SUPPLEMENTARY MATERIAL

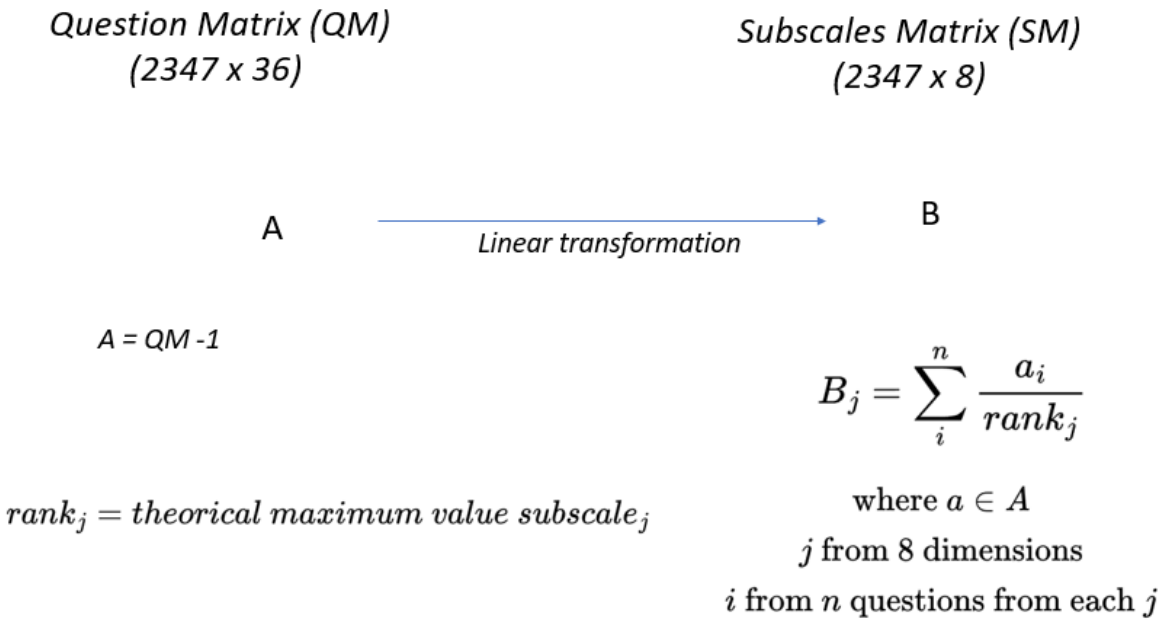
Mathematical arguments based on the work performed and technical specifications on the use of artificial intelligence algorithms

Methods

Linear application of ordered sets

Responses on the SF-36 range from 1 to 6, as shown in **Table 1**, with higher scores corresponding to better health. A subscale is a set of questions related to a specific condition. Therefore, we can define a set of ordered responses for each individual as an ordered 36-dimensional vector. Similarly, subscales define an ordered 8-dimensional vector for each individual. **Figure S1** details the process of computing the set of 8-dimensional vectors of the subscales (matrix B) from the set of 36-dimensional vectors of the responses of SF-36 questionnaires (matrix A) through a linear application. Note that matrix A is rescaled to be in the range between 0 and 5, i.e., the minimum value is 0 instead of 1.

Figure S1.- Diagram of the transition from matrix A to B through a linear application.



While maintaining its properties, this linear application does not guarantee bijectivity since two different 36-dimensional vectors of matrix A could generate the same 8-dimensional vector in matrix B. However, it is relevant to underline that this transformation maintains the properties of the linear application.

The clustering analysis

The clustering analysis was implemented in Python (version 3.7.14). The decoded SF-36 answers and subscale matrices were used and compared. The dimensions were 2347x36 and 2347x8, respectively. To select the optimal number of clusters, some models were fitted with values in the range [2,6] for k (Birch and spectral clustering) by the elbow method. (Bengfort et al. 2022) and the Calinski and Harabasz metric (see **Stable 1**). Three validation metrics are proposed using the scikit-learn package (version 1.0.2) to evaluate the performance of each tested model when the truth labels are unknown:

- Silhouette Coefficient
- Calinski–Harabasz Index (Kozak 2012)
- Davies–Bouldin Index (Halkidi, Batistakis, and Vazirgiannis 2001)

STable 1.- Clustering Algorithms tested

Algorithm	Parameters optimized	Package used
K-means	• Initializer: K-Means++. This method was used to find out optimal initial centers. • Metric distance: Euclidean and Manhattan have been compared. • Optimized using Silhouette score (Rousseeuw 1987) by 200 runs.	Pyclustering (v 0.10.1.2) (Novikov 2019)
Agglomerative	• Links: Centroid, single, complete, and average links were tested, and the maximum silhouette score was chosen	pyclustering
Birch	• Branching_factor and threshold. • Optimized by grid-search using maximum silhouette score.	scikit-learn
DBSCAN	• Eps (The maximum distance between two samples, for one to be considered as in the neighborhood of the other).	scikit-learn

STable 1.- Clustering Algorithms tested

Algorithm	Parameters optimized	Package used
	• Minimum samples. • Left size (this can affect the speed of the construction and query and the memory required to store the tree) using maximum silhouette score.	
K-MEDOIDS	• Initializer: k-medoids++. • Optimized using Silhouette score by 200 runs.	scikit-learn
Fuzzy-C	• Initializer: K Means++. Farthest Centre Candidate option.	pyclustering

Principal Components Analysis

As mentioned before, the set of SF-36 questionnaires consists of 2,347 records or individuals. For each question, we examine the normality and correlation between all pairs of variables. In addition, we use Principal Component Analysis (PCA) to summarize and visualize the interaction between the characteristics. Once the standardization is complete, the analysis is complete and is illustrated by the factor map in **Figure S7**:

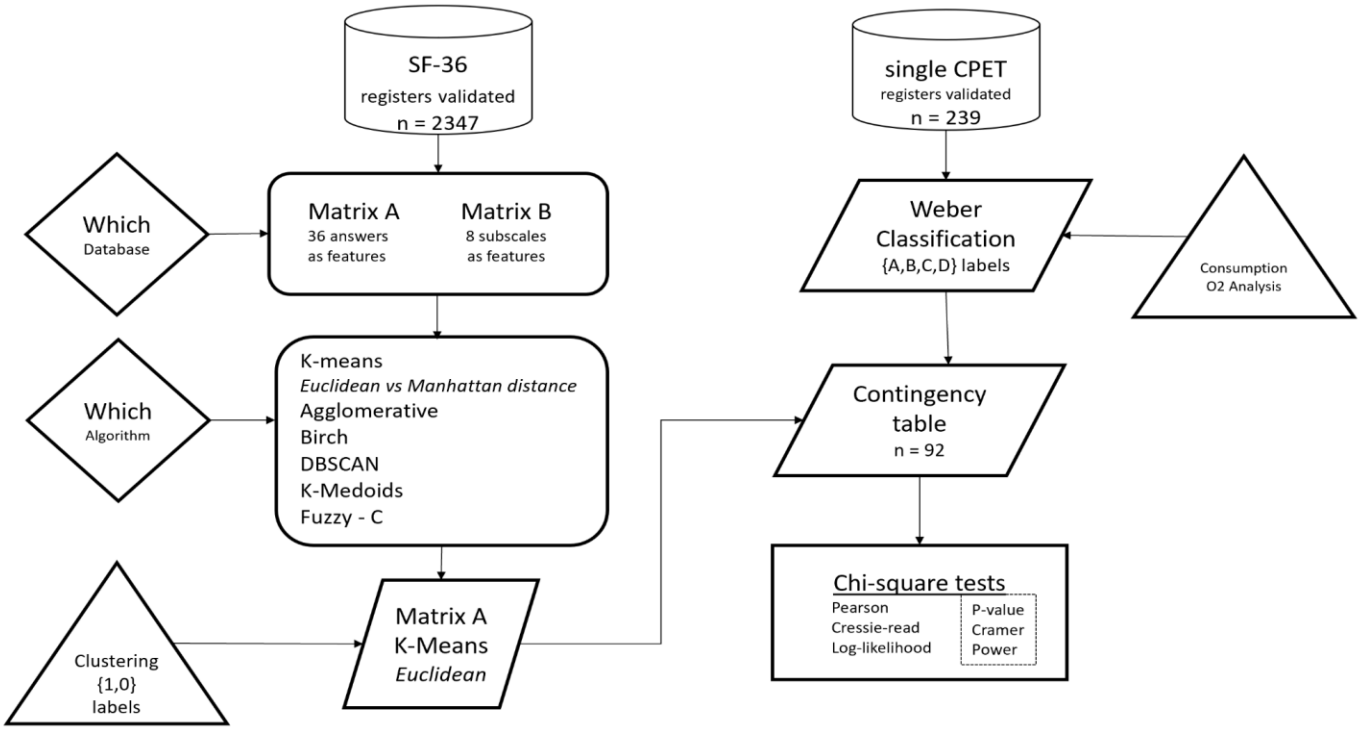
- Eigenvalues and variances for each feature.
- Plot analysis, where positively correlated vectors are grouped in the same quadrant, and negatively correlated vectors are positioned on opposite quadrants.
- Vectors far from the origin (i.e., the center of the coordinate system) are well represented on the factor map.

The cosine square (cos^2) method is computed to measure the quality of each feature, where a high value indicates a good representation of the variable on the principal component and is close to the circumference of the correlation circle (radius equal to 1). This analysis is performed using R (v 4.2.1) with packages factoextra (v 1.0.7) and factominer (v 2.6) (Kassambara 2017).

Contingency table difference analysis

A contingency table was created for patients who completed the single CPET test and the SF-36 questionnaire. We selected 92 patients who met the criteria **(Figure S2)** from 239 single CPET registries. The contingency table is constructed using the clustering labels and Weber's classification based on VO2 levels.

Figure S2.- Steps for the contingency table.

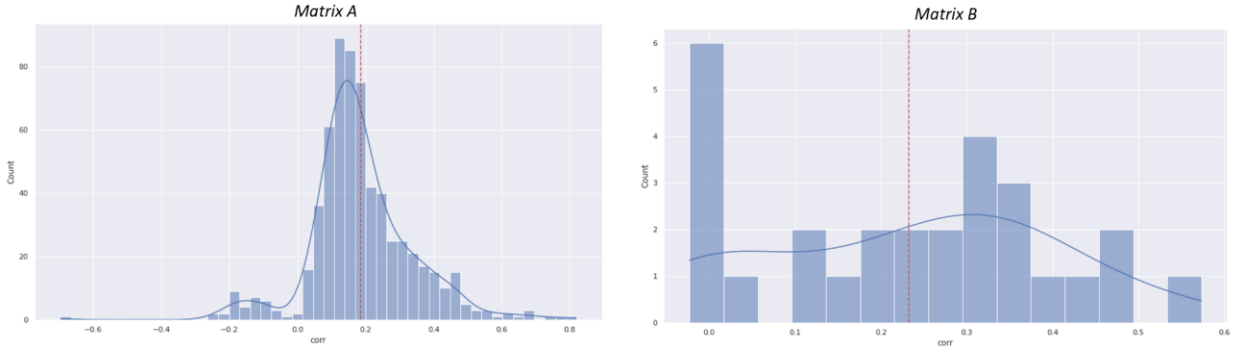


Results

Matrix Analysis

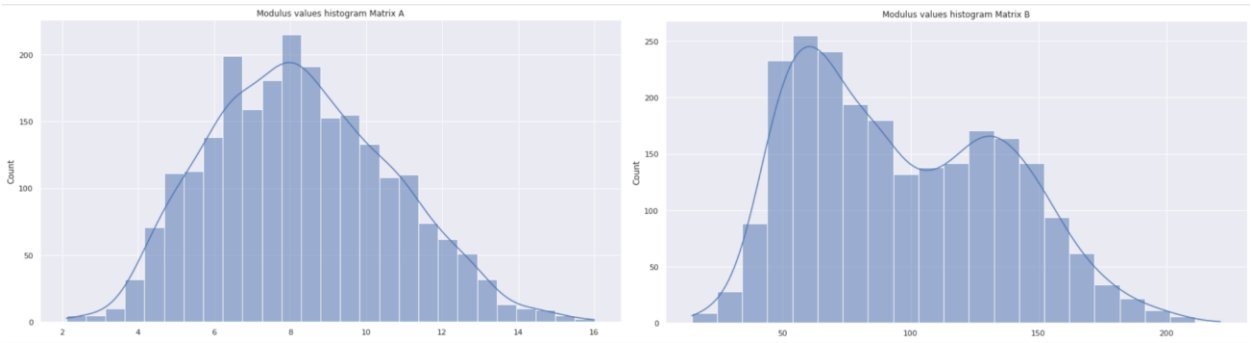
The d'Agostino and Pearson tests are used to analyze the normality of each SF-36 question response and its aggregated subscales (D'agostino & Pearson, 1973). Pearson's linear correlation is calculated between all pairs of variables, and the histogram of the results is shown in **Figure S3**, where the maximum value is 0.8 in the RE1-RE2 pair corresponding to the emotional role. None of the distributions fits a normal distribution.

Figure S3.- Histograms of the series of pairs of linear correlations of variables by matrix. The number of correlations is x/2 (x-1), being x the number of features of each matrix. The red dotted line represents the mean value.



Each data set can be interpreted as a vector, and as mentioned above, it is an ordered set of vectors that represent the state of health of the patient. The modulus of each vector is calculated to create a new ordered set of values. The histograms of these ordered sets are shown below in **Figure S4**.

Figure S4.- Histograms of vector's modulus by matrix.



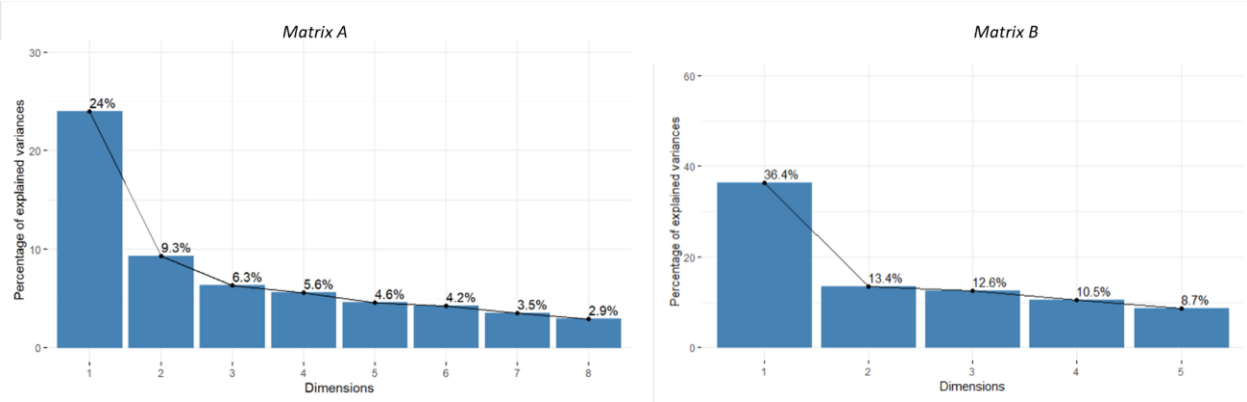
The modulus histograms are not equivalent to matrix A and matrix B. The order of each register was analyzed by percentile performance. If the order is similar, the percentiles would be similar as well. We analyzed the difference in the percentiles of each record in the two matrices, giving an error margin of ± 0.1. We noted that the records with the higher margin of error, 40.14% of the records, indicate a change in the percentile < 0.1 in both matrices, which concludes that the order changes in both matrices. Furthermore, a very differentiated histogram structure is observed, showing a distribution compatible with normality in matrix A and a significant

concentration in lower values (patients with worse health status) between values of 30-70 and less when the modulus value is more significant than 100 in matrix B.

PCA analysis

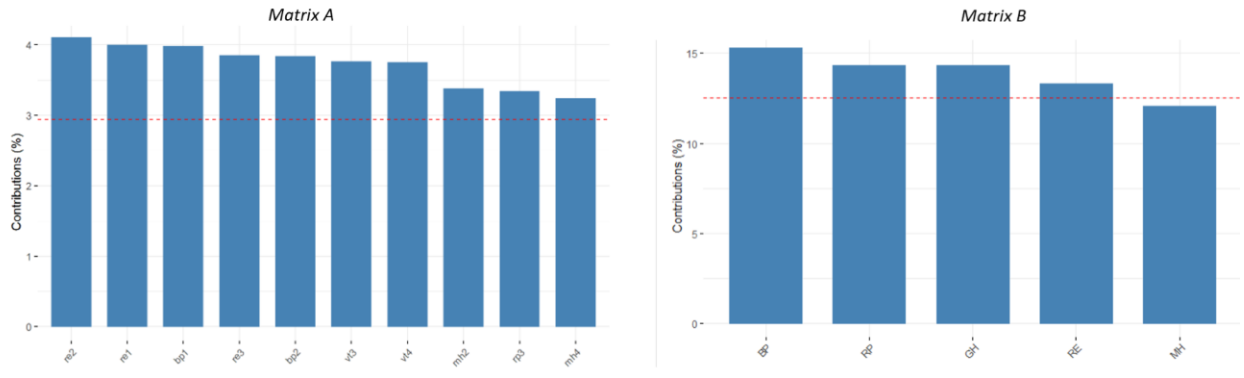
The eigenvalues measure the amount of variation retained by each principal component. The variance contributed by each dimension is calculated with a PCA of 8 components for matrix A and five components for matrix B, as shown in **Figure S5**.

Figure S5.- Percentage of explained variances by matrix.



Matrix A needs eight dimensions or components (reduced from 36) to explain 60.4% of the variance, and matrix B needs five dimensions (reduced from 8) to explain 81.6%. The contributions of these variables for the variability in a given principal component are expressed in percentages. Variables that are correlated with any dimension are the most important in explaining the variability in the dataset. The contribution of variables is shown in **Figure S6**.

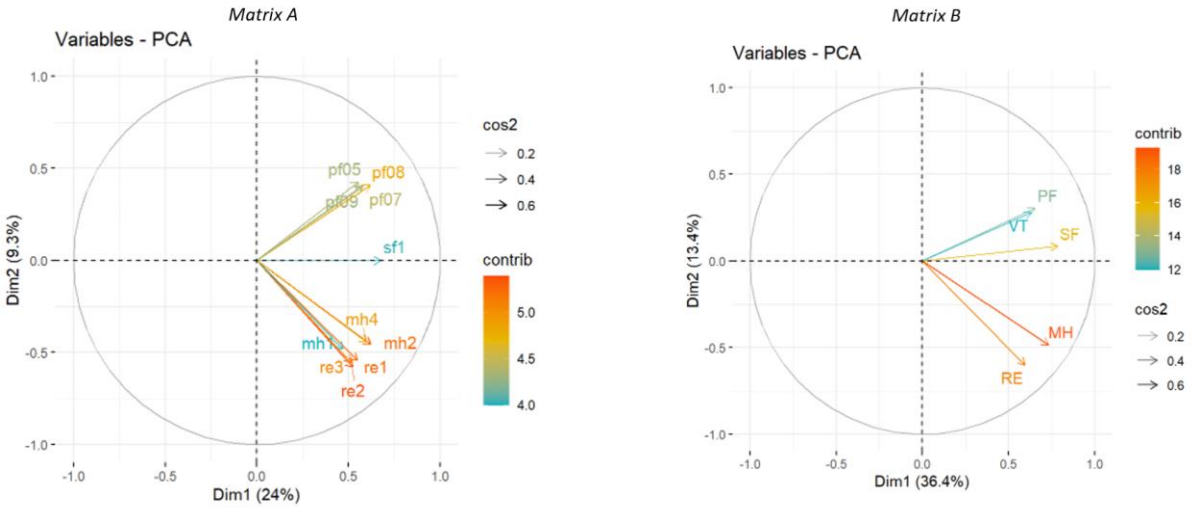
Figure S6.- Contribution of variables by Matrix. The contribution is calculated for an 8-dimensional PCA in matrix A and a 5-dimensional PCA in matrix B. The red dashed line on the graph above indicates the expected average contribution. In matrix A, each column represents the name of each response. In matrix B, each column represents the name of each subscale.



Critical variables have larger contributions. Significant differences are observed between the matrices in the representativeness of the variance of the dimensions. In Matrix A, the variables related to the emotional role (ER) have the most significant influence. On the other hand, in matrix B, somatic pain (BP) contributes the most, with two variables (bp1, bp2) among those that contribute the most in matrix A. The physical part significantly influences Matrix B, with the scales of Body Pain (BP) and Physical Role (PR).

The correlation plot of variables shows the relationship between all pairs of variables, with positively correlated variables grouped in the same quadrant and negatively correlated variables in opposite quadrants. The distance between the variables and the origin measures the quality of the variables, where a greater distance from the origin of the coordinates implies better quality. The cos^2 value indicates the goodness of the variable's representativeness. In **Figure S7**, we depict the most relevant vectors, i.e., the vectors with $cos^2 > 0.45$.

Figure S7.- Correlation plot of variables. Note that the display takes the two main dimensions of each matrix, which contribute 33.3% of the variance in the case of matrices A and 39.8% in the case of matrices B. The color indicates the contribution to the variance. The color indicates the contribution to the variance. The redder the color, the more significant the variance contributed. Each column represents the name of each response in Matrix A. Each column represents a subscale name in Matrix B.



The scales and the physical variables they contain are divided into two groups. In both matrices, Emotional Roles (ER) and Mental Health (MH) are correlated and symmetrically arranged in the first and fourth quadrants of the figure, and so are the more physical parts of the questionnaire (bodily function and vital capacity). The correlation of variances should not be confused with Pearson's linear correlation. These are different concepts in this analysis.

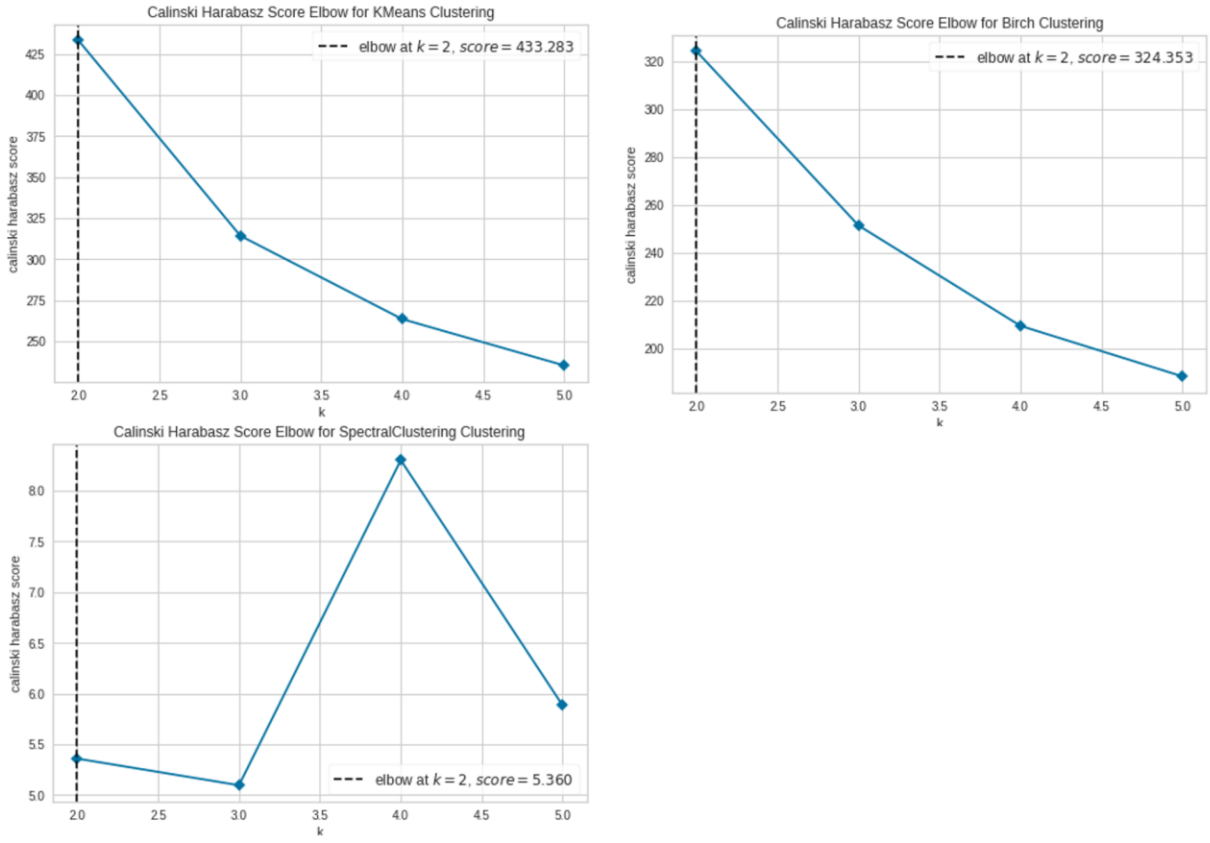
Clustering Analysis

Hopkins clustering test (Lawson & Jurs, 1990) is used to assess the clustering tendency of a data set by measuring the probability that a uniform data distribution generates a given data set. The data is not uniformly distributed if the test is positive (Hopkins score $\simeq 0$). Hence, clusterings can help to classify the observations. However, if the score is too high (≥ 0.5), the data is uniformly distributed, and clustering cannot solve the problem. The Hopkins test scores for both matrices are 0.3045 (matrix A) and 0.2509 (matrix B); consequently, they can be classified.

Determination of the number of clusters.

The Calinski-Harabasz metric has been used to determine the optimal number of clusters. It is tested with three algorithms (K-means, Birch, and Spectral) whose results are consistent and are shown in **Figure S8**.

Figure S8.- Graphs of the determination of the number of clusters of the three algorithms. The graphs represent the optimal number of clusters to be analyzed, and in all algorithms, k=2 is the best result.



Silhouette and Calinski-Harabasz have optimized all algorithm parameters. These metrics are for internal validation since the labels are unknown at the outset. They measure the compactness (distance between objects in the same group) and separation (distance among different clusters or groups). The best scoring algorithms (Agglomerative and DBSCAN) lack functional partitioning because they virtually eliminate a cluster. **Table S2** and **Table S3** show that label 0 is composed of modulus values with lower responses in the remaining algorithms, and therefore the health status is worse. Thus, K-means Euclidean distance manages to separate up to an average of 6.99 in 1,407 records in a cluster. As for the metrics, there is also no unanimity regarding the best algorithm since Birch but if we classify by Calinski-Harabasz, Birch would be far behind K-means or Fuzzy-C. No conclusive study defines one metric better than another (Xiong & Li, n.d.); in our case, one must choose which matrix to use.

Table S2.- Results from Clustering Algorithms Matrix A. Validations metrics: Scores of each metric. Labels count: Number of items of each class. Modulus mean: mean for each class of modulus values.

	Validations metrics			Labels Count		Modulus mean	
Algorithm	Silhouette	Calinski - Harabasz	Davies Bouldin	Label 0	Label 1	Mod 0	Mod 1
K-means Euclidean	0.17	433	2.24	1,407	940	6.99	10.21
K-means Manhattan	0.20	394	2.15	1,751	596	7.43	1.80
Agglomerative	0.50	32	0.92	2,341	6	8.27	14.00
Birch	0.21	284	2.46	1,861	486	7.72	10.46
DBSCAN	0.52	12	0.89	2,345	2	8.28	14.29
K-Medoids	0.18	404	2.27	1,535	812	7.07	10.58
Fuzzy - C	0.15	426	2.47	1,237	1110	6.73	10.01

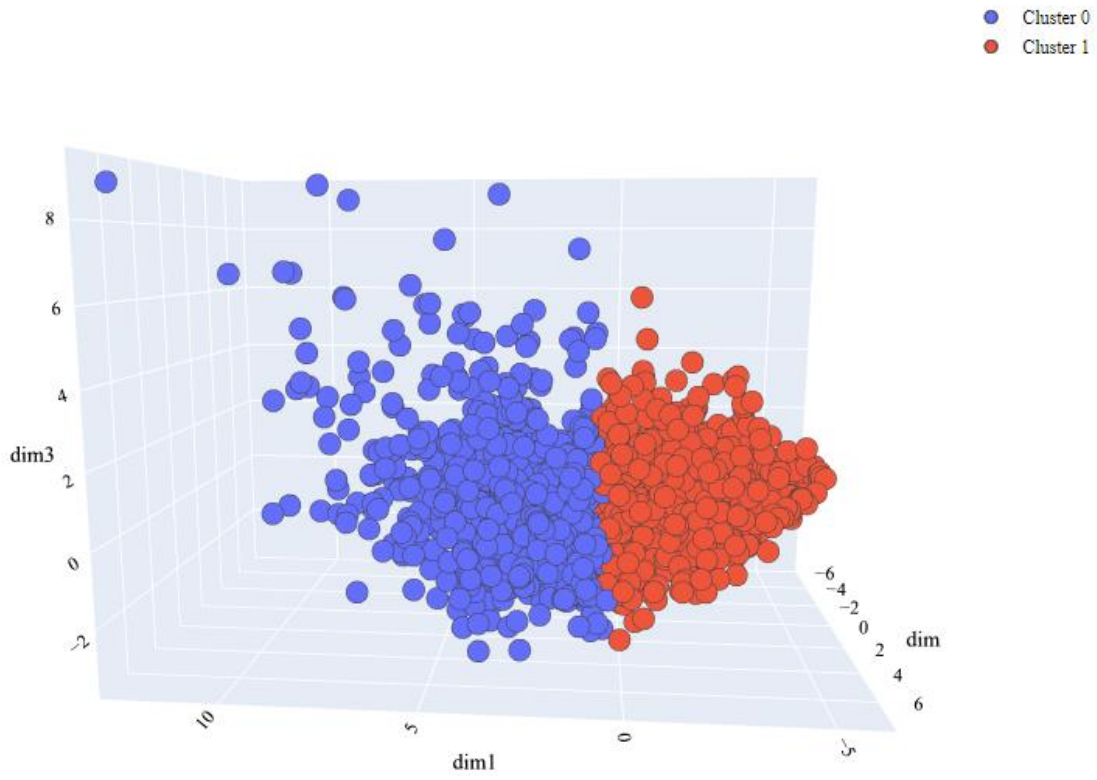
Table S3.- Results from Clustering Algorithms Matrix B. Validations metrics: Scores of each metric. Labels count: Number of items of each class. Modulus mean: mean for each class of modulus values.

	Validations metrics			Labels Count		Modulus mean	
Algorithm	Silhouette	Calinski - Harabasz	Davies Bouldin	Label 0	Label 1	Mod 0	Mod 1
K-means Euclidean	0.24	739	1.69	1389	958	70.23	135.43
K-means Manhattan	0.26	667	1.65	635	1712	144.23	79.27
Agglomerative	0.60	11	0.28	2346	1	96.80	207.82
Birch	0.29	252	1.61	2180	167	94.46	128.01
DBSCAN	0.55	20	0.85	2344	3	96.73	183.95
K-Medoids	0.24	717	1.68	814	1533	138.37	74.80

Fuzzy - C	0.23	736	1.69	1295	1052	67.94	132.44

It can be observed that there are small differences between the scores of the two matrices. For the purpose of comparison, the order of the scores for each metric is ranked, and the lower score is added to obtain the better score, taking into account that neither DBSCAN nor Agglomerative is a valid alternative. With K-means, Matrix A could be a good alternative. With Matrix B, it would be Birch, although Fuzzy-c should be analyzed from a different perspective. Calinski-Harabasz performed well in the study of internal validation metrics (Xiong & Li, n.d.). This metric is less sensitive to noise and the different shapes it can take in an n-dimensional field. It considers the difference between the mean module values of the two clusters, making using the K-means algorithm optimum on the 36-dimensional database. The result obtained is shown in **Figure S9**, where the colored cluster represents each item. The three-dimensional values obtained from a PCA on the original database define the three-dimensional representation.

Figure S9.- 3-D clustering plot K-means Euclidean distance with matrix A. The three axes have been defined using three-dimensional PCA for visualization. Each item is assigned a cluster by color.



REFERENCES

Bengfort, Benjamin, Larry Gray, Rebecca Bilbro, Prema Roman, Patrick Deziel, Kristen McIntyre, Molly Morrison, et al. 2022. *Yellowbrick v1.5*. https://doi.org/10.5281/zenodo.7013541.

D'agostino, Ralph, and E. S. Pearson. 1973. "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √b1." *Biometrika* 60 (3): 613–22.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. 2001. "On Clustering Validation Techniques." *Journal of Intelligent Information Systems* 17 (2): 107–45.

Kassambara, Alboukadel. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA.

Kozak, Marcin. 2012. "'A Dendrite Method for Cluster Analysis' by Caliński and Harabasz: A Classical Work That Is Far Too Often Incorrectly Cited." *Communications in Statistics - Theory and Methods* 41 (12): 2279–80.

Lawson, Richard G., and Peter C. Jurs. 1990. "New Index for Clustering Tendency and Its Application to Chemical Problems." *Journal of Chemical Information and Computer Sciences* 30 (1): 36–41.

Novikov, Andrei. 2019. "PyClustering: Data Mining Library." *Journal of Open Source Software* 4 (36): 1230.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (November): 53–65.

Vallat, Raphael. 2018. "Pingouin: Statistics in Python." *Journal of Open Source Software* 3 (31): 1026.

Xiong, and Li. n.d. "Clustering Validation Measures." *Data Clustering*. https://doi.org/10.1201/9781315373515-23/clustering-validation-measures-hui-xiong-zhongmou-li.