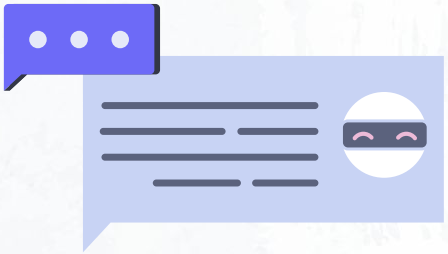
UnSupervised Learning Exam

Università degli Studi di Milano-Bicocca

28-06-2023

Monaco Filippo – 840089
Picione Marco - 827116



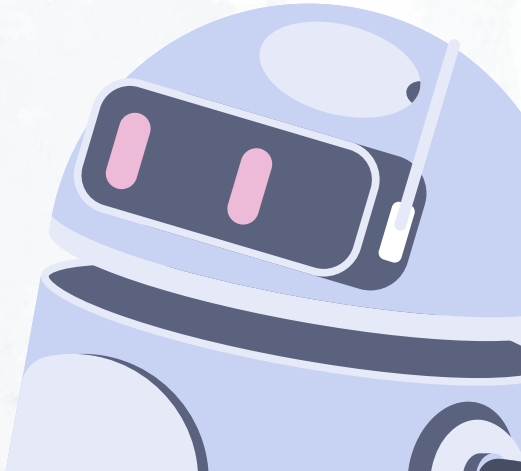


Table of contents

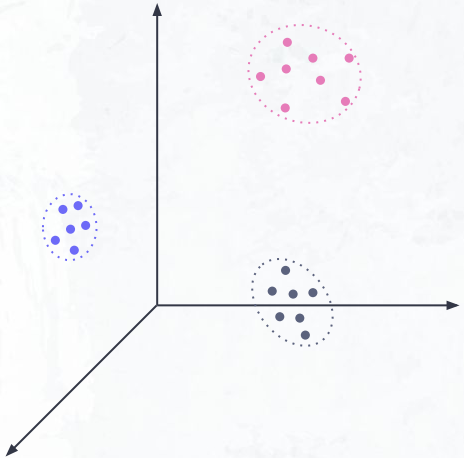
00

Introduction

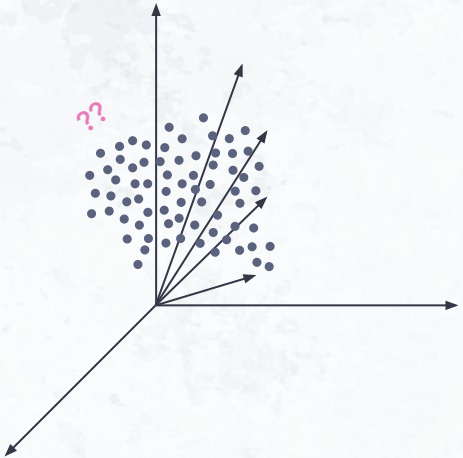
AIM: Clustering of the provided medical dataset
MODELS: Hierarchical & k-prototypes

Cluster Analysis

Group data points in clusters so that objects in a group are similar to one another, and different from the objects in other groups.



Usually more like ...



01

Data Pre-processing

a. Dataset exploration
b. Feature Selection
c. Distance Matrix
d. Outliers Detection

a. Dataset Exploration

Diabetes Prediction Kaggle Challenge (Nov 2022).

40108 objects, 18 attributes each (3 target variables).

Mixed data types:

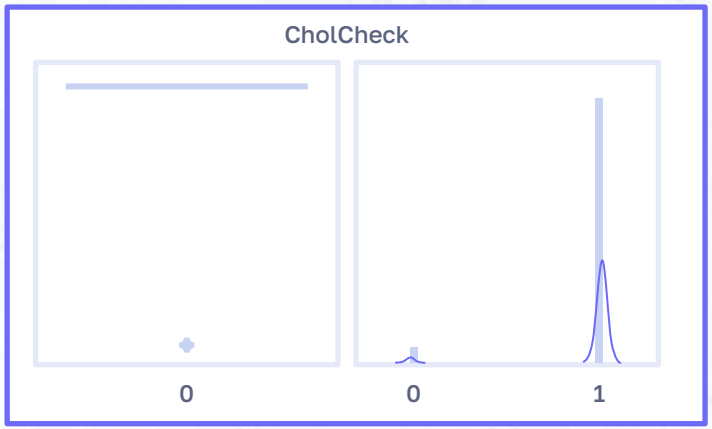
- Nominal: Sex, HighChol, CholCheck, Smoker, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlchoholConsump, DiffWalk
- Ordinal : Age, GenHlth
- Ratio : BMI, MenHlth, PhysHlth

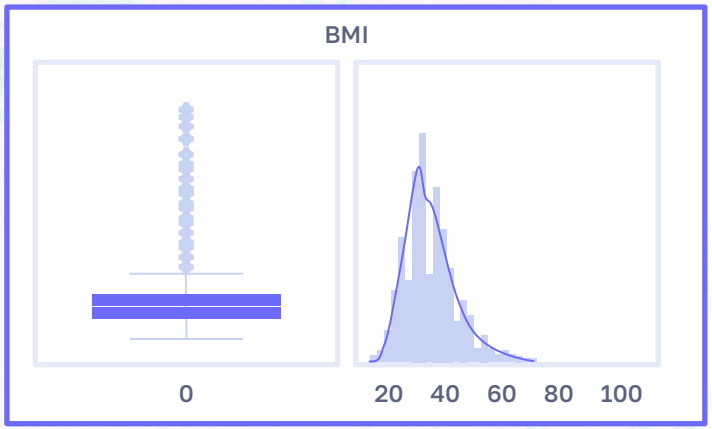
No missing values, 2456 duplicate objects.

Sub-sampling (75%) to reduce computational times and RAM usage.

b. Feature Selection

Study univariate distributions, and discard irrelevant features.

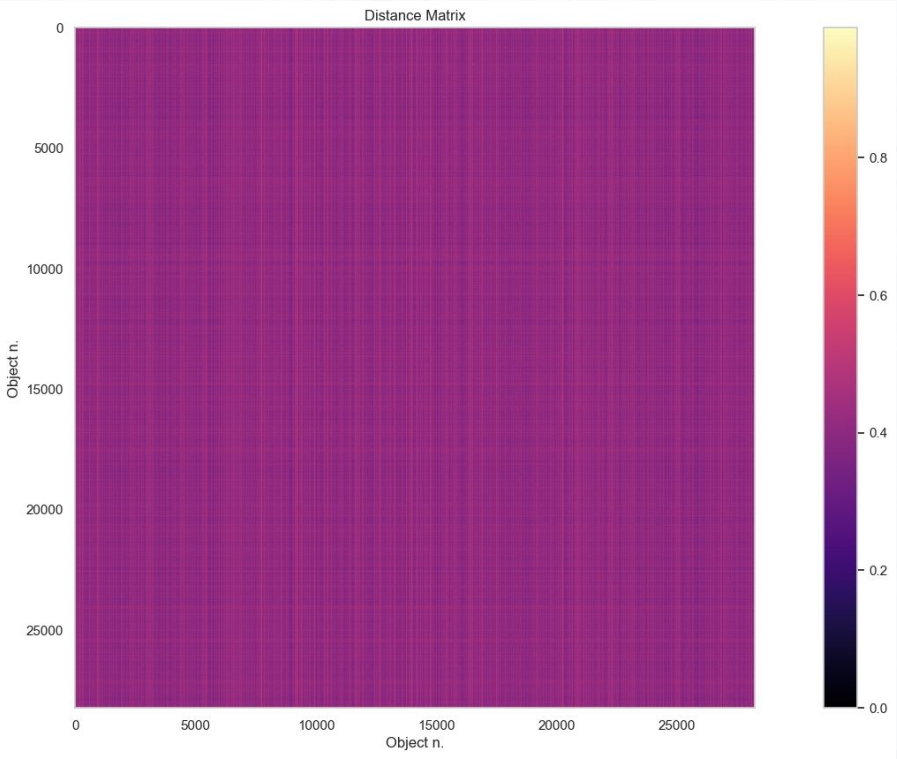




c. Distance Matrix

Gower distance works with mixed data sets to compute distance between points:

- consider each attribute separately
- compute distance one by one using specific distance
- average over all features

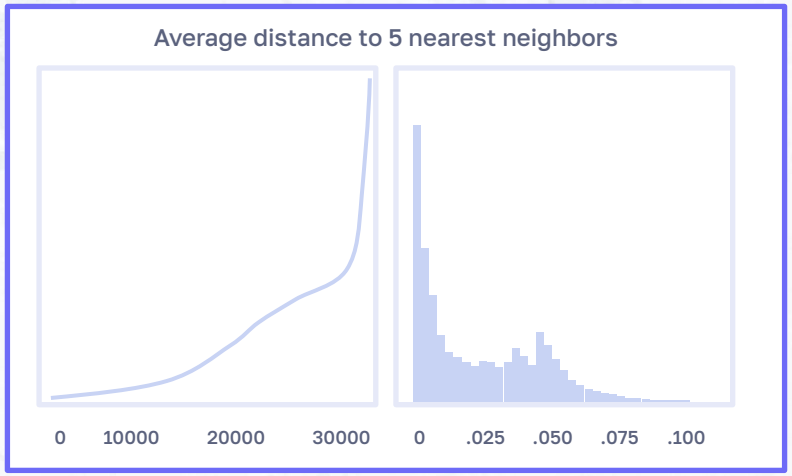


Distance Matrix

- Too many objects
- Too homogeneous
- No single row/column stands out

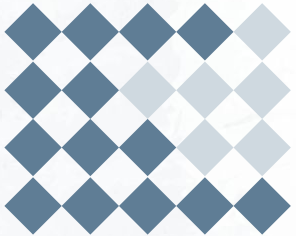
d. Outlier Detection

Consider k nearest neighbors of each point.
Objects that have much higher average distance to neighbors are outliers and removed.



02

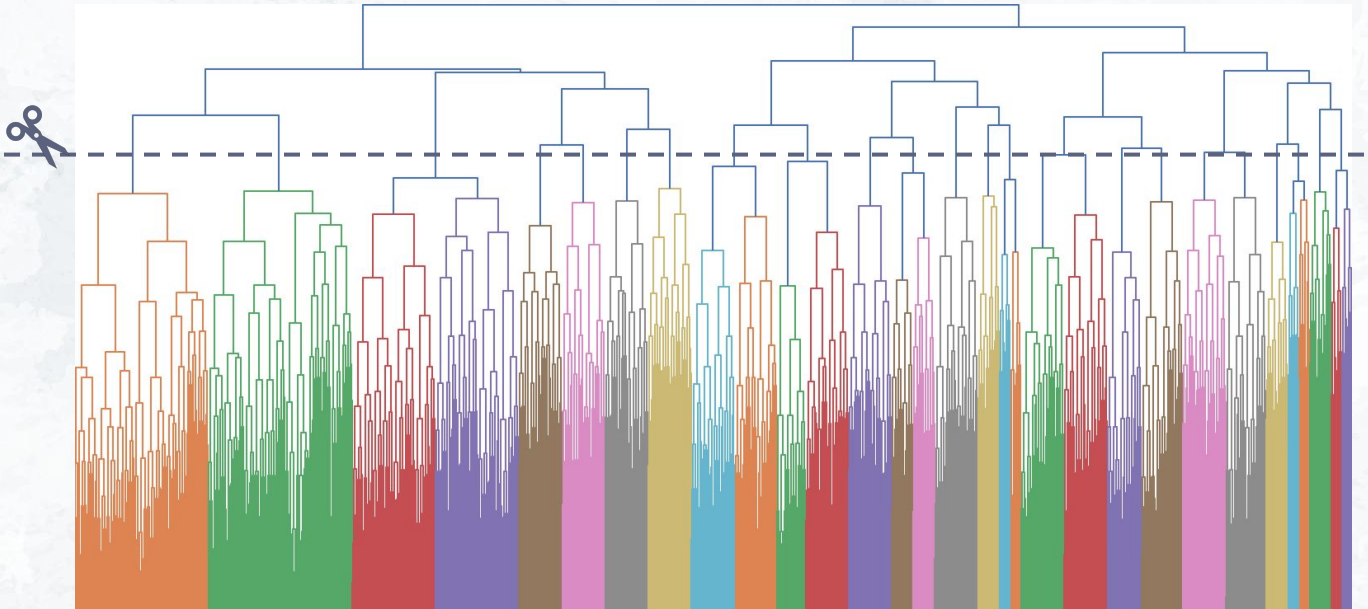
Clustering Models

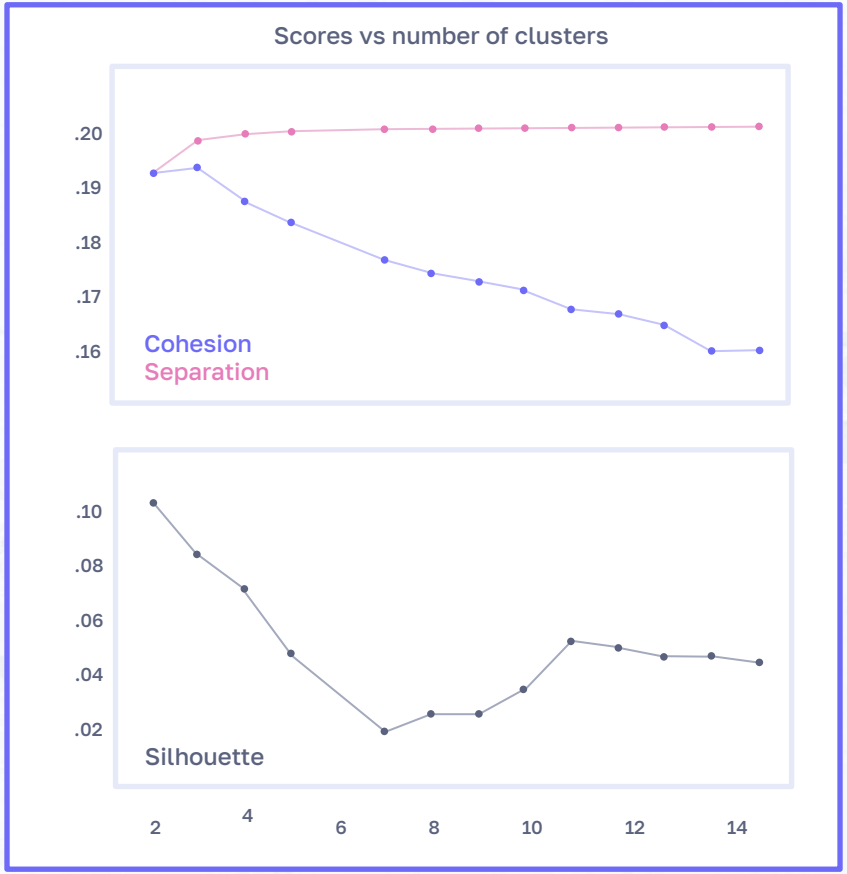


a. Hierarchical clustering
b. k-prototypes clustering

a. Hierarchical

Organizes data into tree structure (dendrogram) by iteratively merging most similar clusters.
Compute solution by choosing a cutting point and picking associated clusters.





Number of clusters optimization

Find optimal number of clusters:

- Cohesion (within cluster)
- Separation (between clusters)
- Silhouette (mix)

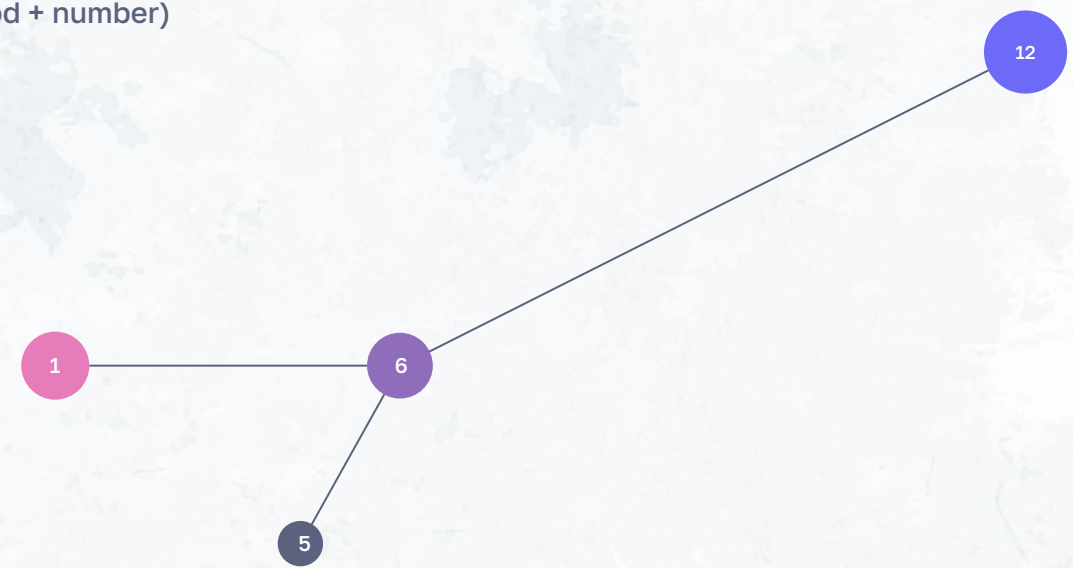
k = 2

b. k-prototype

k-means variant that deals with mixed data.

Need to tune:

- Initialization (method + number)
- Number of Clusters

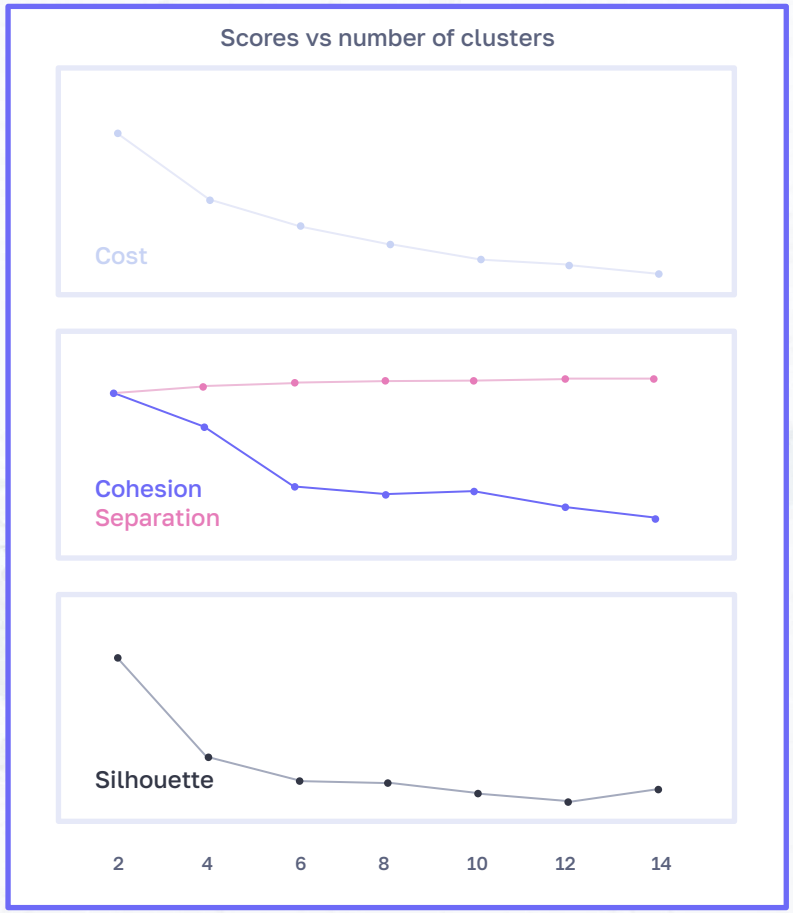


Initializations optimization

Find optimal number of initializations and method:

- Cao
- Huang

Huang
6



Number of clusters optimization

Find optimal number of clusters:

- Score (inertia-like)
- Cohesion (within cluster)
- Separation (between clusters)
- Silhouette (mix)

k = 2

03

Evaluation

Supervised Scores

Utilize supervised scores for evaluation, making use of ground-truth labels of target variables:

- Rand Score: consider pairs of points

$$\mathrm{R} = \frac{a+d}{a+b+c+d}$$

a: same cluster and same class
b: same cluster but different class
c: different cluster but same class
d: different cluster and different class

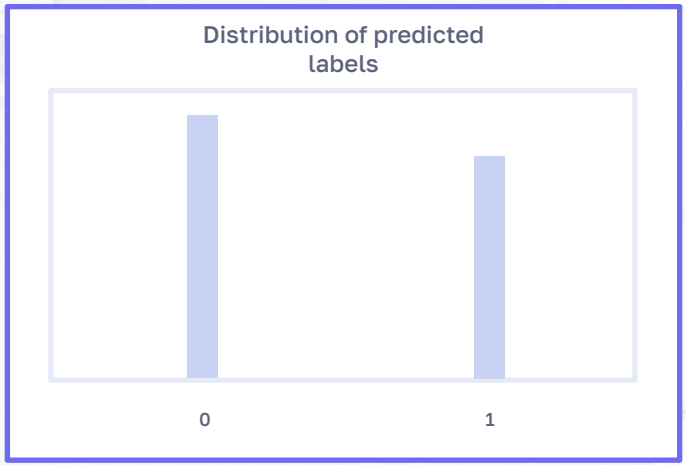
- Fowlkes-Mallows Score

$$\mathrm{FM} = \sqrt{Precision \cdot Recall}$$

a. Hierarchical

Poor performance with all three target variables.
No clear interpretation for number of clusters.

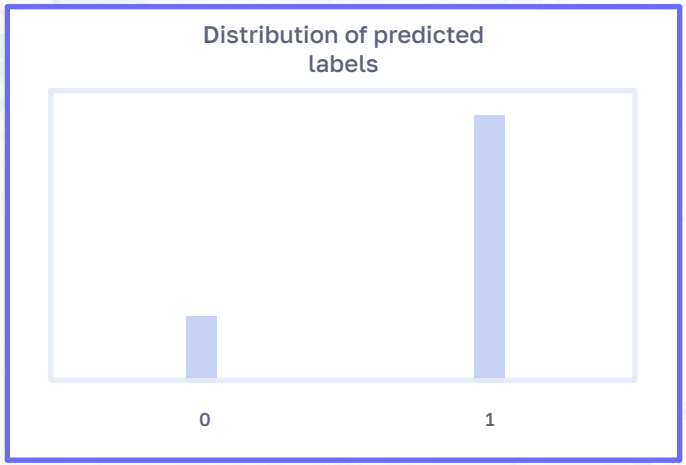
	Rand	Fowlkes-Mallows
Diabetes	0.52	0.52
Hypertension	0.52	0.52
Stroke	0.51	0.67

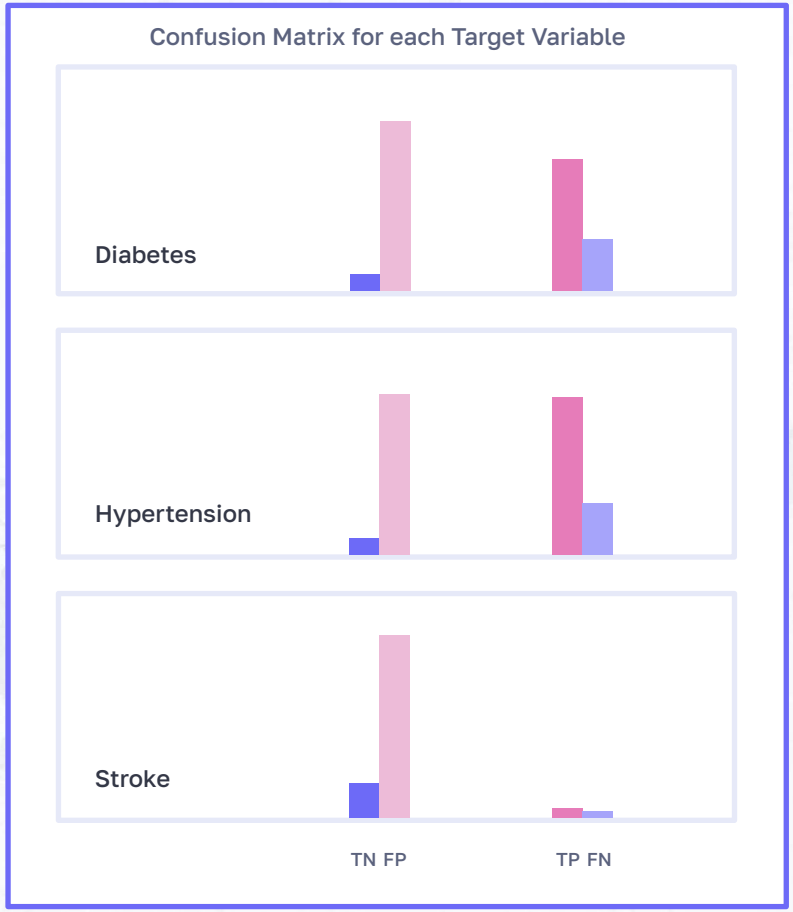


b. k-prototypes

Again, bad performances made exception for stroke, as well as no clear interpretation.

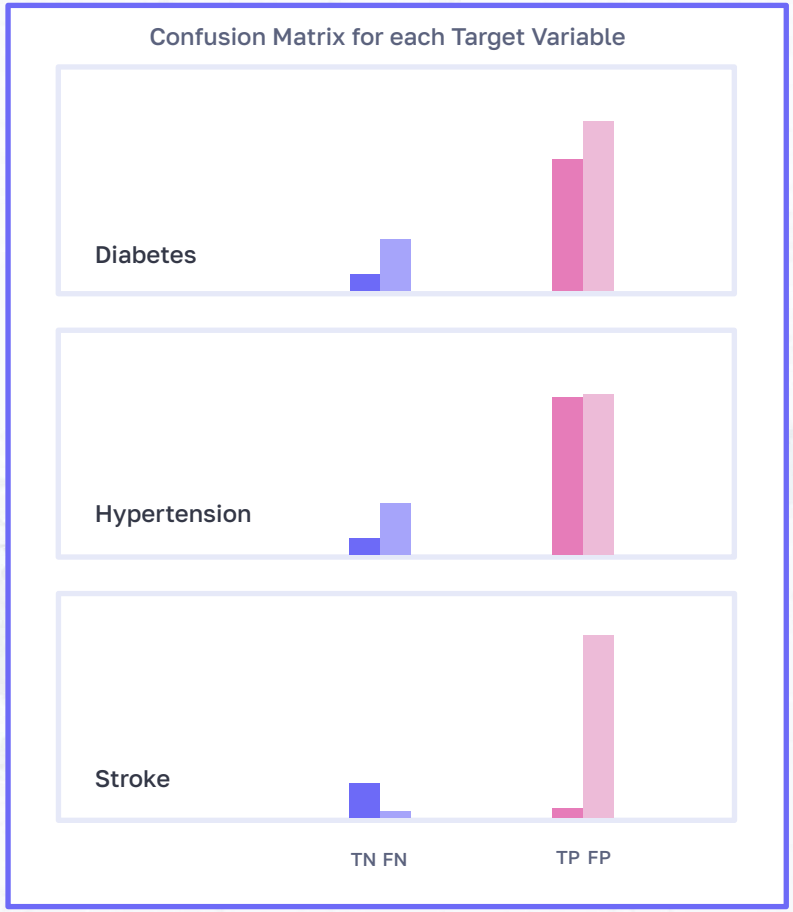
	Rand	Fowlkes-Mallows
Diabetes	0.52	0.61
Hypertension	0.50	0.60
Stroke	0.70	0.82





Confusion Matrix

Precision	Recall	
0.45	0.74	Diabetes
0.50	0.76	Hypertension
0.04	0.56	Stroke

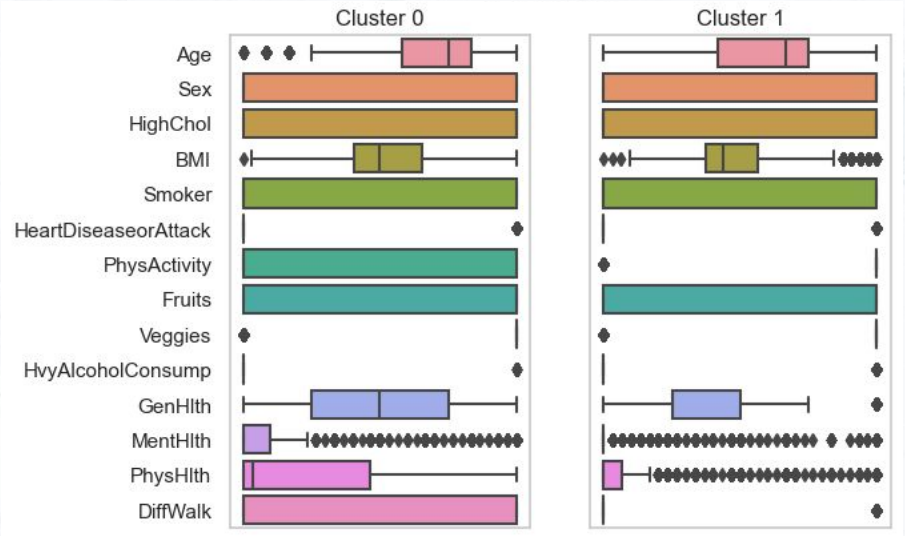


Confusion Matrix

Precision	Recall	
0.45	0.74	Diabetes
0.50	0.76	Hypertension
0.04	0.56	Stroke

Feature selection?

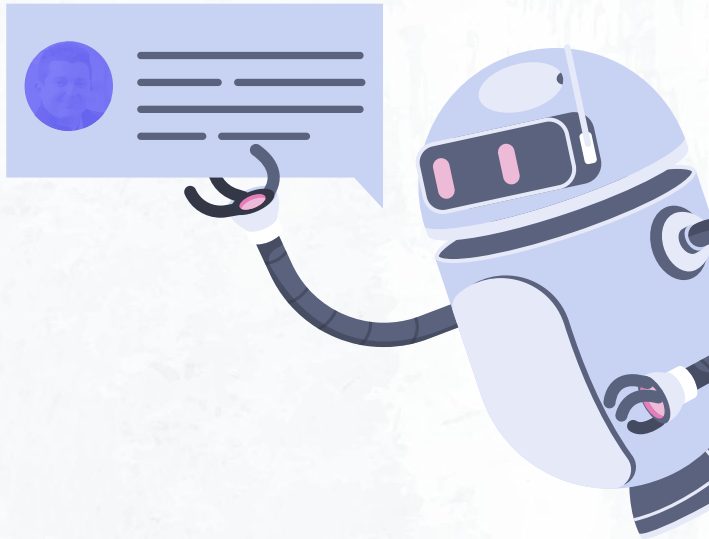
Consider boxplot of each variable in the two clusters.
Hypothetically, draw conclusions on most significant features for target prediction.



inconclusive.

04

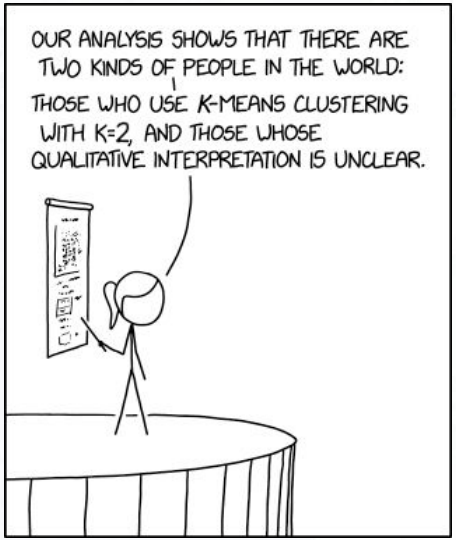
Conclusion



No reasonable solution found, probably due to the nature of the dataset.
In the original challenge, 2 out of 3 target variables were used as attributes.

Nonetheless, we are proud of our work.

Thanks!



(This comic would have been funny if the solution we found made any sense)

Credit:
Filippo Monaco & Marco Picione
under the wise supervision of Prof. Fabio Stella & Giulia Cisotto.

(We did not make us of ChatGPT or any other Natural Language Processing models for this project)