Estimation of nutritional levels of citrus leaves by non-destructive analysis using machine learning techniques

Guillem Miralles[1], Isabel Rodríguez-Carretero[2], Sergio Cubero[3], Marcelino Martínez[1], Fernando Mateo[1], Ana Quiñones[2], José Blasco[3], Juan Gómez-Sanchís[1*]

[1] IDAL, Department of Electronic Engineering, University of Valencia, Av. de la Universidad , S/N, 46100Burjassot, Valencia (Spain). Email: Juan.Gomez-Sanchis@uv.es
[2] Centre for the Development of Sustainable Agriculture. Instituto Valenciano de Investigaciones Agrarias (IVIA), CV-315, km 10,7, 46113 Moncada (Valencia), Spain.
[3] Centro de Agroingeniería, Instituto Valenciano de Investigaciones Agrarias (IVIA), CV-315, km 10,7, Moncada (Valencia) 46113, Spain.

Abstract: One of the characteristics of all living beings is that adequate nutrition has a positive impact on health. In the case of plants, and specifically in fruit trees, adequate nutrition is also essential for them to grow healthy and produce fruits in the highest quantity and quality possible. Therefore, optimal nutrition is key for any farmer. However, excessive use of fertilisers can harm the environment and be a waste of resources for farmers. One of the keys to achieving adequate fertilisation is an accurate diagnosis of the nutritional status of the tree. Traditionally, this diagnosis is made by destructive ionomics analysis, which represents a high economic cost and a delay in obtaining the results. This work proposes Vis-NIR hyperspectral imaging and machine learning regression models to estimate the concentrations of macronutrients (N, P, K, and Ca) and micronutrients (Mn and Fe) in citrus leaves. The methodology involved the application of several machine learning regression methods (linear regression, partial least squares, random forest, support vector regression, and Ada Boost). Data were normalised with standard normal variable (SNV), and principal component analysis (PCA) was used to reduce dimensionality. The results were promising in estimating nutrients with R^2 greater than 0,50 in all cases, especially nitrogen (R^2 of 0.77).

Key words: machine learning, hyperspectral imaging, spectroscopy, precision agriculture, citrus fruit

1. Introduction

The integration of information technology (ICT) in agriculture has optimized intelligent data-driven decision making, boosting productivity and minimizing the ecological impact of agricultural processes [1]. Plants require essential nutrients to grow healthily and provide quality products [2]. Therefore, optimal nutrition is key for any farmer. However, excessive use of fertilizers can damage the environment and represent a waste of resources for farmers [3]. One of the keys to achieve adequate fertilization is an accurate diagnosis of the nutritional status of the tree. Traditionally, this diagnosis is performed by destructive ionomic analysis, which represents a high economic cost and a delay in obtaining the results.

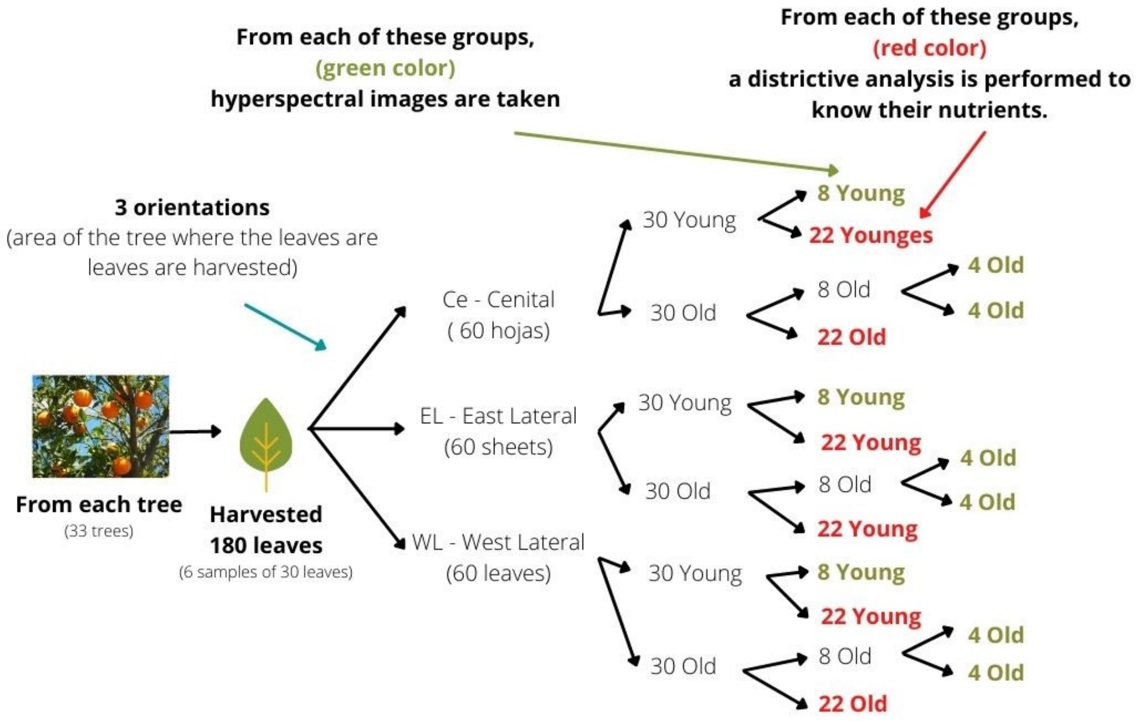
Hyperspectral imaging is a technique that captures information in multiple bands of the electromagnetic spectrum. This allows identifying and distinguishing variations in the reflectance or energy emission of objects in the scene related to the state of the plant, which is useful for crop analysis [4]. Due to their wide spectral range, hyperspectral images can provide rapid information on biochemical properties of crops, such as chlorophyll concentration, moisture content, vegetative state, and nutrient content. This facilitates crop condition monitoring, early detection of diseases or stresses, and optimization of irrigation and fertilization programs. This work proposes the use of Vis-NIR hyperspectral images and machine learning regression models to estimate the concentrations of macronutrients (N, P, K and Ca) and micronutrients (Mn and Fe) in citrus leaves, from the data obtained from the images and machine learning regression models.

2. Materials and methods

2.1. Experimental Design

The plant material came from a citrus plot in Almenara (Castellón) of the Clemenules variety. In May 2021, 33 trees were sampled, 3 samples of 60 leaves in different orientations: east, west and zenith. From each orientation, 30 young leaves (spring budding) and 30 old leaves were selected, for a total of 5940 leaves. From each sample, 22 leaves were used to obtain nutrient reference levels by leaf ionomics analysis and another 8 were used to acquire hyperspectral images. Due to their different sizes, in the case of old leaves, 2 images were taken in groups of 4 leaves; and in the case of young leaves, a single image of the 8 leaves was taken. Thus, 3 hyperspectral images were obtained per tree and orientation: young leaves, first partition of old leaves and second partition of old leaves. The process is described in Figure 1 below.

Figure 1. Diagram of the hyperspectral and nutritional data collection process.



2.2. Image acquisition and processing

A hyperspectral image acquisition system in the Vis-NIR range, consisting of an industrial camera and two liquid crystal tunable filters (LCTF), was used. Images of 1392 x 1040 pixels with a spatial resolution of 0.14 mm/pixel were captured in the spectral range from 400 nm to 1050 nm, in 10 nm steps. Each image was composed of 65 bands.

A system based on diffuse halogen illumination was used, consisting of 12 35 W halogen lamps placed inside a hemispherical hood. To avoid problems of blurred images due to chromatic dispersion of the lens, the focus of the lens was set to the central wavelength (730 nm) of the working range. A hyperspectral image of a certified reference target was acquired to optimize the dynamic range, avoid image saturation and correct the spectral sensitivity of the system.

After obtaining the hyperspectral images, three digital processing steps were performed to extract the spectrum of each leaf. First, the registered images were calibrated by black and white reference. Then, the spatial displacement between images was corrected. Since the images were obtained with two LCTFs, there was a slight variation between them caused by the geometry of the system. Finally, the leaf was segmented from the rest of the scene using a thresholding mask and the average spectrum of the leaf pixels was obtained. This average spectrum constituted the input to the spectral preprocessing and machine learning stage. The preprocessing methodology is detailed in Fazari et al [5].

2.3. Calibration and test sets, spectral preprocessing and feature engineering.

To evaluate the performance of the regression models, the initial set was divided into a calibration set (Cal) (75 %) and an independent test set (Test) (25 %). Simple spectral preprocessing was performed. In particular, the spectral bands were normalized by standardizing them in order to standardize their range of variation. The method used was standard normal variance (SNV) [6]. A simple feature engineering (RCI) option was also explored using principal component analysis (PCA) [7]. Regression models were evaluated using different number of principal components as input (5, 7, 9, 12, 15, 15, 18, 20, 25 and 30, these numbers were determined empirically until no improvement in the performance of the regression methods was achieved) and with all bands without applying PCA.

2.4. Models

In this study, five regression techniques from the field of machine learning were used including Linear Regression, partial least squares (PLS), support vector regression (SVR), Random Forest and Ada Boost. All of them were implemented in the Python programming language with the help of the Scikit-learn library.

- **Linear regression** was used as a reference method to compare with the other techniques used in the work. It bases its operation on assuming a linear relationship between the independent variables and the dependent variable, its main advantage lies in the interpretability of the model [8].
- **PLS** is a regression method especially indicated when the input variables of the problem are highly correlated. The method allows modeling the problem by reducing the set of variables to a smaller set of uncorrelated components [9].
- **SVR,** is a technique based on support vector machines. This regression technique is able to handle nonlinear relational between the input variables and the independent variable of the problem, providing a robust solution against possible outliers [10].
- **Random Forest** is a regression method based on the combination (ensemble methods) of decision trees. It is characterized by aggregating multiple solutions to improve the prediction and robustness of the solution [11].
- **Ada Boost** is a boosting-based regression method based on the idea of creating a strong model from several weak models [12].

2.5. Construction of the models.

To ensure a correct construction of the models, the calibration set was used. In order to avoid overfitting of the models and to achieve an optimal combination of hyperparameters in the training phase, a K-Fold cross-validation approach [13] with 3 subsets (folds) was adopted. Grid Search technique [14] was used to achieve a suitable combination of hyperparameters. The ranges of variation of the hyperparameters (in those models with free hyperparameters) chosen empirically for each model were:

- **PLSRegression**: *n_components* [1-25], *scale* [False, True], *max_iter* [100-1000], *tol* [0,1-0,00001].
- **SVR**: *kernel ['linear', 'rbf', 'sigmoid'], gamma ['scale', 'auto'], C* [0,1-250], epsilon [0,1-0,00001].
- ***RandomForest***: *n_estimators* [5-500], *max_depth* [10, 50, 100, None], *min_samples*_split [2-10], *min_samples_leaf* [1-4], *max_features ['auto', 'sqrt', 'log2', None], bootstrap [True, False].*
- ***AdaBoost***: *n_estimators* [5-500], *learning_rate* [0,001-1], *loss ['linear', 'square', 'exponential'].*

The analyses and experiments were carried out in the Google Colab cloud environment, which provides computational resources for the execution of data processing tasks. This cloud environment allowed to perform calculations and run models in an efficient and scalable manner.

3. Results and discussion

3.1 Macronutrients.

Table 1 shows the results related to the estimation of the macronutrients nitrogen, phosphorus, potassium and calcium. Different model performance indices are shown, in particular the coefficient of determination R2, the mean absolute error (MAE) and the root mean square error (RMSE) [15], both in the calibration and test set.

It is observed that, for nitrogen, all models show an improvement with respect to linear regression. This trend will be maintained for all the nutrients included in the study. The methodology based on Random Forest using all bands as input, presents the best results in the test set for nitrogen (R2 of 0.77). Regarding phosphorus, the methodology that provides the best results is the Ada Boost combination using 18 principal components as inputs, achieving a coefficient of determination R2 of 0.66. For potassium, the best methodology is SVR using 9 principal components with an R2 of 0.61. Similarly, Ada Boost with 20 principal components is the methodology that provides the best results for calcium with an R2 value of 0.62. The estimation capacity of the best methodologies for all macronutrients is good, with coefficient of determination values above 0.6 in all cases.

The best estimation methodologies extracted from the analysis of the coefficient of determination for the macronutrients are maintained if the other calculated yield indices are analyzed. Thus, if the MAE and RMSE values for each macronutrient are examined, it is observed that the minimum values of the estimation errors coincide with the maximum values of R2. This highlights the robustness of the results in terms of model performance estimation.

Table 1. Results of the models for the macronutrients N, P, K and Ca. The FE column shows the best feature engineering choice (ICR), indicating the number of principal components (PC) or using all bands (No). The best methodology for each macronutrient based on the R2 value in the test set is shown in bold.

	MODELO	ICR	R² Cal	MAE Cal	RMSE Cal	R² Test	MAE Test	RMSE Test
N	LinearRegresion	5 CP	0,585	0,132	0,180	0,335	0,145	0,223
	PLSRegression	7 CP	0,724	0,120	0,149	0,726	0,112	0,143
	SVR	No	0,792	0,102	0,129	0,677	0,122	0,155
	RandomForest	**No**	0,813	0,087	0,123	**0,772**	**0,102**	**0,130**
	AdaBoost	No	0,793	0,101	0,129	0,761	0,105	0,133
P	LinearRegresion	5 CP	0,115	0,030	0,050	0,210	0,035	0,055
	PLSRegression	7 CP	0,667	0,021	0,032	0,604	0,025	0,038
	SVR	9 CP	0,742	0,017	0,028	0,586	0,027	0,039
	RandomForest	12 CP	0,686	0,020	0,031	0,562	0,026	0,034
	AdaBoost	**18 CP**	0,700	0,020	0,030	**0,657**	**0,025**	**0,033**
K	LinearRegresion	5 CP	0,482	0,195	0,274	0,268	0,257	0,331
	PLSRegression	5 CP	0,575	0,171	0,246	0,618	0,153	0,241
	SVR	**9 CP**	0,623	0,144	0,232	**0,614**	**0,144**	**0,239**
	RandomForest	No	0,658	0,132	0,222	0,510	0,179	0,271
	AdaBoost	12 CP	0,596	0,156	0,241	0,597	0,167	0,246
Ca	LinearRegresion	5 CP	0,340	0,724	1.214	0,528	0,791	1,123
	PLSRegression	7 CP	0,730	0,504	0,763	0,537	0,675	1,113
	SVR	12CP	0,794	0,353	0,662	0,595	0,623	1,041
	RandomForest	No	0,747	0,399	0,732	0,533	0,693	1,118
	AdaBoost	**20 CP**	0,739	0,405	0,741	0,**620**	**0,612**	**1,008**

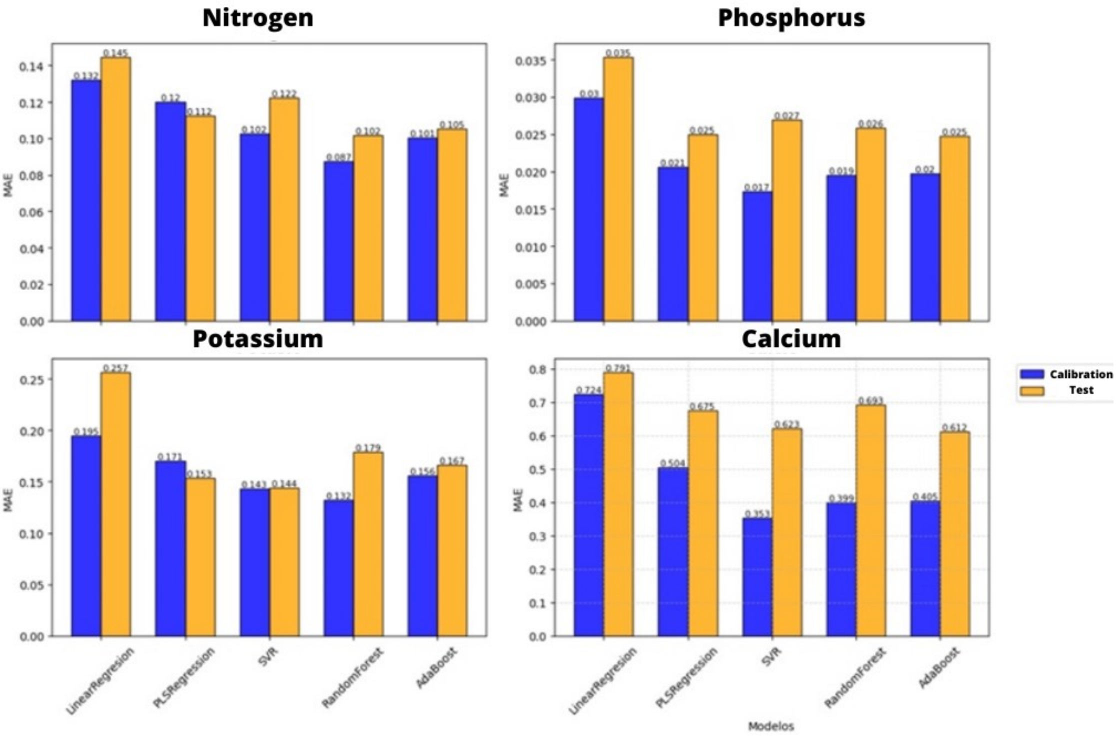


Figure 2. Comparison of the R2 obtained by calibration and testing of the different models predicting macronutrients.

Figure 2 shows, in a bar graph, the comparison between the MAE of the calibration and test sets for each macronutrient with each of the methodologies employed. The general pattern is that the error produced in the calibration and test sets are very similar (within each methodology). This fact shows that there has been no overfitting in the construction of the models and that the generalizability of the models is optimal.

3.2 Micronutrients.

Table 2 shows the results related to the estimation of the micronutrients manganese and iron. The same yield indices are shown as in the case of macronutrients. It can be observed that for manganese all models show an improvement with respect to the linear regression. This trend is also maintained for iron. The best performing methodologies for manganese and iron are Random Forest and Adaboost respectively (R2=0.57 and R2=0.52). The estimation capacity of the best methodologies for micronutrients is lower than in the case of macronutrients, with no coefficient of determination values higher than 0.57 in the test set.

As in the case of macronutrients, the best estimation methodologies extracted from the analysis of the coefficient of determination for micronutrients are maintained if the other calculated yield indices are analyzed. If the MAE and RMSE values for each micronutrient are studied, it is observed that the minimum values of the estimation errors coincide with the maximum values of R2, this highlights the robustness of the results in terms of estimating the performance of the models.

Table 2. Model results for the micronutrients Mn and Fe. The ICR column shows the best feature engineering option, indicating the number of principal components (PC) or using all bands (No). The best methodology for each micronutrient based on the R2 value in the test set is shown in bold.

	MODELO	**ICR**	**R^2 Cal**	**MAE Cal**	**RMSE Cal**	**R^2 Test**	**MAE Test**	**RMSE Test**
Mn	LinearRegresion	5 CP	0,331	4,838	6,081	0,277	5,384	8,321
	PLSRegression	5 CP	0,525	3,866	5,206	0,569	3,988	4,810
	SVR	5 CP	0,517	3,713	5,236	0,566	3,861	4,862
	RandomForest	**No**	0,535	3,777	5,174	**0,570**	**3,832**	**4,808**
	AdaBoost	No	0,543	3,671	5,130	0,391	4,463	5,718
Fe	LinearRegresion	5 CP	0,318	14,337	18,542	0,367	16,070	20,537
	PLSRegression	25 CP	0,410	12,990	17,359	0,465	15,590	18,876
	SVR	5 CP	0,430	12,518	17,089	0,371	14,989	20,479
	RandomForest	25 CP	0,511	12,351	15,816	0,496	14,675	18,326
	AdaBoost	**25 CP**	0,495	12,973	16,091	**0,518**	**14,298**	**17,917**

Figure 3 shows the comparison between the MAE of the calibration and test sets for each micronutrient with each of the methodologies used. As was the case for the macronutrients, the models do not show overfitting since the calibration and test error are very similar.

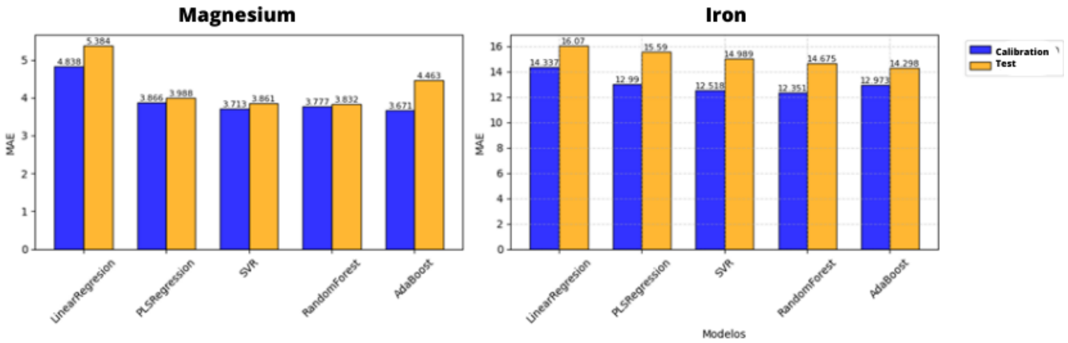


Figure 3. Comparison of R^2 calibration and testing of models estimating iron and manganese.

The mean R^2 of the test results for micronutrients is lower than for macronutrients (0.48 and 0.66, respectively). This is due to the greater complexity and variability in accurately determining micronutrient levels. The estimation of micronutrients presents challenges due to their lower concentration and interaction with other elements. Even so, the results obtained with nondestructive analysis techniques show promising progress in the determination of nutrient levels in citrus leaves, which supports their potential application in future agronomic and nutritional research.

4. Conclusions

The proposed methodology, based on the use of a Vis-NIR hyperspectral vision system and machine learning regression techniques, allows estimating nitrogen content with an R2 of 0.77. For phosphorus, potassium and calcium, R2 values of 0.66, 0.61 and 0.62, respectively, were obtained. In the case of the two micronutrients tested, manganese and iron, values of 0.57 and 0.51, respectively, were obtained, indicating greater difficulty in their estimation.

In the case of micronutrients, the use of hyperspectral vision and machine learning models do not prove to be a really useful alternative for the estimation of nutrient levels in citrus leaves, since in most cases the R2 values are close to 0.5 and even below. This indicates that the models are able to explain at most 50% of the variability in the output data; that is, about 50% of the variation in the predicted values can be attributed to the variables used in the model, while the remaining 50% is due to unknown factors.

The use of the Vis-NIR hyperspectral imaging system and machine learning regression models prove to be a useful alternative for estimating nutrient levels in citrus leaves. These methods offer a non-destructive, cheaper, faster and more accurate solution compared to destructive, costly and time-consuming techniques, thus fulfilling our main objective of efficiency and reduction of environmental impact.

5. Acknowledgments

This work has been partially funded by projects TED2021-130117B-C31 and C33, funded by MCIN/AEI/1 0.13039/501100011033 and by the EU "NextGenerationEU"/PRTR, and projects GVA-IVIA 52203, 52204 and the EU through the ERDF of the Generalitat Valenciana 2021-2027.

Referencias:

1. Geovanny Rambauth-Ibarra, "Agricultura de Precisión: La integración de las TIC en la producción Agrícola", J. Comput. Electron. Sci.: Theory Appl., vol. 3 no. 1 pp. 34-38. January - June, 2022, doi: http://dx.doi.org/1 0,17981/cesta.03.01.2022.04
2. Petra Marschner , "Marschnner's Mineral Nutrition of Higher Plants", (3rd ed.), Elsevier, 2012, doi: https://doi.org/1 0,1016/C2009-0-63043-9
3. Jiao Chen, Shaoyu Lü, Zhe Zhang, Xuxia Zhao, Xinming Li, Piao Ning, Mingzhu Liu, "Environmentally friendly fertilisers: A review of materials used and their effects on the environment", Science of The Total Environment, Volumes 613–614, 2018, Pages 829-839, ISSN 0048-9697, doi: https://doi.org/1 0,1016/j.scitotenv.2017.09.186
4. Rafael Alejandro Casillas-Peñuelas, "Hyperspectral imaging analysis and applications for food quality", CRC Press, 2018, doi: https://doi.org/10.1201/9781315209203
5. Antonio Fazari, Oscar J. Pellicer-Valero, Juan Gómez-Sanchıs, Bruno Bernardi, Sergio Cubero, Souraya Benalia, Giuseppe Zimbalatti, Jose Blasco, "Application of deep

convolutional neural networks for the detection of anthracnose in olives using VIS/NIR hyperspectral images", Computers and Electronics in Agriculture, Volume 187, 2021, 106252, ISSN 0168-1699, doi: https://doi.org/1 0,1016/j.compag.2021.106252

6. Emily Grisanti, Maria Totska, Stefan Huber, Christina Krick Calderon, Monika Hohmann, Dominic Lingenfelser, Matthias Otto, "Dynamic Localized SNV, Peak SNV, and Partial Peak SNV: Novel Standardization Methods for Preprocessing of Spectroscopic Data Used in Predictive Modeling", Journal of Spectroscopy, vol. 2018, Article ID 5037572, 14 pages, 2018, doi: https://doi.org/1 0,1155/2018/5037572
7. Dong, Y., Shan, Y., Li, P., Jiang, L., & Zhang, X, "Nondestructive Characterisation of Citrus Fruit by near-Infrared Diffuse Reflectance Spectroscopy (NIRDRS) with Principal Component Analysis (PCA) and Fisher Linear Discriminant Analysis (FLDA)". Foods, 11(4), 1-16. 2022, doi: https://doi.org/1 0,1080/00032719.2022.2063306
8. Bishop, Christopher M., "Linear Models for Regression" in Pattern Recognition and Machine Learning, New York: Springer, 2006, doi: https://doi.org/1 0,5555/1162264
9. Vinzi, V. E., Chin, W. W., Henseler, J., & Wang, H, "Handbook of Partial Least Squares: Concepts, Methods and Applications", Springer, 2010, doi: https://doi.org/1 0,1007/978-3-540-32827-8
10. Bishop, Christopher M., "Support Vector Machines for Regression" in Pattern Recognition and Machine Learning, New York: Springer, 2006, doi: https://doi.org/1 0,5555/1162264
11. Trevor Hastie, Robert Tibshirani, Jerome Friedman, "Random Forests" in The elements of statistical learning: data mining, inference, and prediction, New York : Springer, Second Edition, 2009, doi: https://doi.org/1 0,1007/978-0-387-84858-7
12. Trevor Hastie, Robert Tibshirani, Jerome Friedman, "Boosting and Additive Trees" in The elements of statistical learning: data mining, inference, and prediction, New York : Springer, Second Edition, 2009, doi: https://doi.org/1 0,1007/978-0-387-84858-7
13. Refaeilzadeh, P., Tang, L., Liu, H., "Cross-Validation " In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, doi: https://doi.org/1 0,1007/978-0-387-39940-9_565
14. Feurer, M., Hutter, F, "Hyperparameter Optimization". In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds) Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham, 2019, doi: https://doi.org/1 0,1007/978-3-030-05318-5_1
15. Kuhn, M., Johnson, K. (2013). Measuring Performance in Regression Models. In: Applied Predictive Modeling. Springer, New York, NY, doi: https://doi.org/10.1007/978-1-4614-6849-3_5