POLITECNICO
MILANO 1863

FINTECH

Machine Learning Project

Data Driven Customer Profiling

Prof. ***Daniele Marazzina***

Prof. ***Raffaele Zenti***

Omar Abdrabou
(10573522)

Leonardo Mandruzzato
(10806545)

Cristiano Serafini
(10621934)

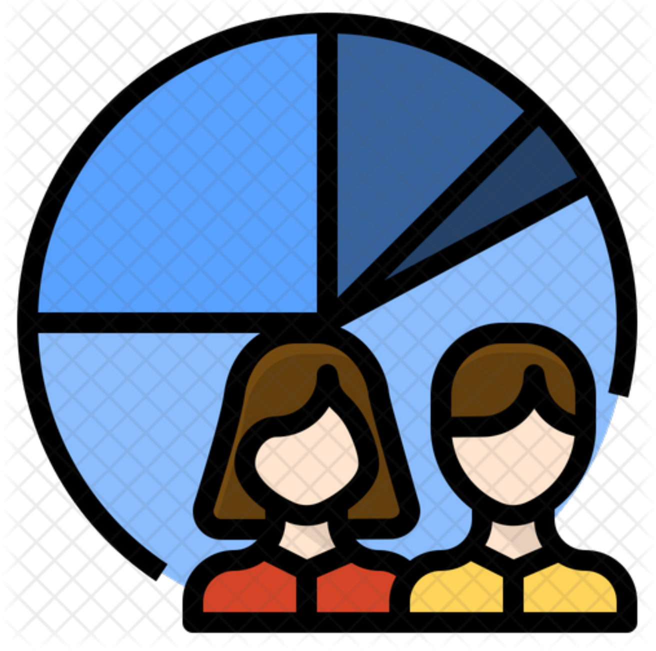
INTRODUCTION

Context Explanation

Customer segmentation is a key for successful targeted marketing, which can target specific groups of customers with different promotions, pricing options, and product placement that catch the interests of the target audience in the most effective way.

Using data to support this activity, gives the opportunity to develop a methodical, well-structured, and efficient process to glean insight on the customers' behavior, needs, and interest, gaining a competitive advantage and using resources.

Through data analysis, unsupervised Machine Learning algorithms can help find customer segments that would be very difficult to spot through intuition and manual examination of data, and even help to find segments that we unknown up to the present.



INTRODUCTION
Objectives Definition

The used dataset describes the information about a bank's customers, offering insight on some important characteristics like income, current investments, and the propensity of using a bank's services.
Different types of analyses were performed on the dataset, like pre-processing the features and studying the correlation between them, exploring it from multiple points of view, and deeply understanding the intrinsic value of each feature and, gain insight.

The objectives of work performed are the following:

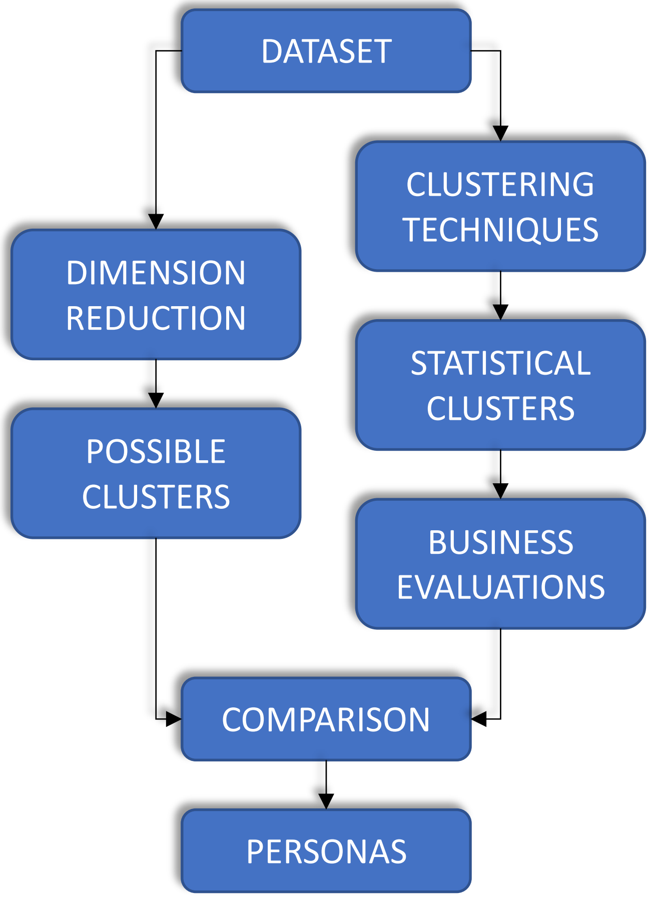
➢Redefinition of current customer base by means of segmentation and identification of customers archetypes (Personas).

➢Support on increasing focus on most profitable segments by summarizing and displaying their characteristics in a graphical manner.

➢Support to the creation of services and strategies tailored to the customers.

WHAT WE DID

- Starting from the **dataset**, we fed it in some **dimensionality reduction algorithms and models** like *t-SNE*, *PCA*, *ICA*, and *autoencoder*
- In this way, we managed to visualize and analyze it, and we individuated some **possible clusters** that we then compared with the ones obtained from the clustering techniques
- We used 3 main **clustering algorithms**:
  - *K-Medoids*
  - *Spectral clustering*
  - *DB Scan*
- To the results obtained from the clustering techniques, we made some **business evaluations** that brought us to increase the number of clusters to consider w.r.t. the ones suggested by the algorithms
- We **compared in detail** the possible clusters individuated through data visualization and the ones found with algorithms (+ business evaluation) to design the 4 main personas describing their respective bank customers' segments



WHAT WE DID

DIMENSIONALITY REDUCTION

While *t-SNE* and *PCA* were already implemeneted in Matlab and we just copy-pasted it, we also exploited 2 other famous and very used techniques:

- *Indipendent Component Analysis (ICA)*
- *Auotoencoder (AE)*

We then compared the results obtained with *PCA*, *ICA*, and *AE* because they were very similar. On the other side, a comparison between *t-SNE* and these other three was a little bit more difficult because the plots were very different, and also from a theoretical standpoint, they are very distant.

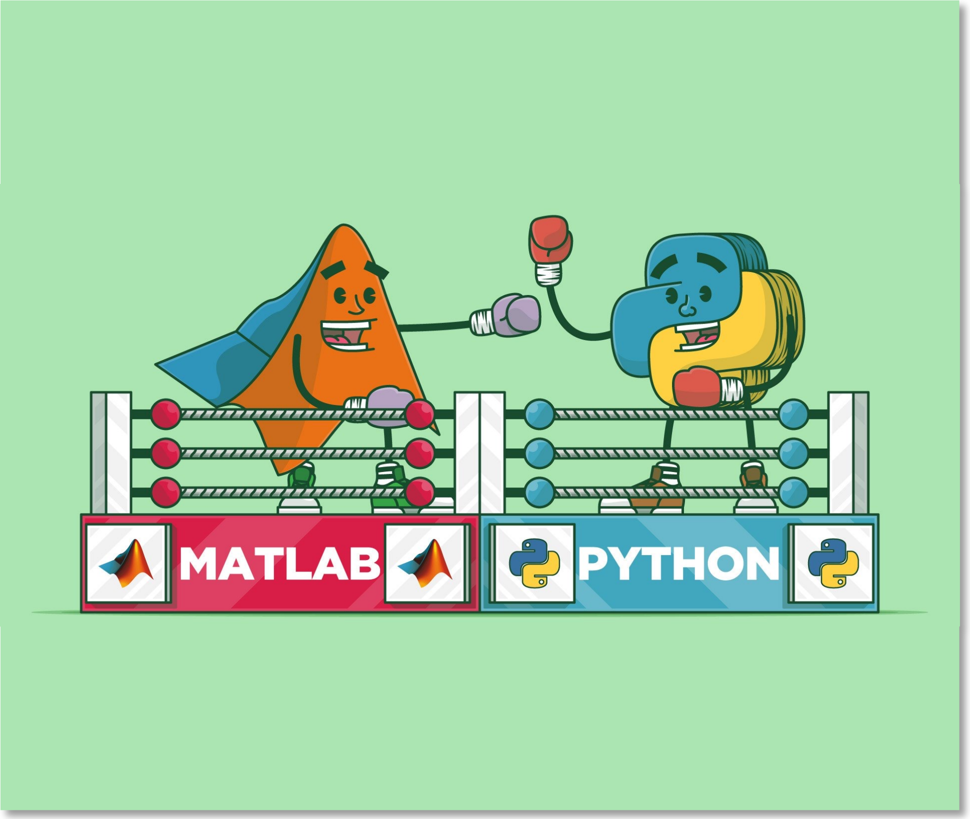
CLUSTERING TECHNIQUES

The 3 techniques we used to achieve our goals are the same exploited in *Matlab*. The difference lays in the implementation. Firstly, we used the *Sklearn* implementation of them. Secondly, we passed different parameters to the algorithms w.r.t. *Matlab*. Finally, for *Spectral Clustering*, we introduced the *Nearest Neighbor Concpet*. Since the latter sometimes is used to implement a classifier *(supervised learning)*, it cannot be used for *unsupervising purposes*, unless in combination with other graph theory techniques, such as *Spectral Clustering*.

For both the tasks accomplished, we did not only used the *Mix Distance* function we have seen in *Matlab*, but we also exploited the *Gower Distance*. So, we repeated all the passages twice: once with the *Mix Distance* and the other with the *Gower* one.

KEY ASPECT

Matlab vs Python



We decided to re-write the code in *Python* for 3 main reasons:

1. **Didactical scope →** since at the beginning some concepts were difficult to understand, deeply look at them from 2 points of view instead of only 1, was very useful to learn how to master them
2. **Convenience →** once we started to write our own code (and not only transfer from *Matlab* to *Python* the one already written), 2 out of 3 members of the group were quite familiar with Python
3. **Knowledge →** in this way, we learned to read and implement code regarding *unsupervised learning* in 2 of the most used languages in this compound

KEY ASPECT

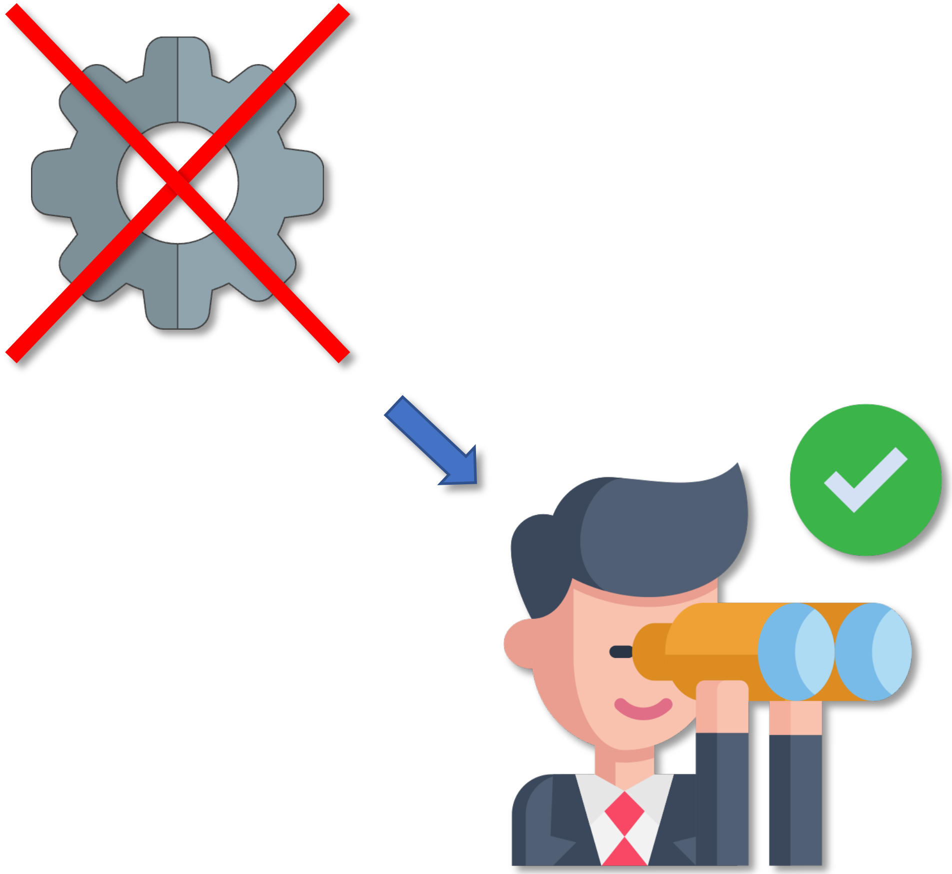
Statistical Criteria vs Business Needs

We gathered all of the analytical results that we have derived from the evaluation methods performed for each of the clustering techniques used. Then, we applied these results to our knowledge of the datasets and comparing them to the groups that we theorized in the beginning, it appears obvious that categorizing all the population of a bank's clients without some sort of adjustments was inappropriate.

Also, from a bank perspective, the results obtained would not have been useful to the bank for creating a plan of action for their customers.

Hence, instead of simply trusting the results derived from the used heuristics, we surmised it would give a far more significant result to set a higher number of clusters for our conclusions.

So, we can say that we do not have only exploited the statistical criteria to accoplish this task, but we have also analyzed the business needs and act to satisfy them.



RESULTS WE GOT

Dimensionality reduction

After having analyzed all the results obtained using the different techniques and different distances, the ones that convinced us the most were those related to the combination ***t-SNE* and *Mix Distance***.
The clusters were easier to distinguish in that plot than in the others and almost all the features were distributed with a meaning inside the graph.

The *t-SNE* results obtained with the Gower Distance were also quite well distributed along the axis of the graph. However, there was a little problem with the *gender* feature: using it, the plot was splitt according to it and for this reason, clustering was more cumbersome.

With PCA, ICA and AE the plots were similar and it was not so easy to distinguish clusters inside.

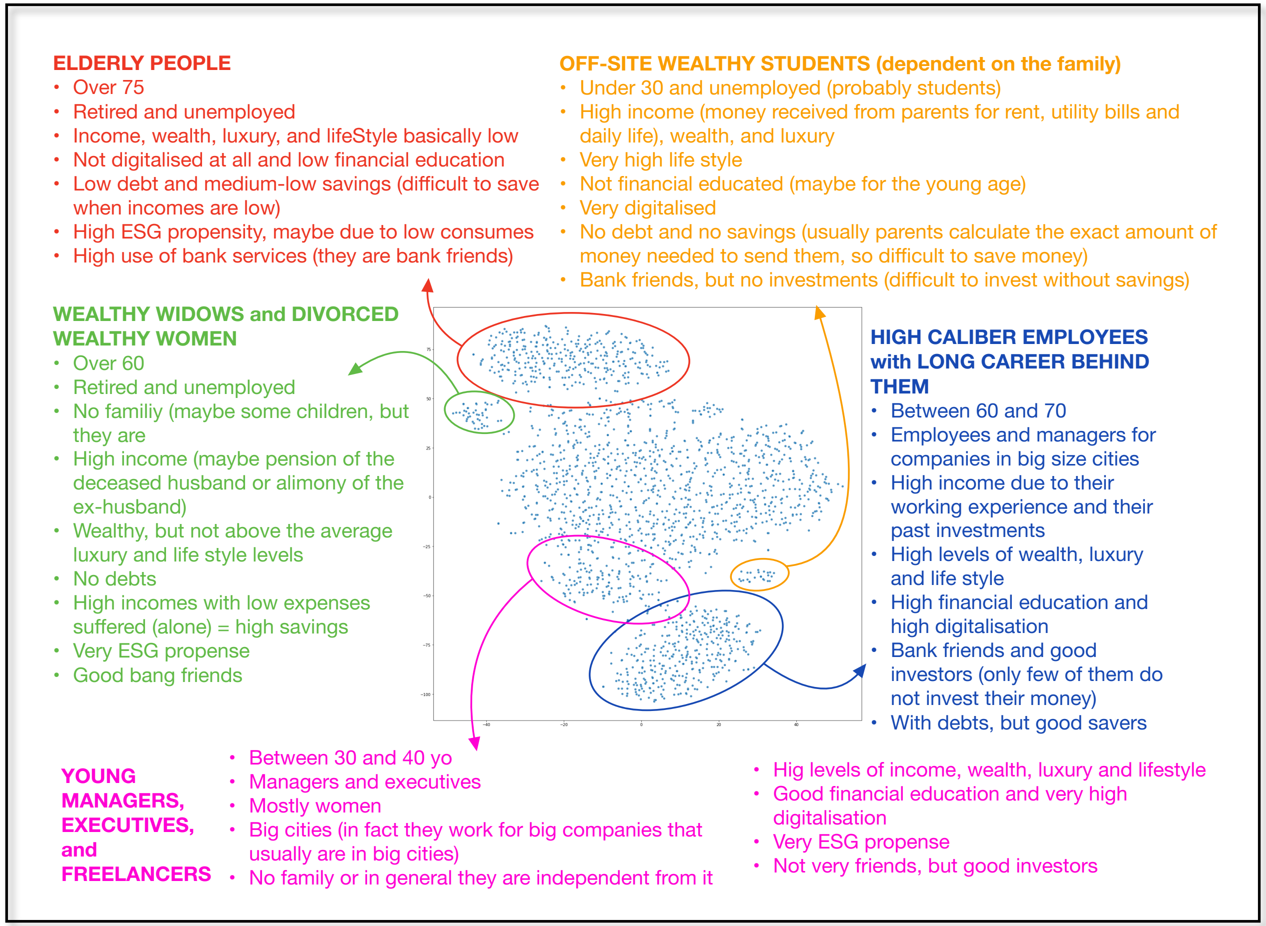
RESULTS WE GOT

Possible clusters

The plot you can see to the side is the one obtained with the combination between t-SNE and the Mix Distance.

We immediately started to search for possible clusters and we reported our intuitions here.

Only afterward we compared the following hypothesis with the results emerging from the use of clustering techniques to reach the final personas the project aims to find.



ELDERLY PEOPLE
- Over 75
- Retired and unemployed
- Income, wealth, luxury, and lifeStyle basically low
- Not digitalised at all and low financial education
- Low debt and medium-low savings (difficult to save when incomes are low)
- High ESG propensity, maybe due to low consumes
- High use of bank services (they are bank friends)

OFF-SITE WEALTHY STUDENTS (dependent on the family)
- Under 30 and unemployed (probably students)
- High income (money received from parents for rent, utility bills and daily life), wealth, and luxury
- Very high life style
- Not financial educated (maybe for the young age)
- Very digitalised
- No debt and no savings (usually parents calculate the exact amount of money needed to send them, so difficult to save money)
- Bank friends, but no investments (difficult to invest without savings)

WEALTHY WIDOWS and DIVORCED WEALTHY WOMEN
- Over 60
- Retired and unemployed
- No familiy (maybe some children, but they are
- High income (maybe pension of the deceased husband or alimony of the ex-husband)
- Wealthy, but not above the average luxury and life style levels
- No debts
- High incomes with low expenses suffered (alone) = high savings
- Very ESG propense
- Good bang friends

HIGH CALIBER EMPLOYEES with LONG CAREER BEHIND THEM
- Between 60 and 70
- Employees and managers for companies in big size cities
- High income due to their working experience and their past investments
- High levels of wealth, luxury and life style
- High financial education and high digitalisation
- Bank friends and good investors (only few of them do not invest their money)
- With debts, but good savers

YOUNG MANAGERS, EXECUTIVES, and FREELANCERS
- Between 30 and 40 yo
- Managers and executives
- Mostly women
- Big cities (in fact they work for big companies that usually are in big cities)
- No family or in general they are independent from it
- Hig levels of income, wealth, luxury and lifestyle
- Good financial education and very high digitalisation
- Very ESG propense
- Not very friends, but good investors

RESULTS WE GOT

Clustering Algorithms

One of our findings after running the K-medoids, DBscan and Spectral clustering algorithm is that these approaches returned far better results when given as input the distance matrix calculated with the Mix Distance function, which also produced more defined clusters overall.
A possible reason for this situation is that the Gower Distance function could be heavily influenced by the Gender feature, as all the observations classified as ‘Male’ in our plots have a tendency of being on the bottom of the plot, and ‘Female’ observations are prevalently on top.

Of all the clustering algorithm produced, the one that produced clusters that were more defined and aligned with our initial consideration is the Spectral Clustering approach. Due to its nature, it does not make strong assumptions on the data distribution, unlike other algorithms. It can correctly cluster observations that actually belong to the same cluster but are farther off than observations in other clusters due to dimension reduction.

RESULTS WE GOT

Clusters evaluation

To evaluate the performances of each cluster from every point of view and in a completely objective manner, we implemented multiple methods that enabled us to understand what was, supposedly, the ideal amount of cluster.
For every algorithm, we used the following evaluation methods:

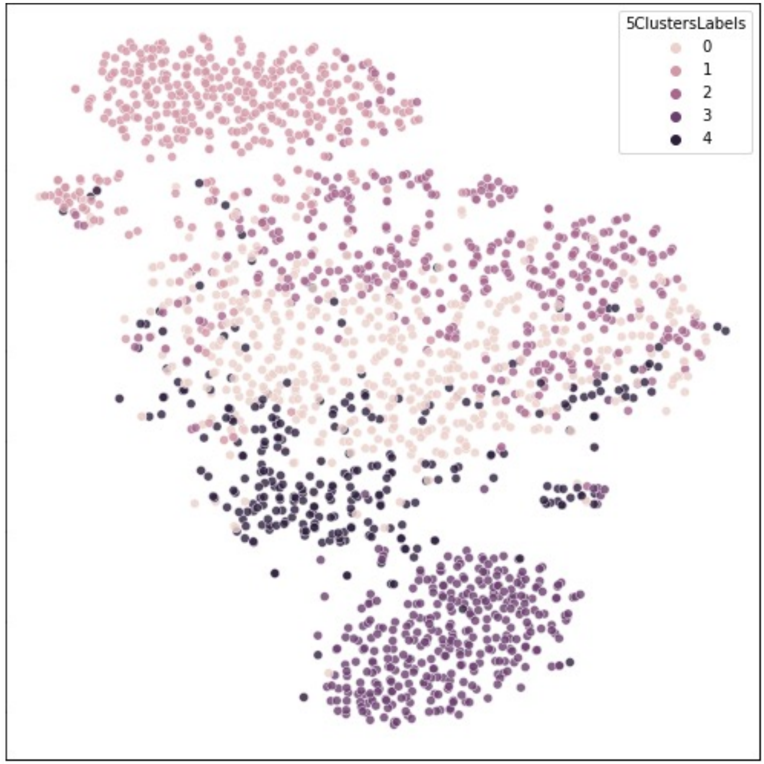
- Inertia Plot and Elbow rule
- Silhouette Score
- Calinski-Harabasz Score
- Davis-Bouldin Score

Also, in case of the Spectral Clustering, a preliminary evaluation using the eigenvalues was also done.
Interestingly enough, all of these metrics applied to the algorithms pointed out that the number of clusters that best grouped the observation belonging to our dataset was between two and three.
As stated before, this result seems to be far off from the actual situation and also lacks any value for banks to gain any kind of insight, as it is highly unlikely that all the customers could be grouped in two groups, three at best.
For this reason, we decided to delve deeper the data and apply our initial findings about the possible number of groups from a business perspective and analyze what the results would be for a higher number of cluster.

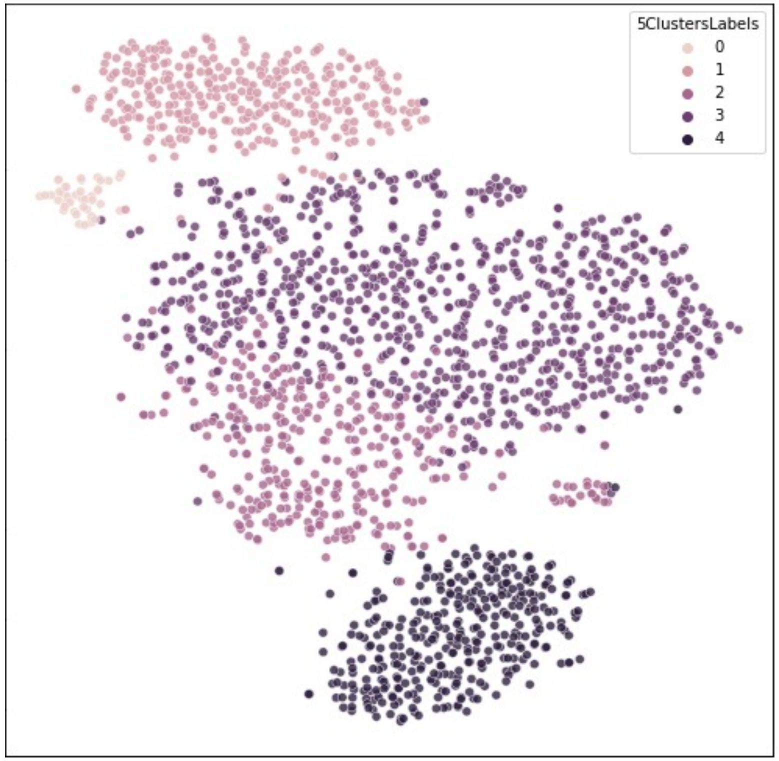
RESULTS WE GOT

Clusters' representations

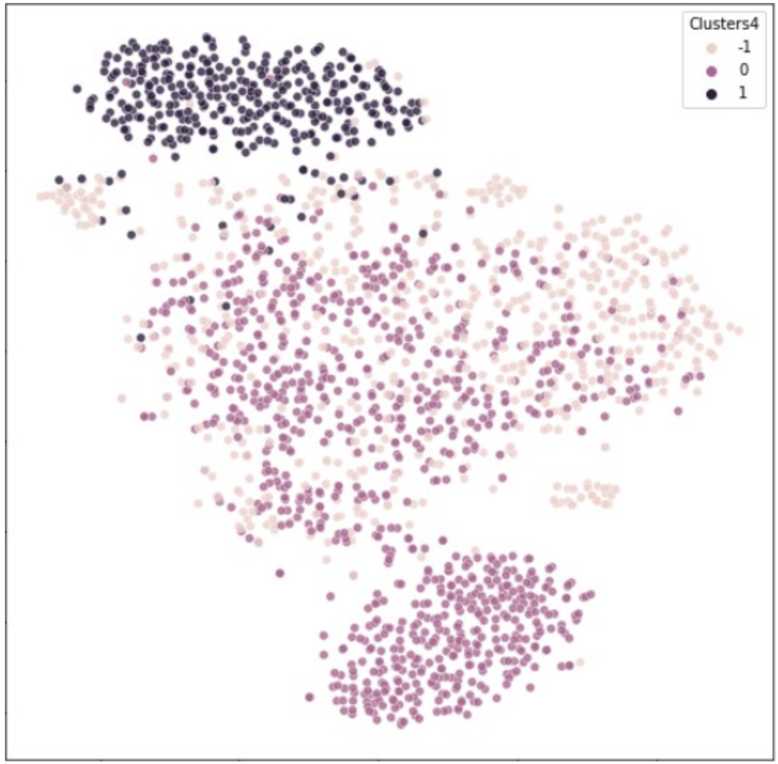
Here are the graphical representation of the data with K-medoids and Spectral clustering algorithms considering five as the chosen number of cluster to use, four for DBscan. All using the Mix Distance function.



K-Medoids Algorithm Results



Spectral Algorithm Results



DB Scan Algorithm Results

COMPARISON

Dimensionality Reduction vs Clustering Techniques

CLUSTER	Dimentionality Reduction	Spectral	K-medoids
Elderly People	X	X	X
Whealty Widow	X	X	
Young Manager	X	X	X
Hight Employee	X	X	X
Off Site whealty Student	X		
Average Person	X	X	X

From the image on the right, we see that the three techniques give different results.
The clustering algorithms can find all manually identified clusters except the one represented by wealthy students. Perhaps because there are too few observations to point out its own cluster.

PERSONAS

PERSONA #1

Lorenzo,60 years old

JOB: CEO of an important company
EDUCATION: Master in management engineering
FREE TIME: Family, Theather
CLUSTER: High level employee
MAIN NEED:
- Long Term Care
- Life Insurance
- Investments (low risky, low but sure capital gain)
- Premium Credit Card
- Asset Manager

FEATURES:

Bank Friend	yes	**Income**	Hight	**Debt**	High
Investments	High	**Wealth**	High	**Saving**	High
Financial Education	High	**Luxury**	High	**Digital propensity**	High
Marital status	married	**Life-Style**	High	**ESG**	High
Family Size	More than two member	**Size of the municipality**	Big		

PERSONA #2

Franco, 77 years old

JOB: City Employee
EDUCATION: Secondary School
FREE TIME: Friends, Grandchildren, Bar
CLUSTER: Elderly poor people
MAIN NEED:
- Long Term Care
- Investments (one-off what the bank recommends)
- Debit Card
- Pension Fund

FEATURES:

Bank Friend	Yes	**Income**	Low	**Debt**	Low
Investments	Low	**Wealth**	Low	**Saving**	Low
Financial Education	Low	**Luxury**	Low	**Digital propensity**	Low
Marital status	Married	**Life-Style**	Low	**ESG**	High
Family Size	At least two member	**Size of the municipality**	Small Medium		

PERSONAS

PERSONA #3

Beatrice, 32 years old

JOB: Head Manager of an important company
EDUCATION: Master in Management and MBA
FREE TIME: Family, friends
CLUSTER: Women career
MAIN NEED:
- Short/Long Term Care
- Job Salary
- Investments (risky, capital gain)
- Debit Card

FEATURES:

Bank Friend	No	**Income**	High	**Debt**	Average
Investments	High	**Wealth**	High	**Saving**	High
Financial Education	High	**Luxury**	High	**Digital propensity**	High
Marital status	Not Married	**Life-Style**	High	**ESG**	High
Family Size	One component	**Size of the municipality**	Big		

PERSONA #4

Carla, 62 years old

JOB: Housewife
EDUCATION: Law Degree
FREE TIME: Art Exhibitions, women’s rallies, parades
CLUSTER: Whealty widow
MAIN NEED:
- Long Term Care
- Inheritance
- Investments (low risk, capital protection)
- Premium credit cards

FEATURES:

Bank Friend	Yes	**Income**	High	**Debt**	Low
Investments	High	**Wealth**	High	**Saving**	High
Financial Education	Low	**Luxury**	Average	**Digital propensity**	High
Marital status	Widow	**Life-Style**	Average	**ESG**	High
Family Size	No more than two component	**Size of the municipality**	Medium Big		

REFERENCES

PCA Article: https://towardsdatascience.com/dimensionality-reduction-with-autoencoders-versus-pca-f47666f80743
K-Medoids Article: https://analyticsindiamag.com/comprehensive-guide-to-k-medoids-clustering-algorithm
Evaluation Article: https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c
Spectral Clustering: https://towardsdatascience.com/spectral-clustering-aba2640c0d5b
Autoencoder vs PCA: https://towardsdatascience.com/dimensionality-reduction-with-autoencoders-versus-pca-f47666f80743
ICA: https://towardsdatascience.com/independent-component-analysis-ica-a3eba0ccec35
Gower Distance: https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553

Python Libraries:

- https://scikit-learn-extra.readthedocs.io/en/latest/generated/sklearn_extra.cluster.KMedoids.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html