Matemática Aplicada, Computacional e Industrial

MACI

Vol. 8

2021

Trabajos presentados al VIII MACI 2021

Proceedings of VIII MACI 2021

La Plata, 3 al 7 de mayo de 2021

EVALUATION OF MACHINE LEARNING ALGORITHMS -K NEAREST NEIGHBORS AND SUPPORT VECTOR MACHINES- FOR STRAWBERRIES CLASSIFICATION DURING FOOD DRYING PROCESS

Juliana Gamboa-Santos[†] & Laura A. Campañone[†, ††]

[†]Grupo de Tecnología de Alimentos, CIDCA (CONICET-CCT La Plata y Universidad Nacional de La Plata). Calle 47 y 116, 1900 La Plata, Argentina, j.gamboa@conicet.gov.ar , www.cidca.com.ar
[††]Departamento de Ingeniería Química, Facultad de Ingeniería (Universidad Nacional de La Plata), Calle 1 y 47, 1900 La Plata, Argentina, lacampa@ing.unlp.edu.ar , www.unlp.edu.ar

Abstract: In the present work, supervised machine learning (ML) algorithms, k-NN and SVM, were applied to classify strawberry samples during a microwave-assisted drying process. A dataset of 1150 strawberry records containing information about images of two pre-treatments types (fresh, FR and osmotically pre-treated, OD), three ranges of drying times (short <40 min; intermediate: 40-70 min and long> 70 min) and three physical characteristics previously selected (shrinkage, brightness and saturation) was used to perform the ML classifiers. The k-NN and SVM models led to good accuracy values, 0.94 for sample type and 0.90 for drying time categories. Since colour and morphological changes are related to changes in the product quality, these results are useful to evaluate the losses of nutritional and sensorial properties taking place during the in-line processing of strawberries, by means of a non-invasive monitoring technique.

Keywords: *machine learning, non-invasive food quality monitoring, classification algorithms, microwaves assisted drying, strawberry, digital image analysis*
2000 AMS Subjects Classification: 68T10

1. INTRODUCTION

During the last decades, Machine learning (ML) emerged as a subset of Artificial Intelligence (AI) with the aim to develop a conceptual understanding of how the human brain works. Despite at the beginning, ML was though as an early stage of robotic era, ML was broadly become as a scientific field making focus on the design of computer models and algorithms to perform specific tasks without the need to be computers explicitly programmed [1]. The most common data science tasks could be to describe patterns and trends, to estimate the value of numerical target variables using a collection of predictors, to resolve classification problems when target variables are categorical, to identify groups (clusters) of records which are similar, to predict the value of certain variables in the future or to associate attributes and uncover rules for quantifying the relationship between two or more attributes [2-6]. ML can be used to learn models from labelled samples, that is usually known as a supervised classification task [4]. Typical supervised classification algorithms include, logistic regression (LG), support vector machines (SVMs), decision trees (DT) and k-nearest neighbors (k-NN), among others [7]. ML and scientific computing applications commonly utilize linear algebra operations on multidimensional arrays, which are computational data structures for representing vectors, matrices, and tensors of a higher order. In recent years, substantial efforts are being spent on the development of improved user-friendly libraries for scientific computing and ML. As an exponent of high-level library, Scikit-Learn, which primarily use NumPy and SciPy is considered one of the most popular open-source libraries for classical ML [1].

This work aims to analyse a dataset of quality features obtained from images of fresh (FR) and pre-treated (osmo-dehydrated, OD) strawberries during a microwave (MW) assisted drying process in order to classify the strawberry samples by type (FR and OD) and drying stage (short, intermediate and long drying times) by using k-NN and SVM algorithms. The evaluation by a non-invasive technique of the loss of quality indicators during drying of strawberries could be of industrial relevance in case of being adapted to specific tasks and food processing stages. In addition, the ML algorithms obtained could be useful as an indirect predictive tool of the loss of sensorial quality, in the conditions assayed, for strawberries during MW drying processing.

2. MATERIALS AND METHODS

Features extracted from strawberry images and collected in the final dataset were: area (loss of cross-

sectional areas) and two colours indicator (saturation and brightness). Area, Brightness and Saturation retention values represent the proportion of the cross-sectional area in segmented strawberry samples maintained at each MW drying time. The latter were calculated by using the corresponding threshold values summarized on **Table 1**.

Table I. Threshold conditions in HSV coordinates applied for feature extraction during MW drying of strawberries

Feature	Hue (H)	Saturation (S)	Value (V)
Area	0-20, 160-179	0-255	1-255
Brightness	0-20, 160-179	0-255	60-255
Saturation	0-20, 160-179	153-255	0-255

Modelling phase was conducted by means of an i5 9400 coffelake microprocessor equipped with a memory DDR4 16 Gb 2400. According to supervised classification methods, data was partitioned using random assignment, into a training dataset (60%) and a test dataset (40%). As it is known, the training set records are complete, but the test set records should have the target variable (temporarily) omitted. So, the data science models learn about the patterns and trends in the data using the training data set. These models were applied to the test set, where they make predictions for the temporarily unknown values of the target variable. These predictions were then evaluated against the (now restored) true target values, using evaluation measures such as overall error rate or mean squared error. Support Vector Machines (SVC) and K-nearest neighbor algorithms (k-NN) were built using Spyder (Python 3.7) platform environment available from Anaconda distribution (4.9.2 version). Data structure was analysed and evaluated using the Pandas 0.25.1 and Numpy 1.16.5 versions. The auxiliary package used to perform the ML classifiers was Scikit-Learn (0.21.3 version). The model parameters set for k-NN and SVM classifiers were included in **Table II**.

Table II. Model parameters applied during machine learning algorithms implementation.

SVM	K-NN
C=1.0; kernel= "rbf"; degree= 3; gamma= "scale"; coef0= 0.0; shrinking= True; probability= False; tol= 1e-3; class_weight= None; verbose= False; max_iter= -1 (no limit); decision_function_shape= "ovr"; break_ties= False; random_state= Seed*	N_neighbors= 5; Weights= "uniform"; Algorithm= "auto"; Leaf_size= 30; p=1 (Manhattan-distance); metric= Minkowski metric; metric_params= None; n_jobs= None

**It was performed by setting a function script "Seed", range (100).*

To avoid data dredging, model validation was addressed by a 5x2 coss-validation study according to previous references [8,9]. Cross-validation method allows to compare experimentally several ML algorithms by obtaining pairs of testing-set classification accuracies (or the metric of choice) [4]. Classification model evaluation measures were conducted by means of the confusion matrix generated by the classification model [4]. Accuracy, as the evaluation measure, is defined in Eq. 1.

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \quad (1)$$

Accuracy represents an overall measure of the proportion of correct classifications being made by the model. TN, FN, FP and TP represent the numbers of true negatives, false negatives, false positives, and true positives, respectively [4]. Additionally, RMSE (Eq. 2) was computed in order to evaluate the classification performance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{ie} - X_{ic})^2}{n}} \quad (2)$$

Where *X* represents the computed variable, *e* and *c* subscripts are the experimental and calculated average values and *n* is the number of experimental data.

3. RESULTS AND DISCUSSION

Figure 1 displays the confusion matrix obtained for SVM classifier according to drying time categories. As can be seen, 91% (short), 83% (intermediate) and 87% (long) of training samples were correctly classified. For testing samples, the same trend was observed since intermediate drying times accounted lower classification performances: 93% (short), 71% (intermediate) and 82% (long).

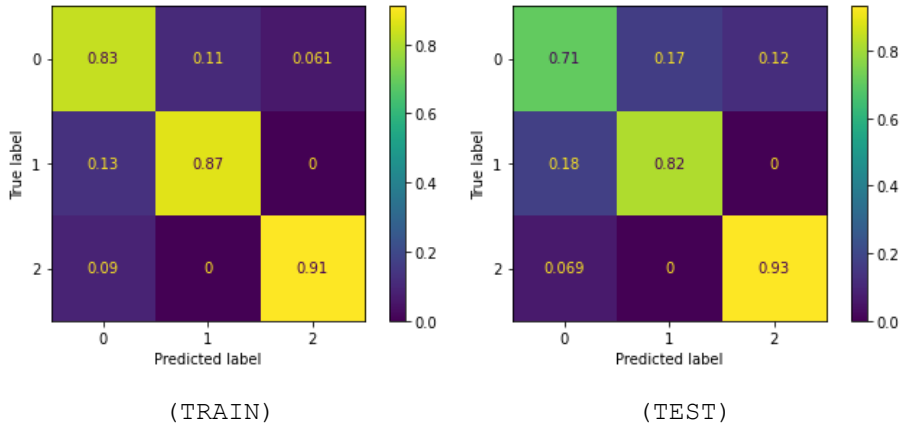

Figure 1. Confusion matrix for SVM classifier according to time categories for training and testing samples. Labels correspond to: "2": short, "0": intermediate, "1": long drying times.

Figure 2 showed the cross-validation results performed up to 10 folds for time categories by applying the SVM classifier. As expected, the higher the k-fold the lower the RMSE value. Since the lower RMSE values were obtained after 6-folds, 10-folds was selected to compare SVM and k-NN algorithms.

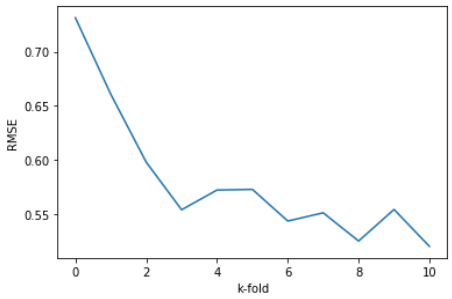

Figure 2. Cross-validation study for drying time categories applying the SVM classifier.

Table III showed the performance parameters obtained for the k-NN and SVM algorithms according to (a) type (OD vs FR) and (b) time categories for FR samples classifications, after 10-fold cross validation study. For all the models tested, fit and score times were at most of centesimal order (seconds, s). K-NN (k=5) model accounted the best fit times. All classification strategies resulted in good performance models with maximum accuracies of 0.94 for sample type (FR vs OD) and 0.90 for time categories (short, intermediate, long) classification tasks. As can be seen, the highest accuracy values 0.94 for type classification were found by using SVM technique. Regarding time categories, SVM resulted also in the best ML method to develop better classification models (0.90). In the case of RMSE metric, SVM presented lower levels than k-NN, thus better performance. It has been previously stated that SVM accounted an improved performance compared to other methods such as CART and Random Forest [10]. SVM classifier is based in

the construction of an optimal separating hyperplane between groups linearly separable [11]. Besides, k-NN (k=5) algorithm has risen test accuracy values up to 0.92 for type classification task. In the case of time categories, k-NN accuracies were notably lower (<0.85). In general, lower model accuracies were obtained for time categories classification in the case of FR samples. For OD samples, time categories classification could not be performed since samples showed a great dispersion along the entire drying process (clusters of time categories could not be identified, data not shown).

Table III. Performance results for classification tasks by means of k-nearest neighbors (k-NN) and support vector machines (SVM) algorithms.

Score	Sample type (FR vs OD)		Time categories (3 clusters*)	
	k-NN (k=5)	SVM	k-KNN (k=5)	SVM
Fit time (s)	0.000612	0.007278	0.000599	0.002089
Score time (s)	0.006278	0.001393	0.003388	0.000801
Test Accuracy**	0.92	0.94	0.85	0.90
RMSE	0.345452	0.342368	0.615233	0.525560

*Clusters: short (<40 min), intermediate (40-70 min), long (>70 min) times). **Maximum accuracy values.

4. CONCLUSION

In the present work, two machine learning classifiers, k-NN and SVM, were performed to classified strawberries according to pre-treatment type (FR and OD) and time categories (short, intermediate, long) with good performance results (accuracy range: 0.85-0.94) and only using three selected morphological and colour features. These achievements could be useful to evaluate the loss of quality indicators during drying of strawberries since shrinkage and colour indicators are commonly related with changes in nutritional and sensorial quality. Moreover, to perform the in-line classification task during processing could be of industrial relevance in case of being adapted to specific tasks and food processing stages. The ML algorithms obtained could be useful as an indirect predictive tool of the loss of quality, in the conditions assayed, for strawberries during MW drying processing by using a non-invasive monitoring technique.

ACKNOWLEDGMENTS

This work has been funded by the ANPCyT Agency, Argentina (project PICT 2162/17).

REFERENCES

[1] S. RASCHKA, J. PATTERSON & C. NOLET, *Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence*, Information 11 (2020), p. 193.

[2] Y. GUPTA, *Selection of important features and predicting wine quality using machine learning techniques*, Procedia Computer Science 125 (2018), pp. 305-312.

[3] T.M. GUNARATNE, C. GONZÁLEZ VIEJO, N.M. GUNARATNE, D.D. TORRICO, F.R. DUNSHEA, S. FUENTES, *Chocolate Quality Assessment Based on Chemical Fingerprinting Using Near Infra-red and Machine Learning Modeling*, Foods 8 (2009), p. 426.

[4] C.D. LAROSE & D.T. LAROSE, *Data Science using Python and R*, Wiley & Sons, 2019.

[5] S.E. REICHEMBACH, C.A. ZINI, K.P. NICOLLI, J.E. WELKE, C. CORDERO, Q. TAO, *Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition*, J. Chromathography A 1595, 158-167.

[6] O. GAZELI, E. BELLOU, D. STEFAS, S. COURIS, *Laser-based classification of olive oils assisted by machine learning*, Food Chemistry 302 (2020), 125329.

[7] A. REN, A. Z. AHMED ZOHA, S. A. SHAH, M. IMRAN, A. ALOMAINY, Q. H. ABBASI, *Machine Learning Driven Approach Towards the Quality Assessment of Fresh Fruits Using Non-Invasive Sensing*, IEEE Sensors Journal 20(4) (2020), p.2075-2083.

[8] W. CASTRO, J. OBLITAS, M. DE-LA-TORRE, C. COTRINA, K. BAZÁN, H. AVILA-GEORGE, *Classification of Cape Gooseberry Fruit According to Its Level of Ripeness Using Machine Learning Techniques and Different Colour Spaces*, IEEE Access 7 (2019), pp. 27389-27400.

[9] X.F. CADET, O. LO-THONG, S. BUREAU, R. DEHAK & M. BESSAFI, *Use of Machine Learning and Infrared Spectra for Rheological Characterization and Application to the Apricot*, Scientific Reports 9 (2019), p. 19197.

[10] A.M. JIMÉNEZ-CARVELO, A. GONZÁLEZ-CASADO, M.G. BOGUR-GONZÁLEZ, L. CUADROS-RODRÍGUEZ, *Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review*, Food Research International 122 (2019), pp. 25-39.

[11] B.V. Canizo, L.B. Escudero, R.G. Pellerano, R.G. Wuilloud, *Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes*, Computers and Electronics in Agriculture 162 (2019), pp. 514-522.