REVISTA BRASILEIRA DE FRUTICULTURA
Published by the Brazilian Society of Fruit Crops
ISSN 0100-2945

Economy - Original Article - Edited by: Alessandro Dal´Col Lúcio

Software for classification of banana ripening stage using machine learning

Angela Vacaro de Souza[1]*, Jéssica Marques de Mello[2], Vitória Ferreira da Silva Favaro[1], Fernando Ferrari Putti[1]

[1] São Paulo State University (UNESP), School of Science and Engineering, Tupã, SP, Brazil.
[2] São Paulo State University (UNESP), Institute of Science and Technology, Sorocaba, SP, Brazil.
*Corresponding author: *angela.souza@unesp.br*

Abstract: Pattern recognition aims to classify some datasets into specific classes or clusters, having several applications in agriculture. The objectification of the process minimizes errors since it reduces subjectivity, allowing a fairer remuneration to the producer and standardized products to the consumer. Thus, this work aimed to develop an embedded system with artificial intelligence to determine the ripening stage of bananas (outputs) from the insertion of physical (i.e., fruit weight, texture and diameter), physicochemical (i.e., pH, titratable acidity (TA), soluble solids (SS) and SS/TA ratio) and biochemical (i.e., total sugars, phenolic compounds, ascorbic acid, quantification of pigments in fruit peel and pulp and antioxidant activity by DPPH and FRAP methods) data (inputs). The bananas were harvested at each evaluated stage according to the Von Loesecke ripening scale, as follows: stage 2, totally green; stage 4, more yellow than green; stage 6, yellow; and stage 7, yellow with brown spots. Subsequently, they were selected and submitted to quality analysis. The data obtained were then mined and the attributes were selected using WEKA software. The classifier software was developed using MATLAB. The most relevant attributes selected in the Bayes Net classifier for the Cross-Validation method were: apical, central, basal and mean textures (between apical, median and basal textures), pH, soluble solids, phenolic compounds, antioxidant activities by the FRAP and DPPH methods, vitamin C, anthocyanins from the pulp, chlorophyll a content in the fruit peel and sugar, resulting in a mean F-measure of 97.0%.

Index terms: Agriculture 4.0, vegetable sorting, data mining, *Musa* spp., post-harvest.

Software para classificação do estado de maturação da banana utilizando aprendizado de máquina

Resumo: O reconhecimento de padrões tem como objetivo classificar alguns conjuntos de dados em classes ou *clusters* específicos, tendo várias aplicações na agri-

Rev. Bras. Frutic., v.46, e-863 DOI: *https://dx.doi.org/10.1590/0100-29452024863*
Received 22 Aug, 2023 • Accepted 02 Feb, 2024 • Published May/Jun, 2024. Jaboticabal - SP - Brazil.

cultura. A objetivação do processo minimiza erros, pois reduz a subjetividade, permitindo uma remuneração mais justa ao produtor e produtos padronizados ao consumidor. Assim, este trabalho teve como objetivo desenvolver um sistema embarcado com inteligência artificial para determinar o estádio de maturação de bananas (*outputs*) a partir da inserção de dados físicos (peso do fruto, textura e diâmetro), físico-químicos (pH, acidez titulável (AT), sólidos solúveis (SS) e relação SS/TA) e bioquímicos (açúcares totais, compostos fenólicos, ácido ascórbico, quantificação de pigmentos na casca e na polpa dos frutos e atividade antioxidante pelos métodos DPPH e FRAP) (*inputs*). As bananas foram colhidas em cada estádio avaliado, de acordo com a escala de maturação de Von Loesecke, a saber: estádio 2, totalmente verde; estádio 4, mais amarela que verde; estádio 6, amarela; e estádio 7, amarela com manchas marrons. Posteriormente, foram selecionados e submetidos a uma análise de qualidade. Os dados obtidos foram então minerados, e os atributos foram selecionados, utilizando o software WEKA. O software classificador foi desenvolvido em MATLAB. Os atributos mais relevantes selecionados no classificador *Bayes Net*, para o método de *Cross-Validation*, foram: texturas apical, central, basal e média (entre as texturas apical, mediana e basal), pH, sólidos solúveis, compostos fenólicos, atividades antioxidantes, pelos métodos FRAP e DPPH, vitamina C, antocianinas da polpa, teor de clorofila a na casca do fruto e açúcar, resultando em uma medida F média de 97,0%.

Termos para indexação: Agricultura 4.0, triagem de vegetais, mineração de dados, *Musa* spp., pós-colheita.

Introduction

The banana tree (*Musa sp* L.) belongs to the Musaceae family and is one of the most important fruit in the world, being appreciated by people of all social classes and ages. As it is a climacteric fruit, the banana changes its organoleptic characteristics of color, flavor, aroma and nutritional parameters throughout the ripening period. For this reason, the stage at which it is harvested is decisive for its storage, commercialization and pricing. The criteria used by producers to predict the moment of harvest are empirical and based on some morphological aspects, such as the disappearance of corners and angles on the fruit surface or the fruit caliber measurement, the physiological degree related to the fruit ripening according to its color, or the distinction of the fruit by age through markings on the plantation based on age difference or bunch bagging date (ALVES, et al., 2004).

The fruit ripening stage classification, as well as its automation, is extremely important, as it prevents products with serious defects such as immaturity or early ripening from reaching the consumer's table – since the quality of a product also implies the quality of life of the people who consume it. By transforming a subjective and slow method into an objective and agile one, errors can be mitigated, causing the producer to be remunerated more precisely according to the quality of their products. In addition, in the trade and pricing of vegetables, these systems can assist in the labeling phase, placing each fruit in its specific class.

To automate the process and classify a product in relation to different quality parameters, sorting is performed using a set of attributes. In order to distinguish the cases among the possible classifications, each one is labeled with a special attribute called 'class', whose values refer to the true classification of cases. The labeled cases are termed 'examples', whereas the sample used by the learning algorithm to induce the classification model is called 'set of training examples' (PRATI et al., 2008). For a more accurate classification, it is necessary to carry out learning processes that can be elaborated by supervised or unsupervised methods, which end up helping the decision-making. To verify the quality of classifications, specific measurements such as the Kappa coefficient, or the general error probability, must be used. The Kappa index is a measure of agreement used in nominal scales that provides an idea of how much the observations deviate from the expected due to chance, thus indicating how legitimate the information is (LOBÃO, 2005; WITTEN and FRANK, 2005). This index is intended to measure the degree of agreement between proportions derived from dependent samples.

Supervised machine learning has several ways to evaluate the performance of learning algorithms and the classifiers they produce (SOKOLOVA et al., 2006). Thus, classification quality metrics are created through a confusion matrix that registers the correctly and incorrectly reorganized examples of each class, that is, a table or matrix that records the results previously mentioned in the text. Supervised learning uses a set of information already described or known where certain outputs are expected in relation to the inputs.

To measure the performance of the learning algorithm in machine learning, there are several acceptable metrics. In this specific work, we used the F-measure, which according to Sokolova et al. (2006) consists of a type of combined metric that benefits algorithms with a higher sensitivity and challenges them so as to measure classification and learning quality. The focus of this work was to develop a software system capable of classifying the ripening stage of bananas elaborated through supervised machine learning techniques using destructive and non-destructive parameters to evaluate 'Nanicão' banana fruits at four different ripening stages. To validate the performance of the classifier, the following methods were used via WEKA software (which has several machine learning algorithms and the necessary tools for data preparation and mining): the resubstitution method or training set (leave-one-out); the sample division method or holdout or percentage split; and the cross-validation method or k-fold. After data mining and attribute selection with the aid of WEKA, the software for the classification of bananas was developed using MATLAB.

Materials and Methods

Experimental design and vegetal material

The bananas were obtained from a commercial area in the municipality of Iacri, São Paulo, Brazil.

They were classified according to a scale proposed by Von Loeseck (1950), being divided into the following ripening stages: stage 2 (more green than yellow); stage 4 (more yellow than green); stage 6 (yellow); and stage 7 (yellow with brown spots), (Figures 1a-d, respetively). From the observation of the existing analogies between the them, the bananas were divided into different classes for data collection in terms of quality and further use for the development of the classifier software. A total of 200 bananas were used, 50 at each ripening stage.

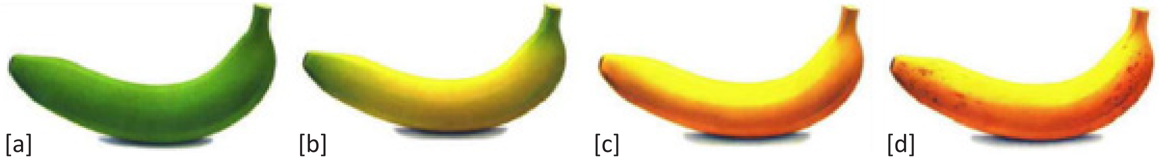


Figure 1: Von Loeseck ripening scale, 1950. Source: Adapted from PBMH & PIF (2006).

The non-destructive and physical-chemical analyses were performed on the day of the experiment setup. For the biochemical analyses, the bananas were macerated in liquid N_2 and subsequently frozen. The fruit analyses were carried out at the Biology and Chemistry Laboratories of the Faculty of Sciences and Engineering of Unesp, Tupã, SP, Brazil, with the following determinations being made:

Physical, chemical, physicochemical and biochemical analyses

Physical analyses: *fruit weight,* performed on a scale, with data expressed in grams (g); *fruit average length and diameter,* determined in centimeters, with the aid of a caliper in the middle region of the fruit; and *fruit texture (firmness),* determined at 3 points, i.e., in the central, apical and basal regions of the fruit, with the aid of a texturometer/digital penetrometer (VICTOR – GY4), at a penetration distance of 10 mm, using a straight circular tip with a diameter of 3.50 mm, with the value obtained in N. At the end, the mean between the apical, central and basal textures was performed, resulting in one more measurement.

Physical-chemical and chemical analyses: *pH (hydrogen potential),* measured in the crushed fruit pulp using a potentiometer (Digital DMPH-2) (IAL, 2008); *titratable acidity (TA),* expressed in grams of citric acid per 100 g of pulp (g of citric acid 100 g-1) and obtained by titration of about 5 g of homogenized and diluted pulp to 100 ml of distilled water with a standardized solution of 0.1 N NaOH using phenolphthalein as an indicator, which occurs when the potentiometer reaches 8.1 (IAL, 2008); *soluble solids (SS),* determined by refractometry on an Atago PR–32 Palette digital refractometer with automatic temperature compensation (IAL, 2008) and expressed in °BRIX; and *SS/TA ratio,* calculated from the ratio between soluble solids and titratable acidity content and expressed as dimensionless values.

Biochemical analyses: For all biochemical analyses, the number of samples required, the extractor and the extractor concentration were standardized, preferably using Sigma-Aldrich reagents (New South Wales, Australia). All samples were homogenized with IKA T 65 basic ULTRA-TURRAX®. Quimis ultrasonic bath model 0335 D version 1.0 was used in some analyses, while SHIMADZU UV-1800 spectrophotometer was employed in some others that required spectrophotometric readings.

Total vitamin C (Vit C) was assessed by titration of ascorbic acid and isomers with iodine solution. For ascorbic acid extraction, 1g of banana was weighed and diluted with the aid of Turax in 50 ml of recently boiled and cooled distilled water. Then, 12.50 ml of 1 M sulfuric acid and 3 ml of 1% starch (used as an indicator) were added. Subsequently, titration with iodine was performed at 0.0005 M and the results were expressed in mg of ascorbic acid per 100 g-1 of fresh mass (IAL, 2008). *Phenolic compounds (PC)* were determined according to the spectrophotometric method and using Folin Ciocalteu reagent (SINGLETON; ROSSI JR., 1965); the results were expressed in mg of equivalent gallic acid per 100 g-1 fresh mass. *Pigments* were quantified according to the methodology adapted from Sims and Gamon (2002) by spectrophotometry through the maceration of the material in liquid nitrogen followed by the addition of Tris-HCL buffer acetone and centrifugation at 6000 rpm for 5 minutes. *Total antioxidant activity* was determined against free radical (DPPH) following the method proposed by Brand-Williams et al. (1995) with some modifications and performing absorbance reading at 517 ηm, and the results were expressed in reduced DPPH %. *Antioxidant activity* (FRAP – Ferric Reducing Antioxidant Power) was assessed according to the method described by Benzie and Strain (1996) and the results were expressed in mmol Fe kg-1. Lastly, *total sugar* was determined based on the dehydration of sugars in an acid medium with concentrated sulfuric acid and the subsequent complex formation with phenol (DUBOIS et al., 1956) followed by spectrophotometric reading at 490 nm.

Data analysis: After the laboratory analysis for the quality of the bananas at different ripening stages, the data were organized in an Attribute-Relation File Format (ARFF) with its own characteristics, containing the attribute domain, the attribute values and the attribute "class". As a starting point, the data used were non-destructive, such as fruit weight, diameter, length, basal, central, apical and median textures, and physical-chemical, such as pH, soluble solids, titratable acidity and SS/TA ratio. These data were chosen because they are simpler to obtain and do not depend on sophisticated equipment and methodologies.

The data obtained were converted to ".arff', the format used in *WEKA*, through an algorithm made in MATLAB called "ArffWriter". In total, 12 attributes were used, with the last one indicating the class of the object. The values ranged from 0 to 3 to represent the fruit ripening stages (or classes), that is, 0 for green, 1 for more green than yellow, 2 for yellow, and 3 for yellow with brown spots.

Classification tests were carried out with different classifiers and techniques on WEKA software. For the classification of these four stages, the classifiers tested were the Bayesian Bayes Net and Naive Bayes, the function algorithms Multilayer Perceptron and SMO (SVM), the lazy classifier IBK (KNN), the classification algorithm OneR (One Rule) and the decision trees J48 and Random Forest.

In addition to F-measure data, WEKA provides the Kappa coefficient (κ). This index is considered a measure of interobserver agreement that allows assessing both whether the agreement is beyond what is expected by chance and the degree of this agreement (SILVA; PAES, 2019). Landis and Koch (1977) suggest the following interpretation for the Kappa coefficient (Table 1):

Table 1: Kappa coefficient values according to Landis and Koch (1977).

Kappa values	Interpretation
<0	No agreement (very bad)
0–0.19	Slight agreement (bad)
0.20–0.39	Fair agreement (reasonable)
0.40–0.59	Moderate agreement (good)
0.60–0.79	Substantial agreement (very good)
0.80–1.00	Almost perfect agreement (excellent)

According to this index, if all classifications are made correctly the Kappa coefficient, κ, will be κ = 1; otherwise, if all observations are classified in the same class, then the value of κ will be κ = 0. Therefore, the value of κ will always be $0 \leq \kappa \leq 1$.

The results obtained were submitted to the Tukey test at $P \leq 0.05$, following the previous work by Souza et al. (2021).

Results and Discussion

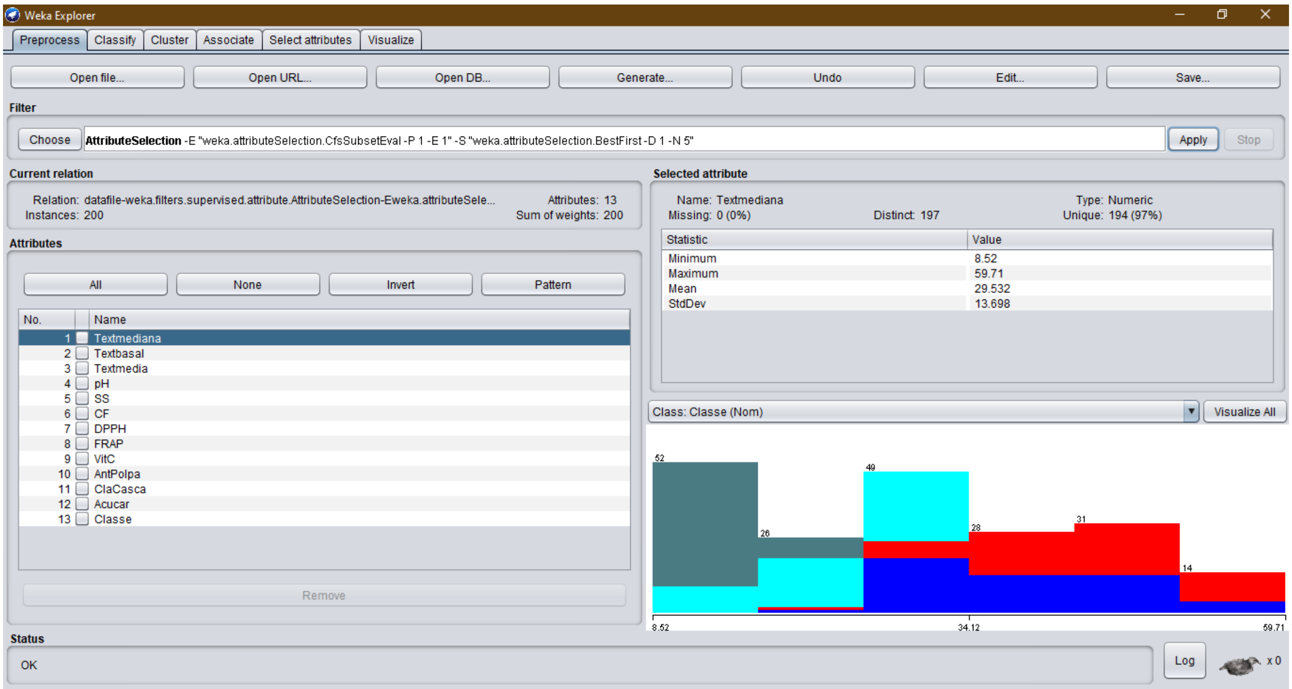
Table 2 shows the mean F-measure values for the dataset used in this work. For the classification test, the algorithms used were: Bayes Net, IBK, J48, Multiplayer Perceptron, Naive Bayes, One R, Random Forest and SMO.

Table 2: Mean F-measure values for the database with 25 attributes for the training set cross-validation and percentage split (66%) test methods using different classifiers available in WEKA software.

Classifier	Mean F-measure for each test method		
	Training set	Cross-validation	*Percentage split* (66%)
Bayes Net	0.990	0.955	0.871
IBK (KNN=1)	1.000	0.955	0.913
J48	0.975	0.910	0.870
Multilayer Perceptron	1.000	0.975	0.942
Naive Bayes	0.920	0.905	0.811
OneR	0.745	0.594	0.563
Random Forest	1.000	0.975	0.928
SMO	0.965	0.960	0.911

As observed in Table 2, Multiplayer Perceptron and Radom Forest reached the highest F-measure values for the three classification methods. In the data normalization process, the results obtained were the same as the non-normalized ones. In order to improve the performance of the classifiers, attributes that were redundant and irrelevant were removed. To this end, the WEKA's attribute selection tool, which utilizes the supervised filter *Cfs*, was used with all of its parameters and default software values. The result was the maintenance of the attributes central, basal and median textures, pH and soluble solids (Figure 2).

The graph in the lower right corner of Figure 2a can be better visualized in Figure 2b, which shows the number of objects in each class (0 - green, 1 - more green than yellow, 6 - yellow and 7 - yellow with brown spots), based on the selected attributes. In this graph, class 0 (green) is represented by dark green color, class 1 (more green than yellow) by light blue color, class 2 (yellow) by blue color, and class 3 (yellow with brown spots) by red color. From this graph it is possible to observe that despite being classified in a specific class, some objects have characteristics that are similar to another class. To avoid redundancy, attributes that had a Kappa coefficient equal to 0.5 were selected for most classifiers and test methods under evaluation. Annex A contains the Kappa coefficients of the attributes not selected by WEKA.



Filter

AttributeSelection -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Current relation

Relation: datafile-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSele... Attributes: 13
Instances: 200 Sum of weights: 200

Attributes

No.		Name
1	□	Textmediana
2	□	Textbasal
3	□	Textmedia
4	□	pH
5	□	SS
6	□	CF
7	□	DPPH
8	□	FRAP
9	□	VitC
10	□	AntPolpa
11	□	ClaCasca
12	□	Acucar
13	□	Classe

Selected attribute

Name: Textmediana Type: Numeric
Missing: 0 (0%) Distinct: 197 Unique: 194 (97%)

Statistic	Value
Minimum	8.52
Maximum	59.71
Mean	29.532
StdDev	13.698

Class: Classe (Nom)

Status

OK

Figure 2. Attribute selection in WEKA with the supervised filter *Cfs* (a) and classification breakdown (b). Subtitle: Textmediana = Median Texture; Textbasal = Basal Texture; textmedia = Average Texture; CF = PC; AntPolpa = Anthocyanin Pulp; ClaCasca = Chlorophyll a Peel; Açúcar = Sugar; Classe = Class.

The classification tests for this dataset tested thirteen attributes (i.e., apical, central, basal and median textures, pH, soluble solids, phenolic compounds, FRAP, DPPH, vitamin C, anthocyanins from the pulp, chlorophyll a content in the fruit peel and sugar) using the same classifiers and test methods. The results are displayed in Table 3.

Moreti (2020) classified different red fruits using the Multilayer Perceptron, IBK, Naive Bayes, Random Forest, SMO and J48 algorithms for the attributes color, shape and texture, obtaining accuracy results for coloring close to 85%, 100%, 85%, 97%, 85% and 92%, respectively. According to the author, for color and shape the algorithm that showed the best results was IBK, while for texture the most accurate was Random Forest. This work evidences the importance of employing different algorithms and parameters in tests (Table 3).

Table 3: Mean F-measure values for the database with 13 attributes defined after the attribute selection process for the training set, cross-validation and percentage split (66%) test methods using the different classifiers available in WEKA software 3.8.3

Classifier	**Mean F-measure for each test method**		
	Training set	**Cross-validation**	***Percentage split* (66%)**
Bayes Net	0.980	0.970	0.898
IBK (KNN=1)	1.000	0.940	0.883
J48	0.965	0.925	0.870
Multilayer Perceptron	1.000	0.955	0.971
Naive Bayes	0.955	0.940	0.870
OneR	0.745	0.594	0.563
Random Forest	1.000	0.960	0.882
SMO	0.955	0.930	0.912

As it can be seen, the Multilayer Perceptron and Random Forest were the classifiers with the highest values for the methods tested, while OneR was the algorithm that obtained the lowest values.

Figure 3 shows a comparison of the mean F-measure values for the two databases tested using the training set method.

According to Figure 4, the database containing all attributes and that containing only the attributes considered relevant by WEKA showed similar results, indicating that all 25 attributes collected are not necessary for determining the fruit class.

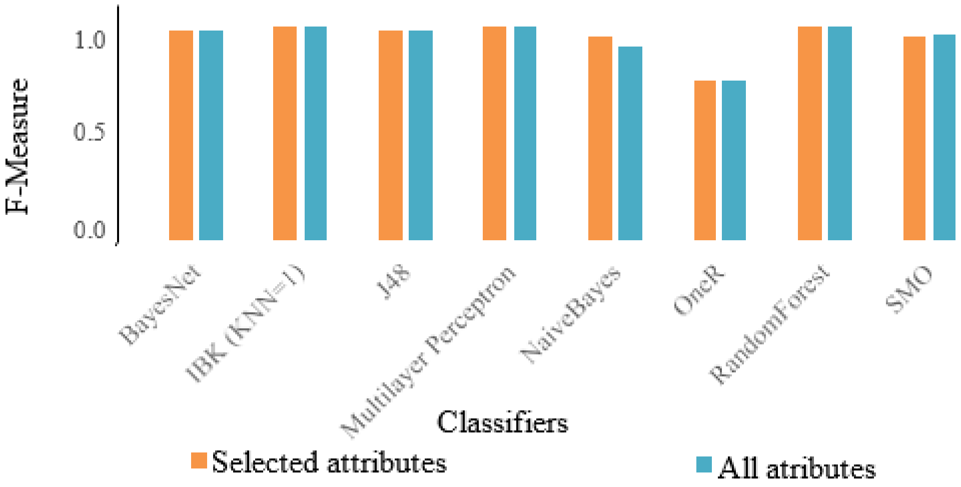


Figure 3: Mean F-measure of the classifiers for the databases tested using the training set method

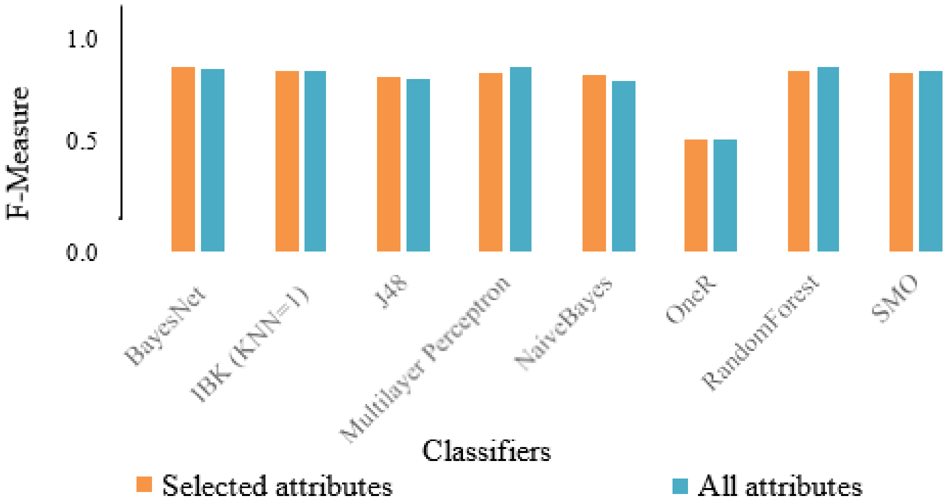


Figure 4: Mean F-measure of the classifiers for the databases tested using the cross-validation method

Arivazhagan et al. (2010) proposed an efficient combination between the attributes color and texture for 14 different species of fruits, with a total of 2633 images obtained, reaching recognition rates of 86% for color and texture combination against 45% for color and 70% for texture. Seng and Mirisaee (2009), on the other hand, developed a fruit recognition system based on the analysis of the fruit color, shape and size. The algorithm used for the classification was KNN (K-nearest neighbor), which measures the distance between the attributes obtained from the unknown object and the database using 50 fruit images, of which 36 are images for training and the rest for testing. The recognition accuracy of the system was 90%. It is noteworthy that both experiments used different numbers of vegetables, attributes and images.

Santos et al. (2020) conducted a study from images of fruits taking into account their color, shape and texture, as well as the combination of all computed vector resources, to compare the spaces of calculated resources using the classifiers MultiLayer Perceptron (MLP), KNN and Naive Bayes. At the end of the experiment, it was verified that the best classifier was Naive Bayes, which provided a combination of color and texture characteristics with a correct classification rate of approximately 71%.

Similar to the training set method, both databases exhibited very close results, demonstrating that the selected attributes are efficient for discretizing classes. Figure 5 shows a comparison of the mean F-Measure values for the two databases tested using the percentage split method (66%).

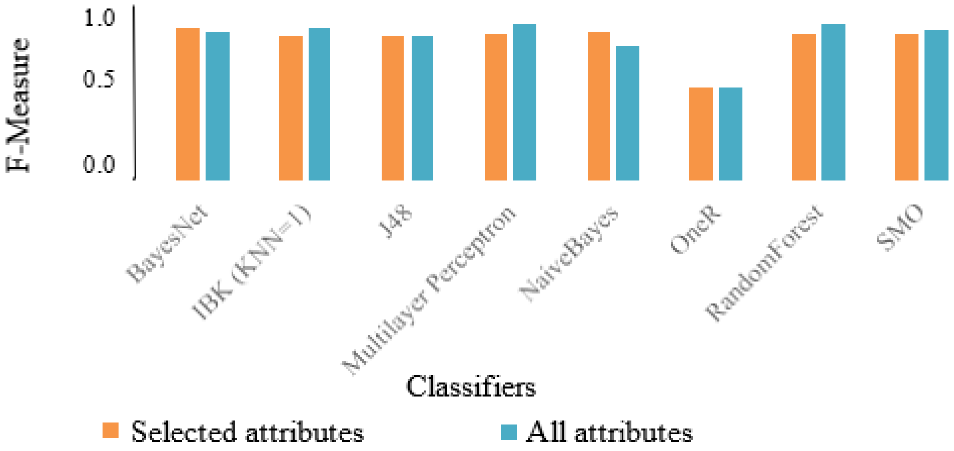


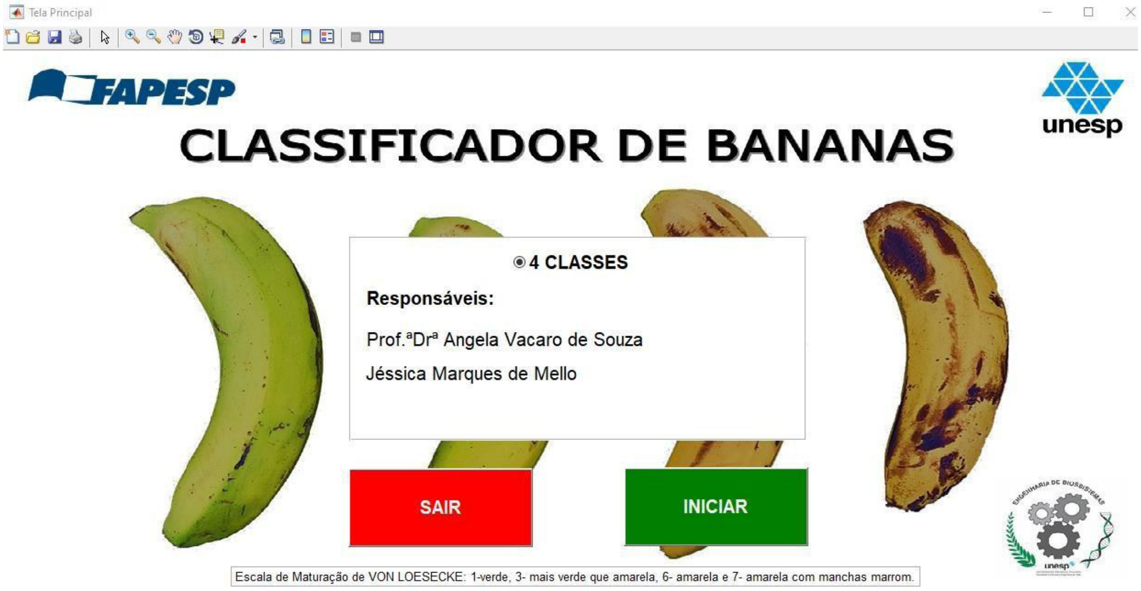
Figure 5: Mean F-measure of the classifiers for the databases tested using the percentage split (66%) method

According to Figure 5, it is possible to verify that the results for the percentage split method were similar to the other methods, that is, close results for both databases. Thus, it can be concluded that the database containing only the relevant attributes had a good performance, which means that it would not be necessary to measure all attributes to classify the fruits into different classes. Regarding the classifiers, the ones that obtained the best results were IBK, Multilayer Perceptron, Random Forest and Bayes Net.

It was possible to note that, in general, the combination of different parameter extraction techniques lead to better results than a single technique. In addition, different classifiers have different performance even when working with the same database.

As observed in Table 3, the best banana classifier among the four classes with attribute selection considering the validation set of cross-validation was Bayes Net, with an F-measure of 97%, being therefore chosen as the ideal classifier for this purpose.

After determining the best classifier for the evaluated parameters, a graphical interface was developed in MATLAB 9.0 environment. Figure 6 shows the graphical interface when the software is started.



CLASSIFICADOR DE BANANAS

◉ 4 CLASSES

Responsáveis:

Prof.ªDrª Angela Vacaro de Souza

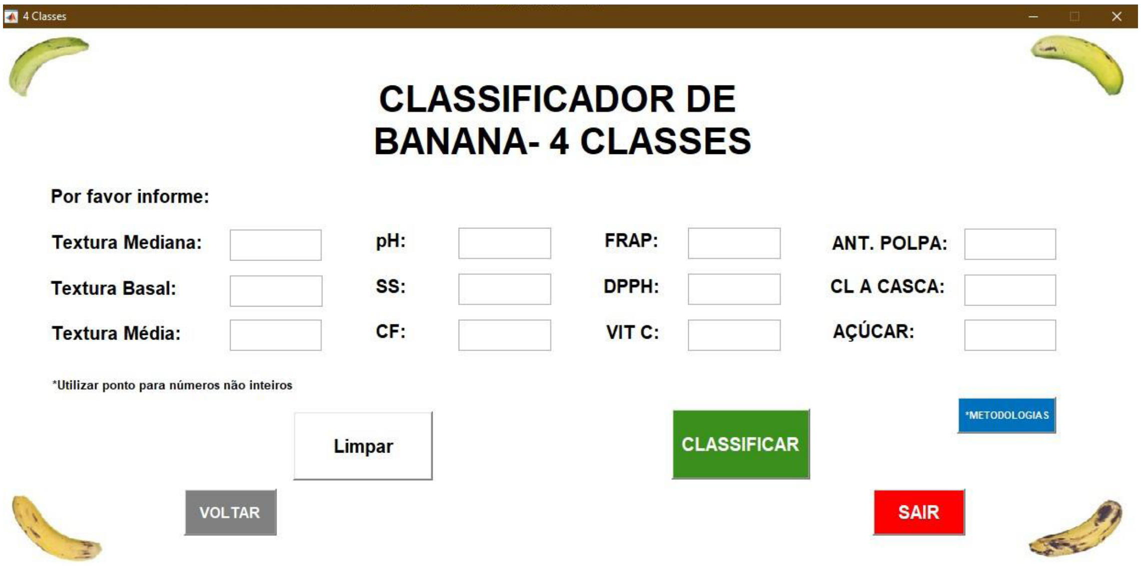
Jéssica Marques de Mello

Escala de Maturação de VON LOESECKE: 1-verde, 3- mais verde que amarela, 6- amarela e 7- amarela com manchas marrom.

Figure 6: Interface of the software home page

Subtitle: Banana Classifier 4 classes; Responsible: Prof. Angela Vacaro de Souza, PhD; Exit / Start; Von Loesecke ripening scale 1 - green 3 - more green than yellow 6 - yellow and 7 yellow with brown spots.

The software home page presents the main information of the classifier, such as the classes that were used from the Von Loesecke scale. Figure 7 shows the screen that appears when the "start" button is clicked. The screen illustrated below displays the data to be classified, in addition to the buttons "clear", which has the function of deleting the values inserted in all text boxes; "return", which returns to the screen shown in Figure 7; "exit", which ends the software execution; "classify", which informs the class in which the banana is, and "methodologies", which presents the methods used to extract each attribute.



CLASSIFICADOR DE BANANA- 4 CLASSES

Por favor informe:

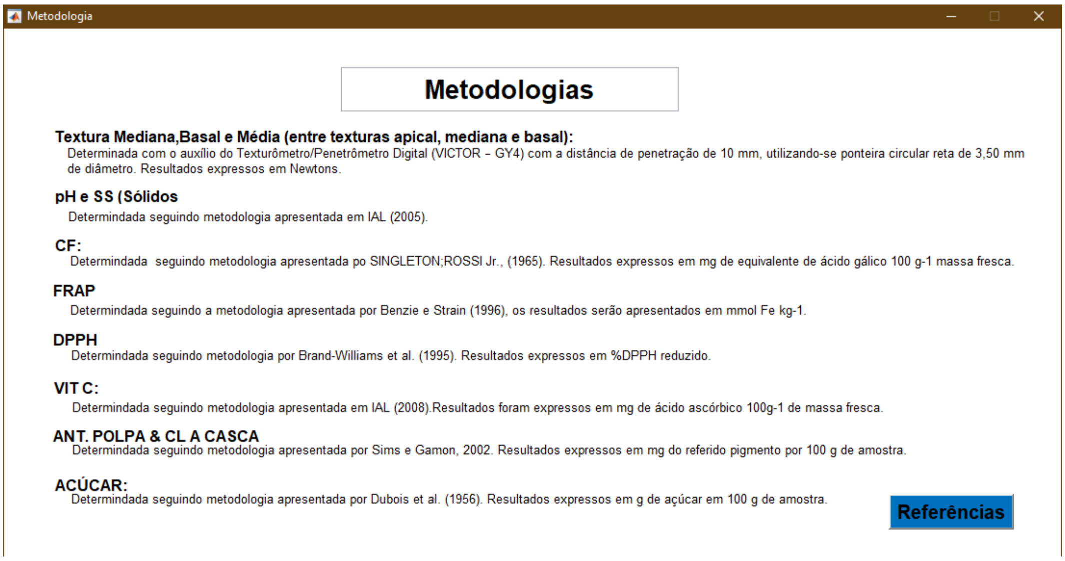
Textura Mediana:		pH:		FRAP:		ANT. POLPA:	
Textura Basal:		SS:		DPPH:		CL A CASCA:	
Textura Média:		CF:		VIT C:		AÇÚCAR:	

*Utilizar ponto para números não inteiros

Figure 7: Interface of the addition of data to be classified

Subtitle: Classifier of bananas - 4 classes; Please insert; median texture, basal texture, mean texture, pH, SS (Soluble Solids), PC (Phenolic Compounds), FRAP, DPPH, VIT C, Anthocyanin. Pulp, Chlorophyll a peel, Sugar.

Figure 8 shows the screen that appears when the user clicks on the “methodologies” button. This screen presents the details of the methodologies used for the quantitative data obtained in the laboratory analysis. In order to guarantee the efficiency of the classifier, the idea was to inform the methodologies used, in addition to the units of each attribute.



Metodologias

Textura Mediana,Basal e Média (entre texturas apical, mediana e basal):
Determinada com o auxílio do Texturômetro/Penetrômetro Digital (VICTOR – GY4) com a distância de penetração de 10 mm, utilizando-se ponteira circular reta de 3,50 mm de diâmetro. Resultados expressos em Newtons.

pH e SS (Sólidos
Determindada seguindo metodologia apresentada em IAL (2005).

CF:
Determindada seguindo metodologia apresentada po SINGLETON;ROSSI Jr., (1965). Resultados expressos em mg de equivalente de ácido gálico 100 g-1 massa fresca.

FRAP
Determindada seguindo a metodologia apresentada por Benzie e Strain (1996), os resultados serão apresentados em mmol Fe kg-1.

DPPH
Determindada seguindo metodologia por Brand-Williams et al. (1995). Resultados expressos em %DPPH reduzido.

VIT C:
Determindada seguindo metodologia apresentada em IAL (2008).Resultados foram expressos em mg de ácido ascórbico 100g-1 de massa fresca.

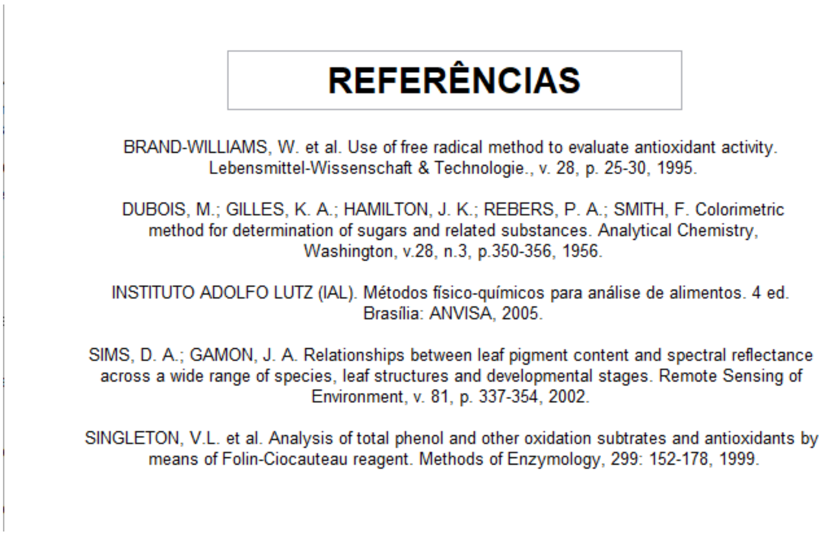
ANT. POLPA & CL A CASCA
Determindada seguindo metodologia apresentada por Sims e Gamon, 2002. Resultados expressos em mg do referido pigmento por 100 g de amostra.

ACÚCAR:
Determindada seguindo metodologia apresentada por Dubois et al. (1956). Resultados expressos em g de açúcar em 100 g de amostra.

Figure 8: Interface that presents the methods used and their respective units to be inserted when classifying the bananas

Subtitle: Methodologies; Median, basal and mean textures (between apical, median and basal textures): Determined with the aid of a texturometer/digital penetrometer (VICTOR – GY4) at a penetration distance of 10 mm using a straight circular tip with a diameter of 3.50 mm. Results expressed in Newtons; pH and SS: Determined following the methodology presented in IAL (2005); PC: Determined following the methodology presented in SINGLETON; ROSSI JR., 1965. Results expressed in mg of equivalent gallic acid per 100 g-1 fresh mass; FRAP: Determined following the methodology presented in Benzie and Strain (1996). Results expressed in mmol Fe kg-1; DPPH: Determined following the methodology presented in Brand-Williams et al. (1995). Result expressed in reduced DPPH %; VIT C: Determined following the methodology presented in IAL (2008). Results expressed in mg of ascorbic acid per 100 g-1 of fresh mass; ANT. (Anthocyanin) PULP & CL. (Chlorophyll) A PEEL: Determined following the methodology presented in Sims and Gamon (2002). Results expressed in mg of pigment per 100 g of sample; SUGAR: Determined following the methodology presented in Dubois et al. (1956). Results expressed in g of sugar per 100 g of sample.

Figure 9 shows the screen that appears when the users clicks on the “references” button. It shows the references for the methodologies used in the extraction of attributes to classify the banana ripening stage.



REFERÊNCIAS

BRAND-WILLIAMS, W. et al. Use of free radical method to evaluate antioxidant activity. Lebensmittel-Wissenschaft & Technologie., v. 28, p. 25-30, 1995.

DUBOIS, M.; GILLES, K. A.; HAMILTON, J. K.; REBERS, P. A.; SMITH, F. Colorimetric method for determination of sugars and related substances. Analytical Chemistry, Washington, v.28, n.3, p.350-356, 1956.

INSTITUTO ADOLFO LUTZ (IAL). Métodos físico-químicos para análise de alimentos. 4 ed. Brasília: ANVISA, 2005.

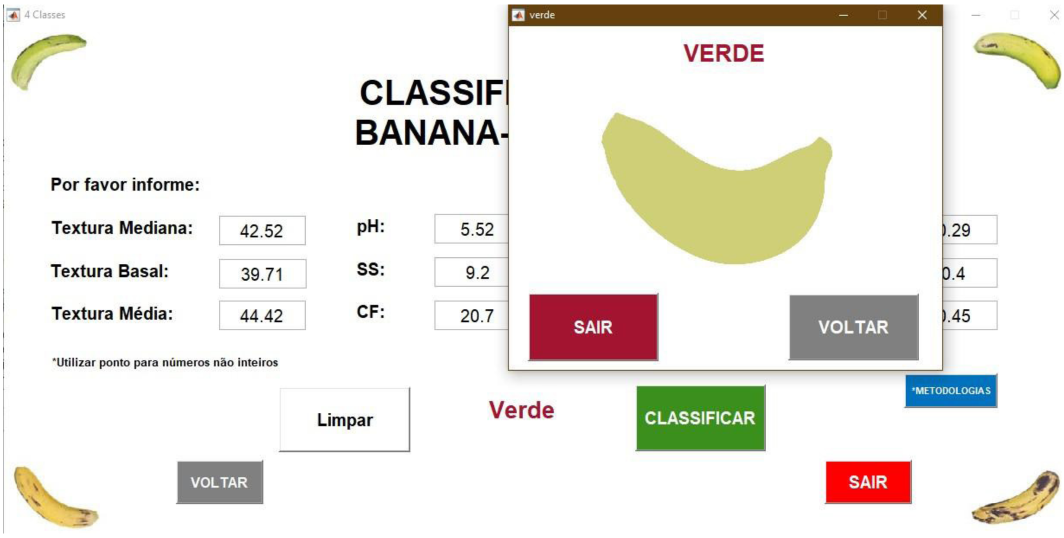
SIMS, D. A.; GAMON, J. A. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sensing of Environment, v. 81, p. 337-354, 2002.

SINGLETON, V.L. et al. Analysis of total phenol and other oxidation subtrates and antioxidants by means of Folin-Ciocauteau reagent. Methods of Enzymology, 299: 152-178, 1999.

Figure 9: Interface of the “references” screen

Subtitle: References.

Figure 10 shows the screen that appears when the user inserts the data to classify the bananas. As observed, when the user clicks on the "classify" button, it indicates at which ripening stage the fruit is and shows the fruit classification according to the data inserted on the screen shown in Figure 7. As seen in Figure 10, the classification takes place in two places, both on the data entry screen and on the classification screen itself.



4 Classes

CLASSIFI
BANANA-

Por favor informe:

Textura Mediana: 42.52
Textura Basal: 39.71
Textura Média: 44.42
pH: 5.52
SS: 9.2
CF: 20.7
.29
0.4
.45

*Utilizar ponto para números não inteiros

Verde

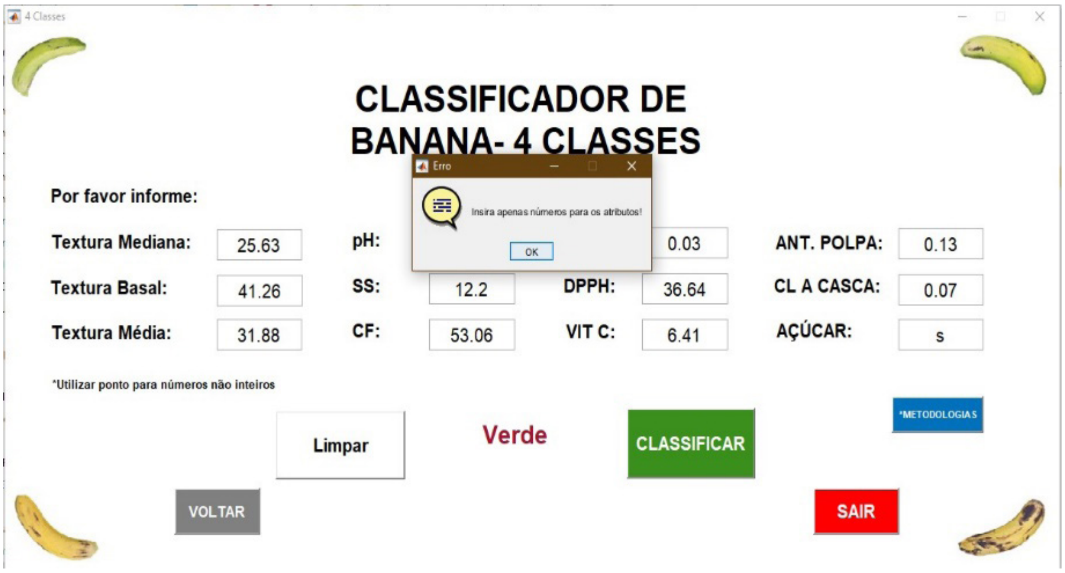
verde

VERDE

Figure 10: Interface of the classification screen

Subtitle: Banana Classifier - 4 classes; Please insert; median texture, basal texture, mean texture, pH, SS (Soluble Solids), PC (Phenolic Compounds), FRAP, DPPH, VIT C, Anthocyanin. Pulp, Chlorophyll a peel, Sugar.

During the software development, two possible problems that could arise for the the user were verified: failure to insert a value, or the insertion, by accident, of a letter instead of a number; in either situation an error message was displayed. Figure 11 shows the error message that was displayed when the error was detected.



4 Classes

CLASSIFICADOR DE
BANANA- 4 CLASSES

Por favor informe:

Textura Mediana: 25.63
Textura Basal: 41.26
Textura Média: 31.88
pH:
SS: 12.2
CF: 53.06
0.03
DPPH: 36.64
VIT C: 6.41
ANT. POLPA: 0.13
CL A CASCA: 0.07
AÇÚCAR: s

*Utilizar ponto para números não inteiros

Verde

Erro

Insira apenas números para os atributos!

Figure 11: Error message during data insertion

Subtitle: Banana Classifier - 4 classes; Please insert; median texture, basal texture, mean texture, pH, SS (Soluble Solids), PC (Phenolic Compounds), FRAP, DPPH, VIT C, Anthocyanin. Pulp, Chlorophyll a peel, Sugar.

As a complementary analysis, the normality and homoscedasticity of the data were applied to the data obtained using the Anderson-Darling and Levene Tests. The data were subjected to analysis of variance using the F test ($P \leq 0.05$) for treatments. The results were subjected to the Tukey test at the P ≤ 0.05 level. The maturation and inherent processes of synthesis of metabolites such as sugars (primary) and ascorbic acid (secondary), as well as degradation of primary and secondary metabolites, such as pigments, could be evidenced throughout the development of the work. At the end of the experiment, it was possible to verify that the fruits of stage 4 presented higher values of firmness and phenolic compounds, which showcases how this stage is the preferred one in relation to *in nature* consumption.

Pearson correlation analysis was performed to investigate the relationships between the study variables, which indicate the existence of a positive or negative relationship between two variables, adopting α = 5% (correlation coefficient) to verify the significance. In this complementary analysis, the concept of correlation developed by Cohen (1988) was used, where values between 0.10 and 0.29 can be considered weak; values between 0.30 and 0.49 can be considered moderate; and values between 0.50 and 1 can be interpreted as strong. These are presented in tables with colours that vary from light to dark grey, according to the intensity of the correlation. Regardless of the sign of the value, the closer it is to 1, the greater the degree of linear statistical dependence between the variables. On the other hand, the closer to zero, the lower the strength of this relationship.

During ripening, the reduction of chlorophyll is directly related to the reduction of magnesium, a nutrient that participates in the synthesis and is a central part of the chlorophyll molecule, as well as linked to 4 other nitrogen atoms. However, it is not possible to notice this behavior during the development of the work. According to Marenco and Lopes (1997), there is a high correlation between photosynthetic pigments and leaf N and Mg concentrations. This high correlation is also attributed to the fact that 5- to 70% of the total N in leaves are part of enzymes that are associated with chloroplasts, and Mg is an enzymatic activator of these.

The average concentration of total carotenoids in the skins of fruits at an advanced senescence stage was 40% higher in fruits at stage 6 than that found in green fruits. Aquino *et al.* (2018), who studied the carotenoid contents of ripe and unripe fruit pulp from 15 banana cultivars, obtained similar results. This colour change is due to enzymatic action on the chlorophyll structure, enabling the expression of carotenoids (Newilah *et al.*, 2009), and occurs both in the pulp and skin of the fruits during ripening. For this reason, the correlation between the main photosynthetic pigment (chlorophyll a) in both skin and pulp, anthocyanins, and flavonoids was negative (also in both skin and pulp in relation to these first pigments).

Another important observation was made in relation to the positive correlation between anthocyanins in peel and pulp and the flavonoids evaluated. Anthocyanins are flavonoids found widely in nature. In this work, the fruits showed increases in the levels of these pigments and flavonoids quantified during ripening and, therefore, a strong positive correlation.

In order to further complement the interpretation and understanding of the obtained results, Cluster Analysis was carried out, which is one of the multivariate analysis techniques that aim to approximate objects, based on their characteristics, where the proximity between them is generally indicated by the distance between the evaluated clusters by the so-called Mahalanobis distance (D2). For both analyses, Minitab software was used. In a second approach, multivariate statistical analysis was applied to the dataset and, for this, a classification model was built, adopting principal component analysis. Multivariate analysis was carried out to verify the grouping of different responses and obtain additional information about the influence of the analysed variables in relation to the banana ripening stages. In this evaluation, it was found that components 1 and 2 (maturation stages and laboratory analyses, respectively) explained 52.34% of the variance in the experiment.

The positions of the vectors in the graph showed that the evaluated attributes were separated into two groups. When evaluating the presented results, it is possible to verify that, in relation to the variables that present a reduction in their respective levels as the stage of maturation increases, there is an approximation of these on the left side of the figure, such as Antioxidant Activity by the DPPH and FRAP methods, Length, Weight, Diameter, and Textures. The opposite behavior, with an increase in levels demonstrated by parameters such as titratable acidity, soluble solids, pH, and total sugars, is showcased in the grouping on the right side of the figure.

To finalize the different ways in which data can be evaluated, a Feedforward Backpropagation ANN (artificial neural network) was created with three layers: input with 8 neurons, intermediate layer with 10, and output with 2 neurons (maturation stages), as shown in Figure 12. The software used was Matlab® (MATHWORKS, 2021).

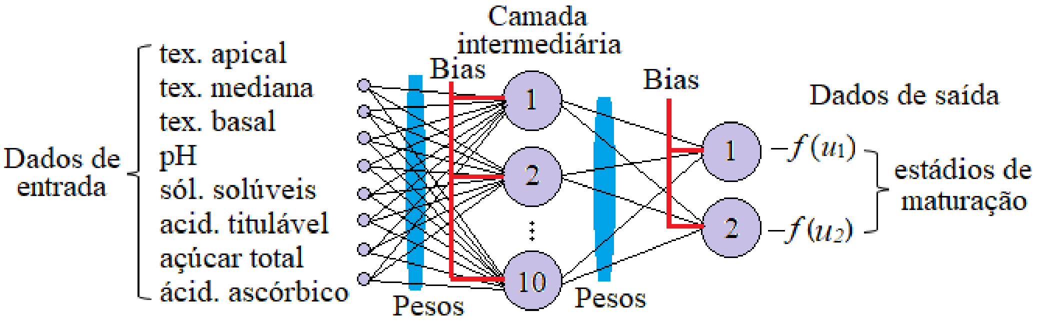


Figure 12- Feedforward Backpropagation ANN used in partial work.

Source: Prepared by the authors (2022)

Subtitle: Dados de entrada = input data; Dados de saída = output data; Camada intermediária = intermediary layer; Pesos = weight; bias = bias; Tex apical, tex mediana, tex basal = apical texture, median tex, basal tex; Sól. Solúveis = soluble solids; Acidez titulável = titratable acidity; Açúcar total = total sugar; Ácido ascórbico = ascorbic acid; Estádios de maturação = maturation stages

In addition to the aforementioned technique, the Generalized Regression ANN was used, which is a statistical technique that extends common linear regression to deal with a wide range of types of dependent variables. Rather than being limited to continuous dependent variables, generalized regression allows you to work with dependent variables that do not fit the normal distribution. This includes binary, count variables, and other non-Gaussian distributions.

From the results referring to the Feedforward Backpropagation ANN, it was observed that the network was well-trained in the three presented configurations, with results close to 100% accuracy. For the validation and test phases, only two samples were classified wrong in the first and second configuration, with 91.6% and 94.4% correct results, respectively. For the third configuration, although the training was excellent, there were five errors in classifying the samples, with 89.5% accuracy in the validation and testing phases, which can be considered good, given that only 60% of the data was used for training. This meant that ANNs could not be as assertive, as there were fewer standards to carry out their training. In this context, the first and second configurations showed better results. In general, the average success rate was 97.5%.

In the results referring to the Generalized Regression ANN, it was observed that the network was also well-trained, even superior to that of the Feedforward Backpropagation ANN, obtaining the same settings (80% training) with results close to 100% accuracy. In the validation phase, only two samples were classified incorrectly, totalling 91.6% accuracy.

In general, both used ANNs showcased excellent performance, when using 80% of the database for training. The second used ANN had a slight increase in performance, which may be due to a greater number of characteristics of the fruit that was used, or perhaps the Generalized Regression ANN had a better affinity with this type of situation.

Conclusion

From the analysis of the mean F-measure values on WEKA software, it was possible to conclude that the best result for the classification of the databases used herein was with 13 attributes, not being necessary the use of all attributes collected in Bayes Net for the cross-validation method. As many attributes were collected, the attributes were selected for classification among the four classes.

In order to classify the bananas into the four classes, the most relevant attributes were filtered, namely: apical, central, basal and mean textures, pH, soluble solids, phenolic compounds, FRAP, DPPH, vitamin C, anthocyanins from the pulp, chlorophyll a content in the fruit peel and sugar, resulting in a mean F-measure of 97.0%. Undoubtedly, the subjectivity inherent to inspection processes based on essentially visual parameters must be taken into account, as it can lead to serious problems between producers/sellers and inspectors.

Acknowledgments

The authors gratefully acknowledge and thank the São Paulo Research Foundation – FAPESP for the financial support (grant number 2020/14166-1) and for the scholarship provided to J. M. Mello (FAPESP - grant number 2020/01711-1) and V. F. S. Favaro (FAPESP – grant number 2021/08901-3).

References

AQUINO, C.F.; SALOMÃO, L.C.C.; CECON, P.R.; SIQUEIRA, D.L.; RIBEIRO, S.M.R. Physical, chemical and morphological characteristics of banana cultivars depending on maturation stages. C**aatinga**, Mossoró, v.30, n.1, p.87-96, 2017. *https://doi.org/10.1590/1983-21252017v30n110rc*.

ARIVAZHAGAN, S., SHEBIAH, R.N., NIDHYANANDHAN, S.S., GANESAN, L. Fruit recognition using color and texture features. **Journal of Emerging Trends in Computing and Information Sciences**, v.1, n.2, p.90-4, 2010.

BENZIE, I.F.F.; STRAIN, J.J. The ferric reducing ability of plasma (FRAP) as a measure of antioxidant power: The FRAP assay. **Analytical Biochemistry**, New York, v.239, p.70–76, 1996.

BRAND-WILLIAMS, W.; CUVELIER, M.E.; BERSET, C.L.W.T. Use of free radical method to evaluate antioxidant activity. **Analytical Biochemistry**, New York, v.28, p.25-30, 1995.

COHEN, J. **Statistical power analysis for the behavioral sciences**. 2nd ed. Hillsdale: Eribau, 1988.

DAMASCENO, M. **Introdução a mineração de dados utilizando o weka**. Macau. Disponível em: *http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/258/207*. Acesso em: 10 nov. 2019.

DUBOIS, M.; GILLES, K.A.; HAMILTON, J.K.; REBERS, P.A.; SMITH, F. Colorimetric method for determination of sugars and related substances. **Analytical Chemistry**, Washington, v.28, n.3, p.350-6, 1956.

HUANG, D.; OU, B.; PRIOR, R.L. The chemistry behind antioxidant capacity assays. **Journal of Agricultural and Food Chemistry**, Easton, v.53, n. 6, p.1841-56, 2005.

IAL - Instituto Adolfo Lutz. **Métodos físico-químicos para análise de alimentos**. 4.ed. Brasília: ANVISA, 2005.

LANDIS, J.R; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, Malden, v.33, n.1, p.159-74, 1977.

LOBÃO, J.S.B.; FRANÇA-ROCHA, W.J.S.; SILVA, A.B. Aplicação dos índices kappa e pabak na validação da classificação automática de imagem de satélite em Feira de Santana-BA. *In*: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12., 2005, Goiânia . **Anais [**...]. São José dos Campos: INPE, 2005. p.1207-14.

MARENCO, R.A.; LOPES, N.F. **Fisiologia vegetal**: fotossíntese, respiração, relações hídricas e nutrição mineral. 2.ed. Viçosa: Editora UFV, 2007.

MATHWORKS. MATLAB (MATrix LABoratory). 2018. Acesso em 28 abril 2023. Disponível em: <*http://www.mathworks.com*>.

MORETI, C.V.P. **Comparação de métodos descritores e métodos de aprendizado de máquina aplicados no reconhecimento de imagens de frutas**. 2020. 54f. Trabalho (Conclusão de Curso de Engenharia de Computação) - Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2020.

NEWILAH, G.N.; DHUIQUE-MAYER, C.; ROJAS-GONZALEZ, J.; TOMEKPE, K.; FOKOU, E.; ETOA, F.X. Evaluating bananas and plantains grown in Cameroon as a potential sources of carotenoids. **Food**, v.2, n.2, p.135-8. 2008.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. IEEE Latin America Transactions [S.l.], v. 6, n. 2, p. 215-222, 2008. Disponível em: <*http://ieeexplore.ieee.org/stamp/stamp.do?arnumber=4609920&isnumber=4609907*>.

SANTOS, A.F.; ELER, D.M.; ARTERO, A.O.; DIAS, M.A.; POLA, I.R.V. **Combining feature extraction techniques for fruit classification**. 2020. Disponível em: *https://www.researchgate.net/profile/DaniloEler/publication/344407154_Combining_Feature_Extraction_Techniques_for_Fruit_Classification/links/5f721db8458515b7cf563403/Combining-Feature-Extraction-Techniques-for-Fruit-Classification.pdf*. Acesso em: 07 jul. 2023.

SENG, W.C., MIRISAEE, S. H. **A new method for fruits recognition system**. Kuala Lumpur: Faculty of Computer Science & Information Technology, University of Malaya, 2009.

SILVA, R. C; PAES, A. T. Teste de concordância Kappa. Educ Contin Saúde Einstein. **2012**;10(4):165-6. Disponível em: *http://apps.einstein.br/revista/arquivos/PDF/2715-165-166.pdf*. Acesso em: 05 dez.2020.

SIMS, D.A.; GAMON, J.A. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. **Remote Sensing of Environment**, New York, v.81, p.337-54, 2002.

SINGLETON, V.L.; ORTHOFER, R.; LAMUELA-RAVENTÓS, R.M. Analysis of total phenol and other oxidation subtrates and antioxidants by means of Folin-Ciocauteau reagent. **Methods of Enzymology**, Madison, v.299, p.152-78, 1999.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, F-Score and ROC: a family of discriminant measures for performance evaluation. *In*: SATTAR, A.; KANG, B.H. (ed.). **AI 2006**: advances in artificial intelligence, lecture notes in computer science. Heidelberg: Springer, 2006. v. 4304. p.1015-21. *https://doi.org/10.1007/11941439_114.2006*.

SOUZA, A.V.; MELLO, J.M.; FAVARO, V.F.S.; SANTOS, T.G.F.; SANTOS, G.P.; SARTORI, D.L.; PUTTI, F.F. Metabolism of bioactive compounds and antioxidant activity in bananas during ripening. **Journal of Food Processing and Preservation**, Oxford, v.45, p.e15959, 2021.

WEKA - Waikato Environment For Knowledge Analysis**. Development University of Waikato. The University of Waikato Hamilton, New Zeland**. Hamilton, 2020. Disponível em: *https://www.waikato.ac.nz/study/?gclid=CjwKCAiA8YyuBhBSEiwA5R3-E_QgUcR2bSuncx910fgtXV0J4q0x8OwJe9FdlWbr3Y77XvvFZ-9g6xoCQ6wQAvD_BwE&gclsrc=aw.ds*. Acesso em: 05 maio 2022

WITTEN, I.H.; FRANK, E. **Data mining pratical machine learning tools and techniques**. 2.ed. San Francisco: Elsevier, 2005. 525p.