PATIENT CONDITION PREDICTION MODELING

Dr Lamrous Sid Sid.Lamrous@utbm.fr
BABA Tchao A. tchaobaba424@gmail.com

1 - Motivation and objective

Goal

- Develop a model for efficient real-time patient monitoring and delivering high-quality personalized healthcare..

Motivation

- Addressing growing medical knowledge, evolving diseases, and COVID-19 challenges for enhanced patient monitoring and personalized care..

Problem

- Increase in chronic diseases

2 - Medical data Analysis

DATASET

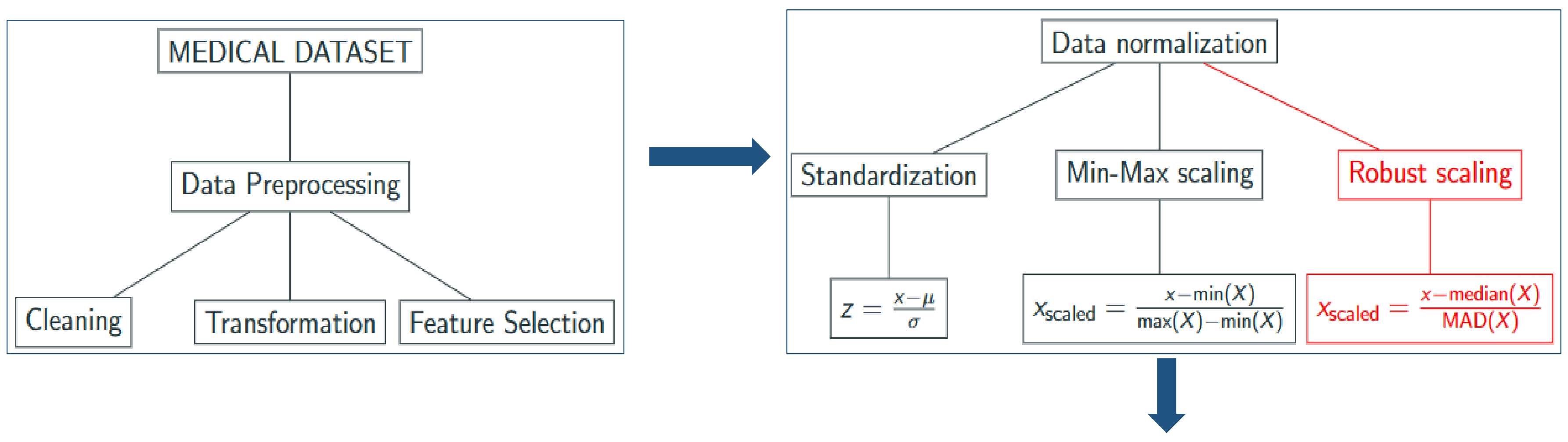
The study utilized the health data of 30 patients who were hospitalized during the COVID-19 period. Regular analyses were conducted on these patients to monitor the progression of their health condition.

Activité par jour	Créatinine	Distance par jour	Fréquence cardiaque
Glycémie	Kaliémie	Natrémie	Poids
Pression artérielle diastolique	Pression artérielle systolique	Saturation en oxygène	Sommeil par jour
Taux d'albumine	Température	EVS	EVA

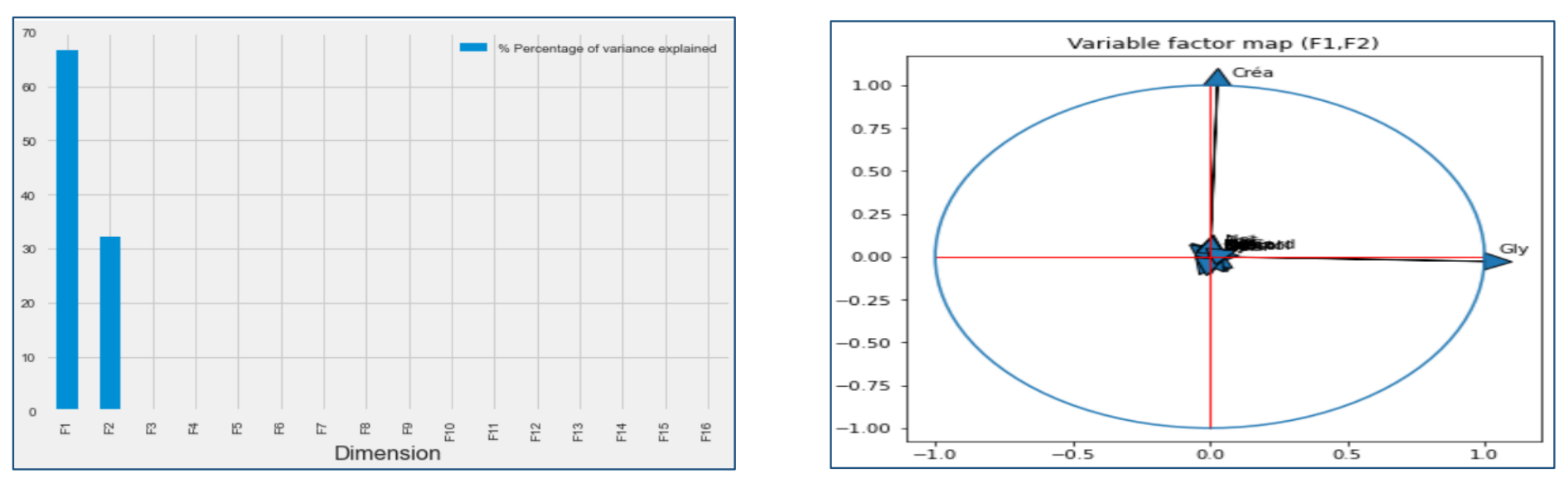
PREPROCESSING

Idea

Preparation and cleaning of patient data to enhance the quality and relevance of subsequent analysis and treatment outcomes.


MEDICAL DATASET
Data Preprocessing
Cleaning
Transformation
Feature Selection
Data normalization
Standardization
Min-Max scaling
Robust scaling
$z = \frac{x-\mu}{\sigma}$
$X_{\text{scaled}} = \frac{x-\min(X)}{\max(X)-\min(X)}$
$X_{\text{scaled}} = \frac{x-\text{median}(X)}{\text{MAD}(X)}$

Dimension reduction (PCA)



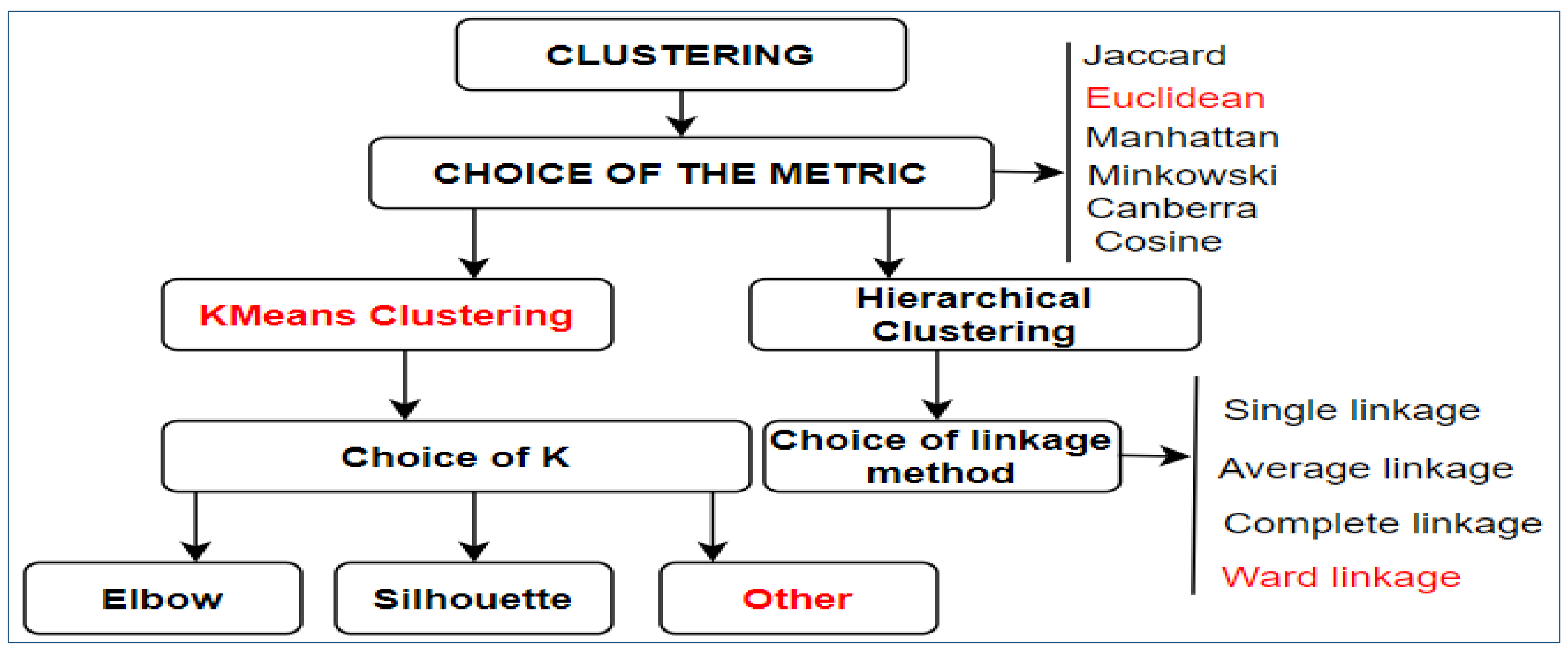
Applying PCA to the data allows for simplifying their representation, identifying the most important variables, facilitating visualization, and enhancing the performance of subsequent analysis techniques.

PROCESSING

Clustering :

Cluster similar patients into homogeneous subgroups. The aim is to identify hidden structures and patterns in the patients' data.

The study bellow chose the Euclidean distance as the metric to assess the dissimilarity between patients' analysis data. This decision aims to achieve a good clustering outcome, enabling the grouping of similar patients based on their data patterns.


CLUSTERING
CHOICE OF THE METRIC
Jaccard
Euclidean
Manhattan
Minkowski
Canberra
Cosine
KMeans Clustering
Hierarchical Clustering
Choice of K
Choice of linkage method
Single linkage
Average linkage
Complete linkage
Ward linkage
Elbow
Silhouette
Other

Choice of the metric

$$\text{Average silhouette score} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\right)$$

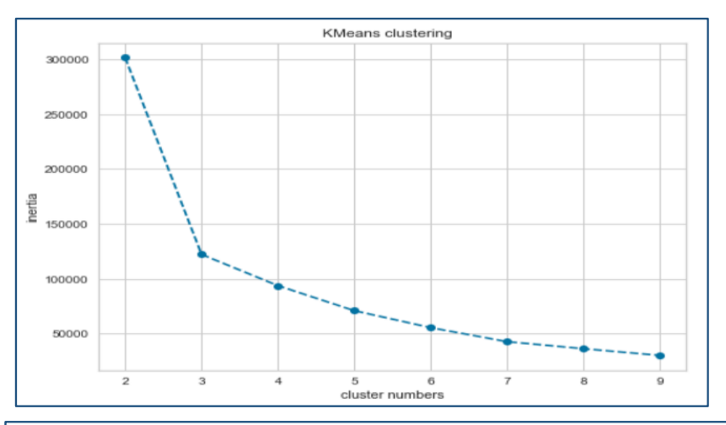
	K = 2	K = 3	K = 4	K = 5	K = 6
Euclidean $d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$	0.74	0.77	0.73	0.74	0.66
Manhattan $d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n}\lvert q_i-p_i\rvert$	0.71	0.70	0.66	0.67	0.60
Minkowski $d(\mathbf{p},\mathbf{q}) = \left(\sum_{i=1}^{n}\lvert q_i-p_i\rvert^p\right)^{\frac{1}{p}}$	0.74	0.77	0.73	0.74	0.66
Canberra $d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n}\frac{\lvert q_i-p_i\rvert}{\lvert q_i\rvert+\lvert p_i\rvert}$	0.09	0.09	0.09	0.06	0.06
Jaccard $d(A,B) = 1 - \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	0.11	0.04	0.03	- 0.03	- 0.01
Cosine $d(\mathbf{u},\mathbf{v}) = 1 - \frac{\mathbf{u}\cdot\mathbf{v}}{\lVert\mathbf{u}\rVert\,\lVert\mathbf{v}\rVert}$	0.14	0.14	0.03	0.06	0.032

Kmeans clustering

Choice of K

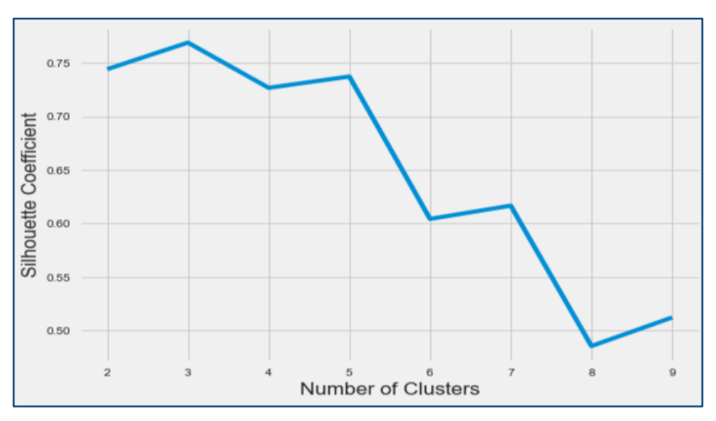
Elbow

$$\text{inertia} = \sum_{i=1}^{n}\sum_{j=1}^{k}\lVert x_i - c_j\rVert^2$$



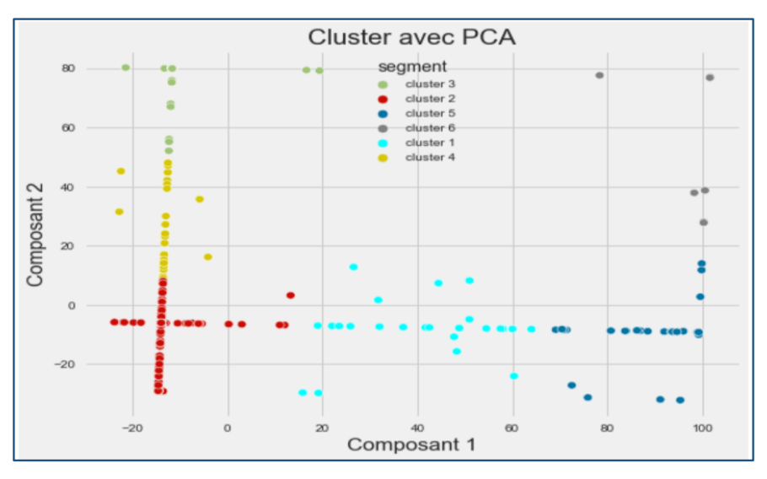
Silhouette

$$\text{Average silhouette score} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\right)$$



Although the elbow and silhouette methods suggested 3 as the optimal number of clusters for is to enable doctors to track, in real-time, even subtle change clustering, we performed clustering into 6 groups. The aims in patients' health conditions.

Kmeans clustering with K = 6



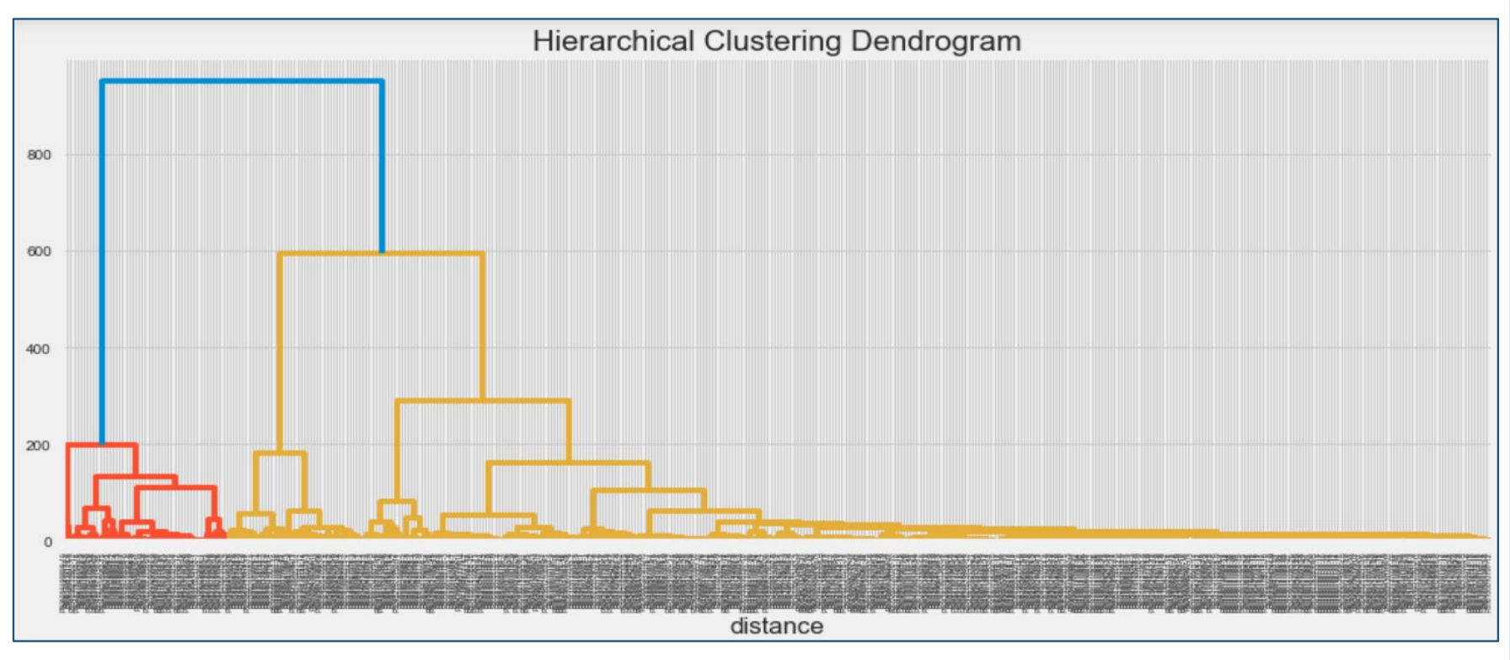
For the rest of the study, only the results of the K-means clustering will be considered.

Hierarchical clustering

Choice of the linkage method

Single linkage	$d_{C_i,C_j} = min_{x,y}\{D(x,y) \mid x \in C_i, y \in Cj\}$
Average linkage	$d_{C_i,C_j} = mean_{x,y}\{D(x,y) \mid x \in C_i, y \in Cj\}$
Complete linkage	$d_{C_i,C_j} = max_{x,y}\{D(x,y) \mid x \in C_i, y \in Cj\}$
Ward's method	$d_{C_i,C_j} = \frac{m_i m_j}{m_i+m_j}\lVert c_i - c_j\rVert^2$

Following the aforementioned study, which aimed to determine the most suitable similarity criterion for hierarchical clustering, the Ward method was ultimately selected. This method enables the formation of clusters that are both compact and homogeneous.



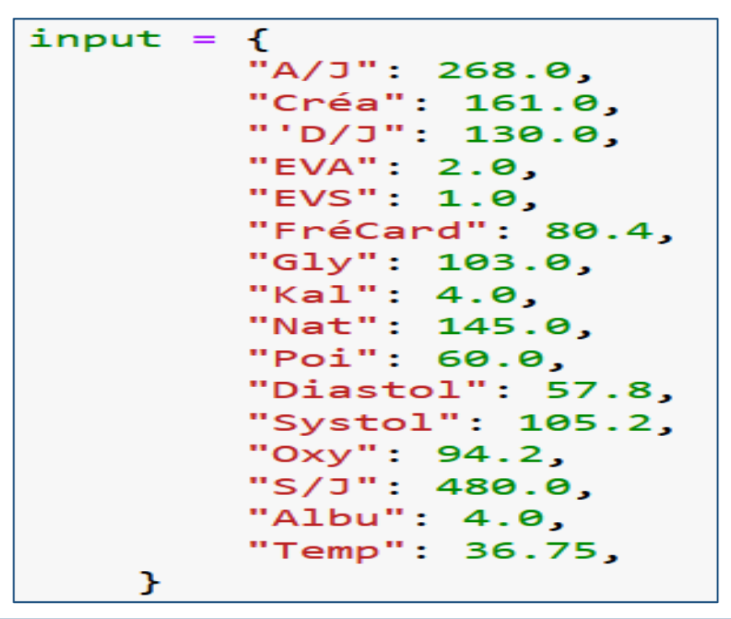
3 - Clustering automation (Supervised learning)

Goal

Assisting doctors in classifying patients based on their test results.

	Formula	X train score	X test score
Logistic Regression	$P(y=1\mid\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}},$	97.14%	76.92%
Random Forest Classifier	$\hat{y}_i = \arg\max_j \sum_{k=1}^{n_{\text{estimators}}} I(x_i \in T_k)\cdot p_{jk},$	100%	96.15%
KNeighbors Classifier	$\hat{y}_i = \arg\max_j \sum_{\substack{k=1 \\ d(x_i,x_k)\leq r}}^{n_{\text{neighbors}}} I(y_k = j),$	99.04%	100%
Extra Trees Classifier	$\hat{y}_i = \arg\max_j \sum_{k=1}^{n_{est}} I(x_i \in R_k)\cdot p_{jk},$	100%	88.46%
LGBMClassifier	$\hat{y}_i = \arg\max_j \sum_{k=1}^{T} w_k \cdot I(x_i \in R_{jk}),$	100%	96.15%

At the end of this study, the LGBM classifier has been chosen to assist in the automatic classification of patients.



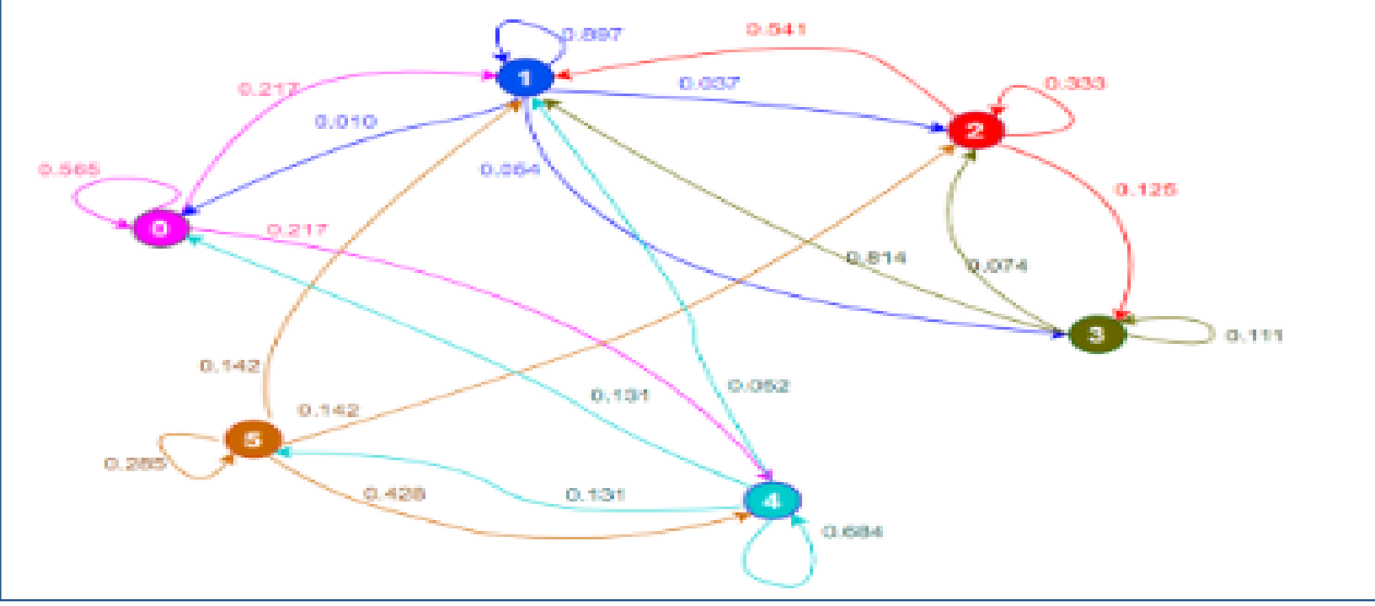
```
input = {
        "A/J": 268.0,
        "Créa": 161.0,
        "'D/J": 130.0,
        "EVA": 2.0,
        "EVS": 1.0,
        "FréCard": 80.4,
        "Gly": 103.0,
        "Kal": 4.0,
        "Nat": 145.0,
        "Poi": 60.0,
        "Diastol": 57.8,
        "Systol": 105.2,
        "Oxy": 94.2,
        "S/J": 480.0,
        "Albu": 4.0,
        "Temp": 36.75,
    }
```

```
Predicted Class: 2
Probability for Class 2: 1.00
```

4 - Predictions with Markov Process

TRANSITION GRAPH AND TRANSITION MATRIX

```
[ [2,1,2,2,1], [4,4,4,4,4,1,1,1], [4,5,4,0,1], [0,0,0,0,4,4,4,4,4,4,4,4,0,1], [2,3,3,2,1],
[1,1,1,1,1,1,1,1,1,1,1,1,1], [1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,1,1,1,1,1,1,1,1,1], [3,1,1,1,3],
[1,1,1,1,1,1,1,1,1,1,1], [3,1,3,1,1,1,1,1,1], [1,0,1,1,1,1,1,1,1,1,1,1,1],
[1,3,1,1,1,3,1,1,1,1,1,1,1,1,1,1,1,1,1], [1,2,3,1,1,1,1,1,1,1,1,1,1,3,1,3,1],
[4,5,5,4,4,0,4,4,1,1,1,1,1,0,0,0,4,4,4,4,4,4,4,4,4,4,5,4,4,4,0,1], [1,1,2,2,2,1,1,1,1,2,1,1,1,1],
[3,1,1,1,1,1,1,1,1,1], [1,1,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,1,1,3,3,1,1,3,1,1,1,1,1,1],
[3,3,1,3,1,1,3,1,1,1,2,2,1,2,1], [3,1,1,1,1,1,1,1,1,1,1,2,1,1,1,3,1,1,1,1],
[1,1,1,1,1,1,1,1,1,1,1,1], [2,2,1,2,1,1,2,1,3,1,1], [2,2,2], [0,1,0,0,0,4,4,4,4,5,1,1,1,1],
[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1], [2,3,1,3,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1],
[3,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1], [1,3,1,1,1,1,1,1], [0,0,0,0,0,0,4,0,0],
[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,2] ]
```



$$M = \begin{bmatrix} 0.565 & 0.217 & 0.000 & 0.000 & 0.217 & 0.000 \\ 0.016 & 0.892 & 0.039 & 0.050 & 0.000 & 0.000 \\ 0.000 & 0.529 & 0.352 & 0.117 & 0.000 & 0.000 \\ 0.000 & 0.777 & 0.111 & 0.111 & 0.000 & 0.000 \\ 0.131 & 0.052 & 0.000 & 0.000 & 0.684 & 0.131 \\ 0.000 & 0.142 & 0.142 & 0.000 & 0.428 & 0.285 \end{bmatrix}$$

TRANSITION MATRIX PROPERTISES

M is a stochastic, regular, irreducible, and diagonalizable matrix. Its largest eigenvalue is equal to 1. According to the Perron-Frobenius theorem, the Markov Process with a transition matrix M has a unique steady state. This steady state is represented by the normalized form of the eigenvector associated with the eigenvalue 1.

Sum of row 1: $0.565+0.217+0.000+0.000+0.217+0.000 = 1.000$
Sum of row 2: $0.010+0.897+0.037+0.054+0.000+0.000 = 1.000$
Sum of row 3: $0.000+0.541+0.333+0.125+0.000+0.000 = 1.000$
Sum of row 4: $0.000+0.814+0.074+0.111+0.000+0.000 = 1.000$
Sum of row 5: $0.131+0.052+0.000+0.000+0.684+0.131 = 1.000$
Sum of row 6: $0.000+0.142+0.142+0.000+0.428+0.285 = 1.000$

$$M^4 = \begin{bmatrix} 0.1742 & 0.4624 & 0.0297 & 0.0269 & 0.2580 & 0.0486 \\ 0.0195 & 0.8566 & 0.0550 & 0.0607 & 0.0072 & 0.0007 \\ 0.0149 & 0.8538 & 0.0640 & 0.0637 & 0.0032 & 0.0001 \\ 0.0167 & 0.8605 & 0.0566 & 0.0615 & 0.0041 & 0.0002 \\ 0.1585 & 0.2933 & 0.0345 & 0.0166 & 0.4041 & 0.0929 \\ 0.1015 & 0.4393 & 0.0504 & 0.0312 & 0.3036 & 0.0737 \end{bmatrix}$$

$$\left\{\begin{array}{ll} \boxed{\lambda_1 \quad 1.000} \\ \lambda_2 & 0.865 \\ \lambda_3 & 0.501 \\ \lambda_4 & 0.301 \\ \lambda_5 & 0.162 \\ \lambda_6 & 0.045 \end{array}\right\} \longrightarrow \pi = \begin{pmatrix} 0.027 \\ 0.83 \\ 0.054 \\ 0.058 \\ 0.024 \\ 0.004 \end{pmatrix}$$

DAILY PATIENT DISTRIBUTION OVERVIEW

Goal

Optimize patient flow and resource management.

$$P(i_K, i_{K-1}, i_{K-2}, \ldots, i_1, i_0) = P(i_K \mid i_{K-1}, i_{K-2}, \ldots, i_1, i_0) \times P(i_{K-1} \mid i_{K-2}, i_{K-3}, \ldots, i_1, i_0) \times \cdots \times P(i_0)$$

According to Markov's property, one can write

$$P(i_k \mid i_{k-1}, i_{k-2}, \ldots, i_1, i_0) = P(i_k \mid i_{k-1})$$

one obtains:

$$P(i_K, i_{K-1}, i_{K-2}, \ldots, i_1, i_0) = P(i_0)\prod_{k=0}^{K-1} P(i_{k+1} \mid i_k)$$

Thus, the following iteration is deduced:

$$\pi_1 = \pi_0 \times M$$
$$\pi_2 = \pi_1 \times M$$
$$\ldots$$
$$\pi_K = \pi_{K-1} \times M$$

FORCAST PATIENT HEALTH CHANGES

Goal

Facilitates proactive interventions and timely treatment adjustments for better patient care.

For $k = 1$: $P(X_n = j \mid X_{n-1} = i) = P_{j,i}(1) = M[i,j]$ $(i, j \in 0,1,2,3,4,5)$
Elements are taken directly in matrix M: $P_{3,4}(1) = M[4][3] = 0.000$

For k greater than 1 : $P(X_n = j \mid X_{n-k} = i) = P_{j,i}(k)$ $(i, j \in 0,1,2,3,4,5)$
The final probability can be obtained by calculating the sum of the probabilities of all paths of length k that start from summit 3 and end at summit 4 on the weighted graph. These probabilities are then multiplied together.

$$P_{3,4}(2) = P_{3,2} \times P_{2,4} + P_{3,4} \times P_{4,4} + P_{3,3} \times P_{3,4} = M[2][3] \times M[4][2] + M[4][3] \times M[4][4] + M[3][3] \times M[4][3]$$
$$P_{3,4}(2) = 0.125 \times 0.000 + 0.000 \times 0.684 + 0.111 \times 0.000 = 0$$

$P_{3,4}(2) = 0$ is confirmed on weighted graph; there is no 2-step path to join cluster 4 from cluster 3.

FORCAST PATIENT TRAJECTORY

Goal

Facilitates personalized care plans and proactive. measures

For example, to calculate the probability of path [2,1,2,3],
The result is as follows:

$$P([2,1,2,3]) = P(2 \mid 1,2,3) \times P(1 \mid 2,3) \times P(2 \mid 3) \times P(3)$$

Applying the Markov property, we obtain the following result:

$$P([2,1,2,3]) = P(2 \mid 1) \times P(1 \mid 2) \times P(2 \mid 3) \times P(3) = M[1,2] \times M[2,1] \times M[3,2] \times \pi[3]$$

Upon examining the values in M, we obtain the following result:

$$P([2,1,2,3]) = 0.037 \times 0.541 \times 0.074 \times 0.058 = 0.0000859$$

5 - Limitations and Perspectives

- Independence assumption neglects key factors impacting health outcomes.
- Assumption of stationarity in Markov models conflicts with the dynamic nature of healthcare

Using advanced modeling techniques such as non-stationary hidden Markov processes or regime-switching Markov models.