models to improve early selection accuracy and identify segregants with superior agronomic performance and fruit quality and quality potential.

## MATERIALS AND METHODS

The experiment was conducted at the Botana Experimental Farm of the University of Nariño, located in the municipality of Pasto, department of Nariño, southern Colombia, at 2,820 m a.s.l. (1°10'11.4" N, 77°18'58" W). The site is characterized by an average temperature of 14 °C, annual rainfall ranging from 800 to 1,000 mm and relative humidity between 70 and 80%. Planting was conducted under greenhouse conditions, with an average temperature of 25.3 °C and relative humidity of 62.84%. Plant spacing was 0.30 m between plants and 0.35 m between rows.

A total of 320 S1 genotypes were derived from six S0 hybrid combinations included in the study. These families were organized according to the breeding program's series classification. Specifically, the hybrids (Amiga × 2846-5)–S0, (Fontanilla × Candonga)–S0, (Fuentepina × Fontanilla)–S0, (Candonga × 1806-1)–S0, (Santa Clara × Aguadilla)–S0, and (Candonga × 2846-5)–S0 contributed 47, 62, 70, 77, 28, and 36 S1 genotypes, respectively, corresponding to Series 100, 200, 300, 400, 500, and 600.

The experiment was established using an Augmented Block Design (Federer & Raghavarao, 1975) with four blocks. In the S1 generation, each plant corresponded to a unique recombinant individual derived from its parental cross; therefore, each genotype was represented by a single plant and treated as one experimental unit with one replication. Accordingly, the 320 S1 genotypes were included as non-common treatments, whereas the three control cultivars (Albion, Monterrey, and Regional) were evaluated as common treatments across blocks. For the control cultivars, 50 plants were established per block, from which 10 plants were sampled for data collection following the methodology proposed by Vieira *et al.* (2017).

Phenological, vegetative, productive, and fruit quality traits were evaluated. Among the phenological variables, days to flowering (DF) were recorded and defined as the number of days elapsed from planting to the opening of the first flower. Vegetative traits included the number of crowns per plant (CP), plant height (PH), and the number of leaves (NL), which characterize plant vigor (PV).

Productive traits included the total number of fruits per plant (NF), the number of commercial fruits per plant (CFP), average fruit weight (FW), equatorial fruit diameter (ED), polar fruit diameter (PD), total yield per plant (TY), commercial yield per plant (C-yield), and harvest losses due to pests and diseases (HL). These traits allowed the assessment of both overall productivity and commercial efficiency of each genotype. Soluble solids content ( $^{\circ}$ Brix, BX) was evaluated as a quality trait using refractometry and is directly associated with perceived fruit sweetness. In addition, for classification analyses, BX was transformed into a binary categorical variable (BX\_cat), with fruits classified as sweet when values were greater than or equal to 8  $^{\circ}$ Brix.

The data analysis was conducted using an integrated approach that combined Analysis of Variance (ANOVA) based on the Augmented Block Design (Zystro *et al.*, 2018) and machine learning (ML) techniques. Initially, linear mixed models were fitted for each evaluated trait, considering block as a fixed effect and genotype as a random effect. From these models, genetic and residual variance components were estimated, and broad-sense heritability ( $H^2$ ) was calculated as the proportion of total genetic variance relative to total phenotypic variance, adjusted by the average number of observations per family. In addition, best linear unbiased predictors (BLUPs) were obtained for each genotype and trait, representing estimates of genetic merit free from systematic environmental effects.

All statistical and machine learning analyses were performed in Google Colab using Python 3.12. Data management and numerical operations were conducted using *pandas* and *numpy*; preprocessing, dataset partitioning, model evaluation, and performance metrics were implemented using *scikit-learn*; mixed linear models and BLUP estimation were performed using *statsmodels*; XGBoost models were fitted using the *xgboost* library; CatBoost models were fitted using the *catboost* library; and class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique) from the imbalanced-learn package. Data visualization was performed using *matplotlib* and *seaborn*.

BLUPs were used to estimate genetic correlations among traits through correlation matrices computed on predicted genetic values. Additionally, ML models based on extreme gradient boosting (XGBoost) (Al-Taai *et al.*, 2023) were implemented to predict FW and yield using the available traits as predictors. Prior to model training, the data underwent a preprocessing pipeline that included standardization of continuous variables. The datasets were randomly split into training (80%) and testing (20%) subsets.

Regression model performance was assessed using the coefficient of determination ( $R^2$ ), mean absolute error (MAE), and root mean square error (RMSE). In addition, the effect of outlier filtering was evaluated using different standard deviation thresholds, and the threshold that maximized predictive ability for each trait was selected.

For soluble solids content, BX was transformed into a binary categorical trait (BX\_cat), with fruits classified as sweet when  $BX \geq 8^{\circ}$  Brix. Classification models were trained using XGBoost and CatBoost, which are suitable for modeling nonlinear relationships and heterogeneous data structures (Friedman, 2001; Chen & Guestrin, 2016; Prokhorenkova *et al.*, 2018). The XGBoost classifier was fitted using  $n\_estimators = 600$ ,  $max\_depth = 6$ ,  $learning\_rate = 0.05$ ,  $subsample = 0.90$ ,  $colsample\_bytree = 0.90$ ,  $reg\_lambda = 1.0$ ,  $objective = \text{"binary:logistic"}$ , and  $tree\_method = \text{"hist"}$ . The CatBoost classifier was fitted

using *iterations* = 800, *depth* = 8, *learning\_rate* = 0.05, *loss\_function* = "Logloss", *eval\_metric* = "AUC", and *random\_seed* = 42; categorical predictors, including Pedigree\_S0, Pedigree\_S1, Trat, and Control were handled directly by CatBoost.

Model performance was evaluated using overall accuracy, the area under the receiver operating characteristic curve (AUC), confusion matrices, and the precision, recall, and F1-score metrics (Powers, 2011). Given the observed class imbalance, SMOTE was applied only to the training subset to avoid information leakage into the testing subset (Chawla *et al.*, 2002). Before applying SMOTE, the original BX\_cat distribution consisted of 63.86% sweet genotypes and 36.14% non-sweet genotypes. Within the training subset, 265 observations were available before balancing. After applying the SMOTE, the training set increased to 338 observations, with a balanced distribution of 169 sweet and 169 non-sweet cases. SMOTE generates synthetic observations of the minority class through linear interpolation between neighboring samples.

Finally, multivariate selection indices were constructed by integrating genetic and predictive information. First, an index based exclusively on BLUPs (SI-BLUP) was calculated by combining key yield and fruit-quality traits using previously defined weights. Second, an index based on machine learning predictions (SI-ML) was developed using the values predicted by XGBoost and the sweetness probabilities obtained from the classification models. Subsequently, both approaches were integrated into a combined index (SI-BLUP-ML), which enabled the prioritization of the most promising genotypes by simultaneously considering genetic merit, expected productive performance, and fruit quality.

## RESULTS

The mean squares from the ANOVA of the Augmented Randomized Complete Block design indicated significant differences for commercial yield (C-yield), the number of commercial fruits per plant (CFP), average fruit weight (FW), and equatorial fruit diameter (ED). The ANOVA (Table 1) revealed significant genotypic effects for commercial yield, the number commercial of fruits per plant (CFP), average fruit weight (FW), and soluble solids content (BX). In contrast, polar fruit diameter (PD), harvest losses due to pests and diseases (HL), and plant vigor (PV) did not show significant variation, suggesting limited genetic variance for these traits.

Broad-sense heritability estimates ( $H^2$ ) revealed moderate to high values for most of the evaluated traits (Table 2). The highest heritabilities were observed for the number of crowns (CP,  $H^2$  = 0.85), equatorial fruit diameter (ED,  $H^2$  = 0.86), and average fruit weight (FW,  $H^2$  = 0.85), indicating a strong genetic contribution to phenotypic expression and a high potential for response to selection (Zareei *et al.*, 2023). Similarly, days to flowering (DF;  $H^2$  = 0.74) showed high heritability, whereas commercial yield exhibited a moderate-to-high value ( $H^2$  = 0.68).

Traits with moderate heritability included polar fruit diameter (PD,  $H^2$  = 0.52), the number of commercial fruits per plant (CFP,  $H^2$  = 0.53), and soluble solids content (BX,  $H^2$  = 0.64). This suggests a relatively greater environmental influence on their phenotypic expression, although with a sufficient genetic basis to be considered in selection schemes.

**Table 1.** Mean squares from the ANOVA of the Augmented Randomized Complete Block design for traits evaluated in 320 S1 genotypes of *Fragaria × ananassa* and three control cultivars.

| Source         | df  | BX                 | ED                 | DP                 | CFP                 | FW                 | HL                   | C-yield                | PV                 |
|----------------|-----|--------------------|--------------------|--------------------|---------------------|--------------------|----------------------|------------------------|--------------------|
| Treatment (IB) | 322 | 2.51*              | 0.31*              | 0.47 <sup>ns</sup> | 859.45*             | 4.61*              | 106.90 <sup>ns</sup> | 47,681.90*             | 0.69 <sup>ns</sup> |
| Check          | 2   | 0.87 <sup>ns</sup> | 0.82*              | 1.14*              | 62.32 <sup>ns</sup> | 25.49*             | 222.97*              | 6,628.61 <sup>ns</sup> | 5.31*              |
| Test           | 319 | 2.52*              | 0.31*              | 0.46 <sup>ns</sup> | 854.50*             | 4.45*              | 106.54 <sup>ns</sup> | 47,647.73*             | 0.67 <sup>ns</sup> |
| Test vs. Check | 1   | 2.41 <sup>ns</sup> | 0.24*              | 0.01 <sup>ns</sup> | 4030.63*            | 13.89*             | 5.35 <sup>ns</sup>   | 140,691.72*            | 0.00 <sup>ns</sup> |
| Block (ET)     | 3   | 0.59 <sup>ns</sup> | 0.07 <sup>ns</sup> | 0.59 <sup>ns</sup> | 51.78 <sup>ns</sup> | 1.46 <sup>ns</sup> | 48.77 <sup>ns</sup>  | 9,609.60 <sup>ns</sup> | 0.04 <sup>ns</sup> |
| Residuals      | 6   | 0.44               | 0.02               | 0.14               | 117.06              | 0.48               | 37.65                | 5,744.43               | 0.22               |
| OAM            |     | 8.74               | 2.30               | 3.30               | 41.02               | 6.81               | 12.59                | 281.40                 | 2.89               |
| CV             |     | 7.62               | 6.48               | 11.42              | 27.21               | 10.09              | 49.51                | 27.71                  | 16.34              |

Note. IB (ignoring blocks), ET (eliminating treatments), OAM (overall adjusted mean), BX (soluble solids content), ED (equatorial fruit diameter), DP (polar fruit diameter), CFP (number of commercial fruits per plant), FW (fruit weight), HL (harvest losses due to pests and diseases), PV (plant vigor), and C-yield (commercial yield). \* = significant at  $p < 0.05$ ; ns = not significant. CV = coefficient of variation (%).

**Table 2.** Broad-sense heritability ( $H^2$ ) and genetic ( $r_g$ , above the diagonal) and phenotypic ( $r_f$ , below the diagonal) correlations among selected traits evaluated in S1 strawberry genotypes (*Fragaria × ananassa*).

| Trait   | $H^2$ | DF       | CP       | PD       | ED       | BX       | CFP      | FW       | C-yield  |
|---------|-------|----------|----------|----------|----------|----------|----------|----------|----------|
| DF      | 0.74  | <b>1</b> | -0.27*   | ns       | ns       | ns       | -0.15*   | Ns       | -0.14    |
| CP      | 0.85  | -0.25*   | <b>1</b> | ns       | ns       | ns       | 0.31*    | Ns       | 0.25*    |
| PD      | 0.52  | ns       | ns       | <b>1</b> | 0.30*    | ns       | 0.19*    | 0.46*    | 0.34*    |
| ED      | 0.86  | ns       | ns       | 0.33*    | <b>1</b> | -0.13*   | 0.20*    | 0.51*    | 0.36*    |
| BX      | 0.64  | ns       | ns       | ns       | -0.14*   | <b>1</b> | -0.12*   | -0.18*   | -0.17    |
| CFP     | 0.53  | -0.14*   | 0.30*    | 0.17*    | 0.20*    | -0.15*   | <b>1</b> | Ns       | 0.90*    |
| FW      | 0.85  | ns       | ns       | 0.47*    | 0.54*    | -0.18*   | ns       | <b>1</b> | 0.32*    |
| C-yield | 0.68  | -0.16*   | 0.24*    | 0.34*    | 0.38*    | -0.19*   | 0.90*    | 0.34*    | <b>1</b> |

Note. \* = significant relationships; ns = non-significant relationships. DF (days to flowering), CP (number of crowns per plant), PD (polar fruit diameter), ED (equatorial fruit diameter), BX (soluble solids content), CFP (number of commercial fruits per plant), FW (fruit weight), C-yield (Commercial yield).

Regarding genetic correlations, positive and significant associations were observed among yield-related traits. A strong genetic correlation was detected between CFP and C-yield ( $r_g = 0.90^*$ ), indicating that increases in fruit number translate directly into higher yield levels. Likewise, FW exhibited positive genetic correlations with PD ( $r_g =$

0.46\*) and ED ( $r_g = 0.51^*$ ), as well as with C-yield ( $r_g = 0.32^*$ ), suggesting that larger fruit size contribute favorably to productivity. Similarly, positive genetic correlations were observed between PD and C-yield ( $r_g = 0.34^*$ ), indicating that fruit structural traits are genetically associated with yield. In contrast, low-magnitude negative genetic correlations were observed between soluble solids content (BX) and several productive traits, including ED ( $r_g = -0.13^*$ ), FW ( $r_g = -0.18^*$ ), and CFP ( $r_g = -0.12^*$ ). This indicates that genotypes with higher soluble solids content tend, at the genetic level, to exhibit reductions in fruit size and number, a relationship frequently observed when multiple quantitative traits are evaluated simultaneously in breeding programs (Falconer & Mackay, 1996). In addition, DF showed negative genetic correlations with CP ( $r_g = -0.27^*$ ) and CFP ( $r_g = -0.15^*$ ), indicating that genotypes with shorter flowering period tend to produce more fruits per plant.

Phenotypic correlations generally followed the same pattern as genetic correlations, although with slightly lower magnitudes, suggesting that the observed associations are mainly driven by genetic factors rather than exclusively by environmental effects. The strong phenotypic correlation between CFP and C-yield ( $r_f = 0.90^*$ ) reinforces the importance of CFP as a key yield component. Likewise, negative phenotypic correlations between BX and fruit size and yield-related traits confirmed the antagonism observed at the genetic level. Overall, these results indicate that the number of commercial fruits per plant (CFP), average fruit weight (FW), and equatorial fruit diameter (ED) constitute strategic traits for the indirect selection of yield, whereas soluble solids content (BX) should be considered with caution because of its negative associations with productive components. This highlights the need for multivariate selection approaches that allow simultaneous balancing of yield and fruit quality.

Regarding the performance of XGBoost-based regression models, substantial differences were detected among the evaluated traits and standard-deviation thresholds (Table 3). The highest predictive ability was obtained for C-yield, which reached a coefficient of determination ( $R^2$ ) of 0.908 and an RMSE of 66.63 under threshold three (3), a pattern consistent with the strong performance of gradient boosting algorithms reported in agronomic studies (Chen & Guestrin, 2016). The FW trait showed an intermediate fit, with its highest value of  $R^2$  (0.46) also observed at this threshold, indicating moderate predictive ability. In contrast, BX exhibited poor performance across all scenarios, with negative  $R^2$  values even under its best configuration ( $-0.10$ ), reflecting the limited capacity of the model to explain variability in this trait.

**Table 3.** Performance of XGBoost models under the optimal threshold (three) for response variables evaluated in S1 strawberry genotypes (*Fragaria* × *ananassa*).

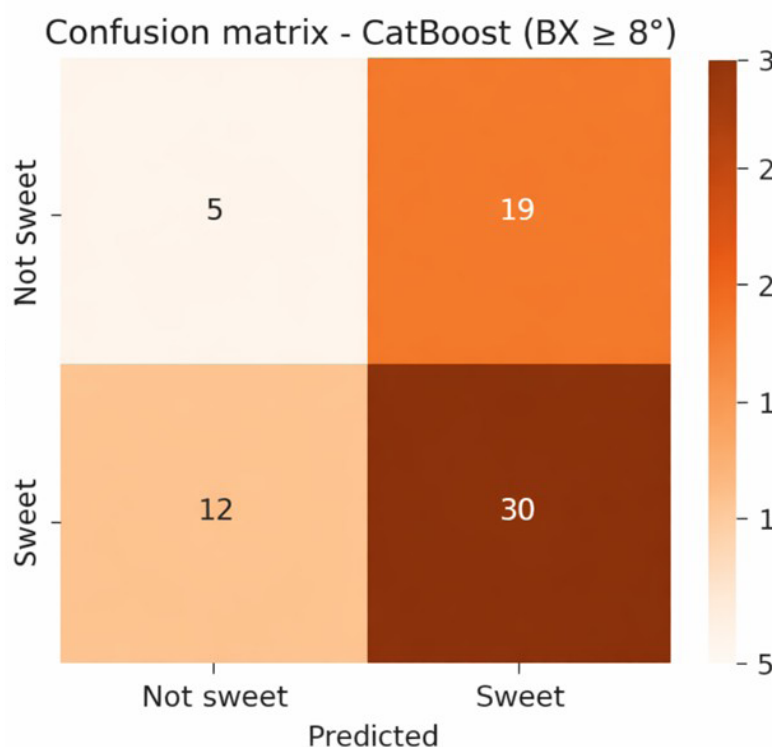
| Trait   | $R^2$ | MAE   | RMSE  | Interpretation              |
|---------|-------|-------|-------|-----------------------------|
| FW      | 0.46  | 1.03  | 1.33  | Moderate predictive ability |
| C-yield | 0.91  | 45.41 | 66.63 | Very high accuracy          |
| BX      | -0.10 | 1.06  | 1.45  | Poor performance            |

The poor performance observed in predicting soluble solids content (BX) as a continuous trait revealed limitations in the model's ability to capture the phenotypic variability associated with this trait. Consequently, improved predictive performance was observed when fruit sweetness was addressed through an alternative classification scheme using the categorical variable BX\_cat, an approach that has been reported as appropriate in contexts where the continuous signal is weak or strongly influenced by environmental factors (James *et al.*, 2021; Apostolopoulos *et al.*, 2023). In this context, the CatBoost model applied at the  $BX \geq 8^\circ\text{Brix}$  threshold showed differential performance between classes (Figure 1).

The algorithm correctly identified most fruits classified as sweet, with a sensitivity of 71%, whereas the ability to recognize non-sweet fruits was considerably lower (21%), evidencing a tendency to overclassify fruits into the sweet category. Overall, these results indicate that, although the classification approach improves the interpretability of the trait relative to direct regression, discriminatory power remains limited and could be strengthened through the incorporation of additional physiological or environmental traits (James *et al.*, 2021; Apostolopoulos *et al.*, 2023). The confusion matrix showed an asymmetric classification pattern. CatBoost correctly classified most sweet fruits, but its capacity to identify non-sweet fruits was limited, confirming a tendency to overclassify genotypes into the sweet category. This result suggests that the  $BX \geq 8^\circ\text{Brix}$  threshold improved the interpretability of sweetness prediction, but did not fully resolve the overlap between sweetness classes.

Reformulating BX as a binary categorical variable allowed a significant improvement in predictive performance. Classification models based on XGBoost and CatBoost achieved relatively high AUC values (0.52) and F1-scores (0.24 for non-sweet and 0.66 for sweet), particularly after applying class-balancing techniques using SMOTE.

**Figure 1.** Confusion matrix for CatBoost classification of fruit sweetness in S1 strawberry genotypes (*Fragaria* × *ananassa*) using  $BX \geq 8^\circ\text{Brix}$  as the sweetness threshold. Rows indicate observed classes and columns indicate predicted classes.



Finally, multivariate selection indices were constructed. The index based exclusively on BLUPs (SI\_BLUP) enabled the identification of genotypes with high genetic merit. The index based on machine learning predictions (SI\_ML) prioritized genotypes with high-expected performance and higher probabilities of sweetness. The combined index (SI\_BLUP\_ML) integrated both approaches, generating a more balanced ranking.

The Pearson correlation matrix (Table 4) quantifies the linear association between each multivariate selection index and the genetic values (BLUPs) of the individual traits included in their construction. This allows assessment of the implicit relative weight of each component within the composite selection formula and evaluation of the coherence among the different index strategies.

Very high phenotypic concordance was observed among the selection index methodologies, as evidenced by nearly perfect correlation coefficients ( $r = 0.99$ ) between the combined index (SI\_BLUP\_ML) and each of its components. This result confirms that the hybrid index effectively represents a synthesis of its two sources. The slightly lower—yet still very high—correlation ( $r = 0.96$ ) between SI\_BLUP and SI\_BLUP\_ML reflects the complementary nature of their underlying principles: whereas the former is based on genetic merit, the latter incorporates integrative phenotypic predictions that capture complex interactions, resulting in a subtly different genotype ranking.

**Table 4.** Pearson correlation matrix between selection indices and genetic values (BLUPs) for the traits of interest.

| SI-trait   | SI_BLUP | SI_ML | SI_BLUP_ML | FW   | C-yield | BX    |
|------------|---------|-------|------------|------|---------|-------|
| SI_BLUP    | 1       | 0.96  | 0.99       | 0.65 | 0.92    | -0.06 |
| SI_ML      |         | 1     | 0.99       | 0.59 | 0.92    | -0.19 |
| SI_BLUP_ML |         |       | 1          | 0.63 | 0.93    | -0.13 |
| FW         |         |       |            | 1    | 0.33    | -0.19 |
| C-yield    |         |       |            |      | 1       | -0.15 |
| BX         |         |       |            |      |         | 1     |

As expected, all three indices showed strong and positive associations with C-yield ( $r = 0.92$ – $0.93$ ), confirming that productivity is the main determinant of selection value in this breeding program. Correlations with fruit weight (FW) were also positive and of high magnitude ( $r = 0.59$ – $0.65$ ), indicating that this quality trait contributes to, but does not dominate, the construction of the selection index. This pattern reflects the weighting structure assigned to the selection indices, in which commercial yield was prioritized as the main breeding objective, whereas fruit weight and soluble solids content were included as complementary traits. Therefore, the strong positive association between the indices and C-yield, together with the low or negative correlations with BX, indicates that the current index formulation favored productive performance over organoleptic quality. This result should be considered in future breeding cycles, where greater relative weight could be assigned to soluble solids content or other sensory traits to achieve a better balance between yield and fruit quality.

The correlations observed between the three selection indices (SI-BLUP, SI-ML, and SI-BLUP-ML) and the predicted genetic effects (BLUPs) for soluble solids content (BX) were negative and low in magnitude, ranging from  $-0.06$  to  $-0.19$ . The most negative value corresponded to SI-ML. No substantial positive correlations were detected between any index and BX BLUPs.

Based on the selection indices (SI) described above, Table 5 summarizes the performance of the twenty S1 strawberry genotypes selected for having the highest SI values. Within the selected fraction, corresponding to 6.25% of the evaluated genotypes, substantial variability was observed for soluble solids content (BX), fruit weight (FW), and C-yield, as well as for the multivariate selection indices derived from BLUPs (SI\_BLUP), machine learning (SI\_ML), and their combination (SI\_BLUP\_ML).

In terms of sweetness, BX values ranged from 5.84 to 12.00 °Brix, with genotypes 326B2, 139B3, and 309B1 standing out for their high soluble solids content, whereas 115B2 and 237B3 exhibited the lowest values. For fruit weight, a range between 6.63 and 12.25 g was observed, with genotypes 606B1, 215B2, and 529B4 being particularly notable, indicating the presence of genetic materials with potential for fresh fruit markets.

Commercial yield displayed marked contrasts among genotypes, with the highest values recorded for 514B2 (1167.7 g plant<sup>-1</sup>), 237B3 (1026.7 g plant<sup>-1</sup>), and 115B2 (972.5 g plant<sup>-1</sup>), whereas 630B4 and 606B1 were located among the lowest-yielding genotypes. This pattern suggests the existence of physiological trade-offs between productivity and quality in certain materials, a phenomenon commonly observed in segregating populations, where contrasting genetic variation among traits determines the response to selection in breeding programs (Bernardo, 2020).

The selection indices enabled the integration of the multivariate response and consistent ranking of genotypes. According to Table 5, 514B2 ranked among the top positions across all three approaches, followed by 115B2 and 237B3, identifying them as promising genetic materials for advanced stages of the breeding program. In contrast, 104B1, 308B1, and 630B4 exhibited the lowest SI\_BLUP\_ML values, reflecting limited overall performance despite intermediate levels in some individual traits.

The partial concordance between SI\_BLUP and SI\_ML, reflected in the combined index SI\_BLUP\_ML, supports the complementarity between mixed models and machine learning algorithms in strengthening decision-making for early selection. Overall, the results summarized in Table 5 confirm the usefulness of integrative approaches for identifying superior genotypes in S1 strawberry populations with multiple breeding objectives.

**Table 5.** Agronomic performance and multivariate ranking of S1 strawberry genotypes based on BLUP indices (SI\_BLUP), machine learning (SI\_ML), and their combination (SI\_BLUP\_ML).

| Genotype | Pedigree                           | Series/parental origin        | BX    | FW    | C-yield | SI_BLUP | SI_ML | SI_BLUP_ML |
|----------|------------------------------------|-------------------------------|-------|-------|---------|---------|-------|------------|
| 514B2    | (Santa Clara x Aguadilla)-S1-514B2 | 500/(Santa Clara x Aguadilla) | 6.51  | 10.15 | 1167.73 | 2.70    | 3.09  | 2.90       |
| 115B2    | (Amiga x 2846-5)-S1-115B2          | 100/ Amiga x 2846-5           | 5.90  | 10.13 | 972.52  | 2.30    | 2.69  | 2.49       |
| 237B3    | (Fontanilla x Candonga)-S1-237B3   | 200/ Fontanilla x Candonga    | 5.84  | 8.56  | 1026.69 | 2.27    | 2.62  | 2.45       |
| 309B1    | (Fuentepina x Fontanilla)-S1-309B1 | 300/ Fuentepina x Fontanilla  | 10.88 | 8.83  | 883.29  | 2.20    | 2.22  | 2.21       |
| 219B2    | (Fontanilla x Candonga)-S1-219B2   | 200/ Fontanilla x Candonga    | 7.18  | 9.26  | 925.77  | 2.11    | 2.13  | 2.12       |
| 625B3    | (Candonga x 2846-5)-S1-625B3       | 600/ Candonga x 2846-5        | 7.07  | 9.70  | 882.47  | 2.01    | 2.22  | 2.11       |
| 529B4    | (Santa Clara x Aguadilla)-S1-529B4 | 500/(Santa Clara x Aguadilla) | 6.38  | 10.71 | 760.06  | 2.01    | 2.19  | 2.10       |
| 525B4    | (Santa Clara x Aguadilla)-S1-525B4 | 500/(Santa Clara x Aguadilla) | 6.86  | 8.21  | 870.30  | 2.00    | 2.07  | 2.04       |
| 626B3    | (Candonga x 2846-5)-S1-626B3       | 600/ Candonga x 2846-5        | 7.49  | 7.71  | 964.22  | 1.88    | 2.13  | 2.01       |
| 624B3    | (Candonga x 2846-5)-S1-624B3       | 600/ Candonga x 2846-5        | 6.94  | 7.88  | 921.62  | 1.71    | 2.00  | 1.86       |
| 428B2    | (Candonga x 1806-1)-S1-428B2       | 400/ Candonga x 1806-1        | 10.68 | 7.03  | 751.96  | 1.70    | 1.82  | 1.76       |
| 215B2    | (Fontanilla x Candonga)-S1-215B2   | 200/ Fontanilla x Candonga    | 7.70  | 11.04 | 695.70  | 1.63    | 1.81  | 1.72       |
| 617B2    | (Candonga x 2846-5)-S1-617B2       | 600/ Candonga x 2846-5        | 9.03  | 10.34 | 651.16  | 1.60    | 1.71  | 1.65       |
| 606B1    | (Candonga x 2846-5)-S1-606B1       | 600/ Candonga x 2846-5        | 8.96  | 12.25 | 538.98  | 1.69    | 1.58  | 1.64       |
| 139B3    | (Amiga x 3846-5)-S1-139B3          | 100/ Amiga x 2846-5           | 11.48 | 6.63  | 835.05  | 1.61    | 1.60  | 1.61       |
| 481B4    | (Candonga x 1806-1)-S1-481B4       | 400/ Candonga x 1806-1        | 8.54  | 9.27  | 648.83  | 1.59    | 1.47  | 1.53       |
| 104B1    | (Amiga x 2846-5)-S1-104B1          | 100/ Amiga x 2846-5           | 7.50  | 8.70  | 661.16  | 1.27    | 1.39  | 1.33       |
| 326B2    | (Fuentepina x Fontanilla)-S1-326B2 | 300/ Fuentepina x Fontanilla  | 12.00 | 8.67  | 615.42  | 1.34    | 1.29  | 1.32       |
| 308B1    | (Fuentepina x Fontanilla)-S1-308B1 | 300/ Fuentepina x Fontanilla  | 8.10  | 8.52  | 672.99  | 1.27    | 1.35  | 1.31       |
| 630B4    | (Candonga x 2846-5)-S1-630B4       | 600/ Candonga x 2846-5        | 10.08 | 9.57  | 507.21  | 1.42    | 1.13  | 1.28       |

Note. BX = soluble solids content; FW = fruit weight; C-yield = commercial yield; SI\_BLUP = selection index based on BLUPs; SI\_ML = selection index based on machine learning predictions; SI\_BLUP\_ML = combined selection index. Series 100 = Amiga x 2846-5; Series 200 = Fontanilla x Candonga; Series 300 = Fuentepina x Fontanilla; Series 400 = Candonga x 1806-1; Series 500 = Santa Clara x Aguadilla; Series 600 = Candonga x 2846-5.

## DISCUSSION

The results obtained confirm that S1 populations of *Fragaria* × *ananassa* exhibit substantial segregation for productive and quality traits, consistent with the highly heterozygous genetic structure of the species and with patterns reported for early generations such as S1 in breeding programs (Chandler *et al.*, 2012; Zeist & Resende, 2019). The detection of significant differences among genotypes for yield, fruit number, fruit weight, and soluble solids content reflects the availability of exploitable genetic variability, an indispensable condition for achieving sustained genetic progress in subsequent selection cycles (Bernardo, 2020; Falconer & Mackay, 1996).

Because the evaluated population corresponded to S1 genotypes, the observed variability also reflects the genetic consequences of selfing in a highly heterozygous and allopolyploid species such as *Fragaria* × *ananassa* (Chandler *et al.*, 2012; Zeist & Resende, 2019). In this context, some individuals may express reduced vigor or unfavorable trait combinations associated with inbreeding depression, whereas others may retain

favorable allelic combinations for productivity and fruit quality. The integration of BLUPs and machine learning predictors helped identify S1 individuals that maintained superior performance despite this genetic burden, supporting their use as candidates for advanced evaluation and recombination in subsequent breeding cycles (Falconer & Mackay, 1996; Bernardo, 2020).

The moderate to high heritability observed for FW, CFP, and CP suggests that these traits are strongly controlled by genetic factors and therefore constitute priority targets for direct selection in strawberry, as documented in other breeding programs (Chandler *et al.*, 2012; Zeist & Resende, 2019). In particular, the relatively high heritability of C-yield supports the potential to obtain substantial gains through early selection if methodologies capable of properly separating genetic effects from environmental variation are applied.

Positive genetic correlations among CFP, fruit size, and yield confirm that these components represent structural determinants of productivity in strawberry, in agreement with previous studies highlighting the central role of these traits in yield architecture (Sturzeanu & Hera, 2019; Sturzeanu *et al.*, 2025). In contrast, negative correlations between soluble solids content and productive variables reveal a physiological antagonism between sugar accumulation and fruit growth, a phenomenon widely recognized in quantitative genetics and particularly relevant in fruit crops, where photoassimilate allocation generates trade-offs between organoleptic quality and total yield (Falconer & Mackay, 1996; Chandler *et al.*, 2012). This pattern justifies the need for multivariate selection approaches capable of balancing contrasting objectives and avoiding unidirectional progress that could compromise key commercial attributes.

The high predictive performance of XGBoost for commercial yield confirms the suitability of gradient boosting algorithms for modeling complex nonlinear relationships in agricultural systems, as previously reported for horticultural crops and cereals (Chen & Guestrin, 2016; Tong & Nikoloski, 2021). In strawberry, where phenotypic expression is strongly modulated by genotype  $\times$  environment interactions, these tools may offer substantial advantages over strictly linear models by capturing implicit interactions among vegetative, phenological, and productive traits.

The contrasting behavior of XGBoost and CatBoost may be associated with the structure of the response variables and the type of information available in the predictors. XGBoost showed high predictive ability for commercial yield, probably because this trait was strongly associated with measurable productive and structural variables, such as number of fruits, fruit weight, and fruit diameter. CatBoost, in contrast, offered the theoretical advantage of handling categorical predictors directly, including pedigree, treatment code, and control status; however, this advantage was insufficient to overcome the weak separability of sweetness classes. Therefore, the limited performance observed for BX classification seems to be more related to the biological complexity and environmental sensitivity of soluble solids accumulation than to the algorithmic structure alone (Chen & Guestrin, 2016; Prokhorenkova *et al.*, 2018).

In contrast, the poor performance obtained for soluble solids content confirms that predicting sensory quality traits remains one of the major challenges in applying machine learning to fruit crops, owing to their high sensitivity to microenvironmental factors, fruit physiological status, and agronomic management (Newerli-Guz *et al.*, 2023). Reformulating BX as a categorical variable improved trait interpretability and partially increased the predictive ability of classification models, consistent with previous reports showing that discretizing highly noisy variables favors supervised learning in complex agricultural contexts (James *et al.*, 2021; Apostolopoulos *et al.*, 2023). Nevertheless, the asymmetry observed between sensitivity and specificity suggests that considerable

phenotypic overlap between sweetness categories still exists, which could be mitigated by incorporating ecophysiological or microclimatic covariates, or even genomic information, in future evaluations.

The low predictive ability for BX may also be interpreted in relation to genotype  $\times$  environment interaction, even though the present evaluation was conducted at a single experimental site. Soluble solids content is highly sensitive to microenvironmental variation during fruit ripening, including radiation, temperature, water availability and source-sink dynamics. Because these traits were not explicitly included as predictors, the models probably captured only part of the biological variation associated with sugar accumulation. Future ML-assisted selection workflows could improve the prediction of fruit quality by incorporating real-time climatic variables, such as solar radiation, temperature, relative humidity, and precipitation, together with physiological indicators more directly related to carbon assimilation and sugar partitioning, such as specific leaf area, chlorophyll indices, or canopy vigor (James *et al.*, 2021; Apostolopoulos *et al.*, 2023).

The construction of selection indices based on BLUPs, machine learning predictions, and their combination enabled simultaneous integration of genetic merit and expected phenotypic performance, generating a more balanced ranking of the selected S1 genotypes. The strong association between the indices and commercial yield confirms that this trait remains the main driver of selection decisions within the program, whereas the moderate weight of soluble solids content within the indices reflects the detected genetic antagonism and the need for careful weighting when quality traits are included. This behavior agrees with classical selection index theory, whereby genetic correlations among traits condition the efficiency of joint response to selection (Falconer & Mackay, 1996; Bernardo, 2020).

Genotypes prioritized through the combined BLUP–ML index represent strategic candidates for advanced evaluation cycles, as they simultaneously concentrate high productivity and acceptable fruit quality levels. This result supports the potential of machine learning as a complementary tool to classical strawberry breeding, particularly at early stages where selection pressure must be exerted on large segregating populations with limited experimental resources.

Overall, the findings of this study demonstrate that integrating mixed models, machine learning algorithms, and multivariate selection schemes constitutes a promising avenue for accelerating genetic progress in strawberry, enabling more efficient management of yield–quality trade-offs and improved decision-making in modern breeding programs.

## CONCLUSIONS

S1 strawberry populations exhibited high genetic variability and moderate to high heritabilities for key traits, with positive correlations among yield components and antagonistic relationships with soluble solids content, justifying the application of multivariate selection strategies. Machine learning models—particularly XGBoost and CatBoost—predicted commercial yield with high accuracy but showed limitations for soluble solids content, highlighting a restriction for selection based on organoleptic quality. Integrating BLUPs and machine learning predictions into a combined index enabled prioritization of genotypes with balanced productivity and quality, identifying promising genetic materials for advanced stages of the breeding program. However, these selected materials should be evaluated across environments and production cycles to confirm their genetic stability and fruit quality performance. This methodology can contribute to reducing field space, evaluation time, and operational costs, thereby improving the efficiency of early selection in strawberry breeding programs.

## AUTHOR CONTRIBUTIONS

Conceptualization: T.C.L.B.; Methodology: L.K.L.S.; Research: L.K.L.S., D.E.D.A and T.C.L.B.; Validation: D.E.D.A; Writing: preparation of the original draft: L.K.L.S., D.E.D.A and T.C.L.B.; Writing: review and editing: L.K.L.S., D.E.D.A and T.C.L.B.; Funding Acquisition: D.E.D.A.

## ACKNOWLEDGMENTS

The authors acknowledge the Sistema de Investigaciones of the Vicerrectoría de Investigaciones e Interacción Social of the Universidad de Nariño for funding the research that led to this scientific article.

## FUNDING

This study was funded by the Universidad de Nariño. The authors express their gratitude for providing the financial support and the facilities required to conduct this research.

## CONFLICT OF INTEREST

The authors declare that there are no financial, commercial, or personal conflicts of interest that could have inappropriately influenced the conduct of the study, the interpretation of the results, or the preparation of the manuscript.

## REFERENCES

- Acuña Caita, J. F., & Fischer Gebauer, G. (2024). Fresa (*Fragaria × ananassa* Duch.): manual de recomendaciones técnicas para su cultivo en el departamento de Cundinamarca. <https://repositorio.unal.edu.co/handle/unal/86761>
- Red de información y comunicación del sector agropecuario colombiano (AGRONET) (2025). Reporte: Área, Producción y Rendimiento Nacional por Cultivo Fresa. <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=1>
- Al-Taai, S. R., Azize, N. M., Thoeny, Z. A., Imran, H., Bernardo, L. F. A., & Al-Khafaji, Z. (2023). XGBoost Prediction Model Optimized with Bayesian for the Compressive Strength of Eco-Friendly Concrete Containing Ground Granulated Blast Furnace Slag and Recycled Coarse Aggregate. *Applied Sciences*, 13(15), 8889. <https://doi.org/10.3390/app13158889>
- Altieri, G., Curcio, D., Lepore, A., Grobler, E., Maffia, A., Gargano, N., Tedesco, A., Graziano, M. L., Mazzei, P., Capocasa, F., Mezzetti, B., & Celano, G. (2025). Yield and quality of new strawberry advanced breeding selections and commercial cultivars, grown under warm-temperate climatic conditions. *Agriculture*, 15(13), 1406. <https://doi.org/10.3390/agriculture15131406>
- Apostolopoulos, I. D., Tzani, M., & Aznaouridis, S. I. (2023). A General machine learning model for assessing fruit quality using deep image features. *AI*, 4, 812–830. <https://doi.org/10.3390/ai4040041>
- Brasileiro, B. P., Marinho, C. D., Costa, P. M. de A., Cruz, C. D., Peterlini, L. A. & Barbosa, M. H. P. (2015). Selection in sugarcane families with artificial neural networks. *Crop Breeding and Applied Biotechnology*, 15(2), 72–78. <https://doi.org/10.1590/1984-70332015v15n2a14>
- Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity*, 125, 375–385. <https://doi.org/10.1038/s41437-020-0312-1>
- Chandler, C. K., Foltz, K., Dale, A., Whitaker, V. M., & Herrington, M. (2012). Strawberry. In: M., Badenes, D. Byrne, (eds) Fruit Breeding. Handbook of Plant Breeding, vol 8. Boston: Springer. [https://doi.org/10.1007/978-1-4419-0763-9\\_9](https://doi.org/10.1007/978-1-4419-0763-9_9)
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002) SMOTE: Synthetic Minority Over-sampling

- Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T. Q., & Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Chen, Q., Li, J., Feng, J., & Qian, J. (2024). Dynamic comprehensive quality assessment of post-harvest grape in different transportation chains using SAHP–CatBoost machine learning, *Food Quality and Safety*, 8, 1-11. <https://doi.org/10.1093/fqsafe/fyae007>
- Falconer, D. & Mackay, T. (1996). *Introduction to quantitative genetics*, 4th edition. New Jersey: Prentice Hall. 464p.
- Federer, W. T., & Raghavarao, D. (1975). On augmented designs. *Biometrics*, 31(1), 29-35. <https://doi.org/10.2307/2529707>
- Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Gutiérrez-Esquivel, D., & Ruiz-Rivas, M. (2025). Cultivares híbridos como herramienta de mejoramiento genético en fresa: Híbridos en el mejoramiento genético de fresa. *Revista C+TEC*. 4, 30-35. <https://dialnet.unirioja.es/servlet/articulo?codigo=10304665>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*, (2nd ed.). New York: Springer. 607p.
- Jones, D., Fornarelli, R., Derbyshire, M., Gibberd, M., Barker, K., & Hane, J. (2023). The pursuit of genetic gain in agricultural crops through the application of machine learning to genomic prediction. *Frontiers in Genetics*, 14, 1186782. <https://doi.org/10.3389/fgene.2023.1186782>
- Khatibi, S. M. H., & Ali, J. (2024) Harnessing the power of machine learning for crop improvement and sustainable production. *Frontiers in Plant Science*, 15, 1417912. <https://doi.org/10.3389/fpls.2024.1417912>
- Lin, F., Chen, D., Liu, C., & He, J. (2024). Non-Destructive detection of golden passion fruit quality based on dielectric characteristics. *Applied Sciences*, 14(5), 2200. <https://doi.org/10.3390/app14052200>
- M'hamdi, O., Takács, S., Palotás, G., Ilahy, R., Helyes, L., & Pék, Z. A. (2024). A Comparative analysis of XGBoost and neural network models for predicting some tomato fruit quality traits from environmental and meteorological data. *Plants*, 13(5), 746. <https://doi.org/10.3390/plants13050746>
- Newerli-Guz, J., Śmiechowska, M., Drzewiecka, A., & Tylingo, R. (2023). Bioactive ingredients with health-promoting properties of strawberry fruit (*Fragaria x ananassa* Duchesne). *Molecules*, 28(6), 2711. <https://doi.org/10.3390/molecules28062711>
- Powers, D. M. W. (2011) Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2, 37-63. <https://doi.org/10.48550/arXiv.2010.16061>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased Boosting with Categorical Features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, 3-8, 6639-6649. [https://papers.nips.cc/paper\\_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf](https://papers.nips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf)
- Şimşek, Ö. (2024). Machine learning offers insights into the impact of in vitro drought stress on strawberry cultivars. *Agriculture*, 14(2), 294. <https://doi.org/10.3390/agriculture14020294>
- Sturzeanu, M., Hera, O., Militaru, M., & Vîjan, L. E. (2025). Improving strawberry fruit quality through breeding: cultivar performance and biochemical diversity. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, 53(3), 14704. <https://doi.org/10.15835/nbha53314704>
- Sturzeanu, M., & Hera, O. (2019). Variability and heritability for yield and fruit weigh traits of F1 strawberry hybrids. *Fruit Growing Research*, 35, 13-16. <https://doi.org/10.33045/fgr.v35.2019.02>
- Takahashi, M., Kawasaki, Y., Naito, H., Lee, U., & Yoshi, K. (2025) Fruit size prediction of tomato cultivars using machine learning algorithms. *Frontiers in Plant Science*, 16, 1516255. <https://doi.org/10.3389/fpls.2025.1516255>
- Tong, H., & Nikoloski, Z. (2021). Machine Learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of Plant Physiology*, 257, 153354. <https://doi.org/10.1016/j.jplph.2020.153354>
- Vieira, S. D., de Souza, D. C., Martins, I. A., Ribeiro, G. H. M. R., Resende, L. V., Ferraz, A. K. L., Galvão, A. G., & de Resende, J. T. V. (2017). Selection of experimental strawberry (*Fragaria x ananassa*) hybrids based on selection indices. *Genetics and Molecular Research*, 16(1), gmr16019052. <http://dx.doi.org/10.4238/gmr16019052>
- Zareei, E., Karami, F., Aryal, R., & Saed-Moucheshi, A. (2023). Genotypic by phenotypic interaction affects the heritability and relationship among quantity and quality traits of strawberry (*Fragaria x ananassa*). *New Zealand Journal of Crop and Horticultural Science*, 51(4), 594–613. <https://doi.org/10.1080/01140671.2022.2039725>
- Zeist, A. R., & Resende, J. T. V. (2019). Strawberry breeding in Brazil: current momentum and perspectives. *Horticultura Brasileira*, 37, 007-016. <http://dx.doi.org/10.1590/S0102-053620190101>
- Zystro, J., Colley, M., & Dawson, J. (2018). Alternative Experimental Designs for Plant Breeding. *Plant Breeding Reviews*, 42, 87–117. <https://doi.org/10.1002/9781119521358.ch3>