

The background image shows a person in blue medical scrubs holding a smartphone. Overlaid on the image is a complex digital network of glowing blue lines and various medical icons, including a heart, lungs, a stethoscope, and a smartphone. The overall theme is medical technology and data analysis.

# Analysis of a medical database

Abderrahmen Borchani

27/04/2024

# Outline

---

- Introduction
- Methodology
- Results
- Conclusion

# Introduction

---

- Lung cancer is one of the leading causes of death worldwide, according to the World Health Organization (WHO), with an estimated incidence of 2,206,771 in 2020. This situation requires an effective and automated solution to enable early detection.
- Using analysis we will help doctors in their decision-making regarding lung cancer.



Section 1

# Methodology

# Methodology

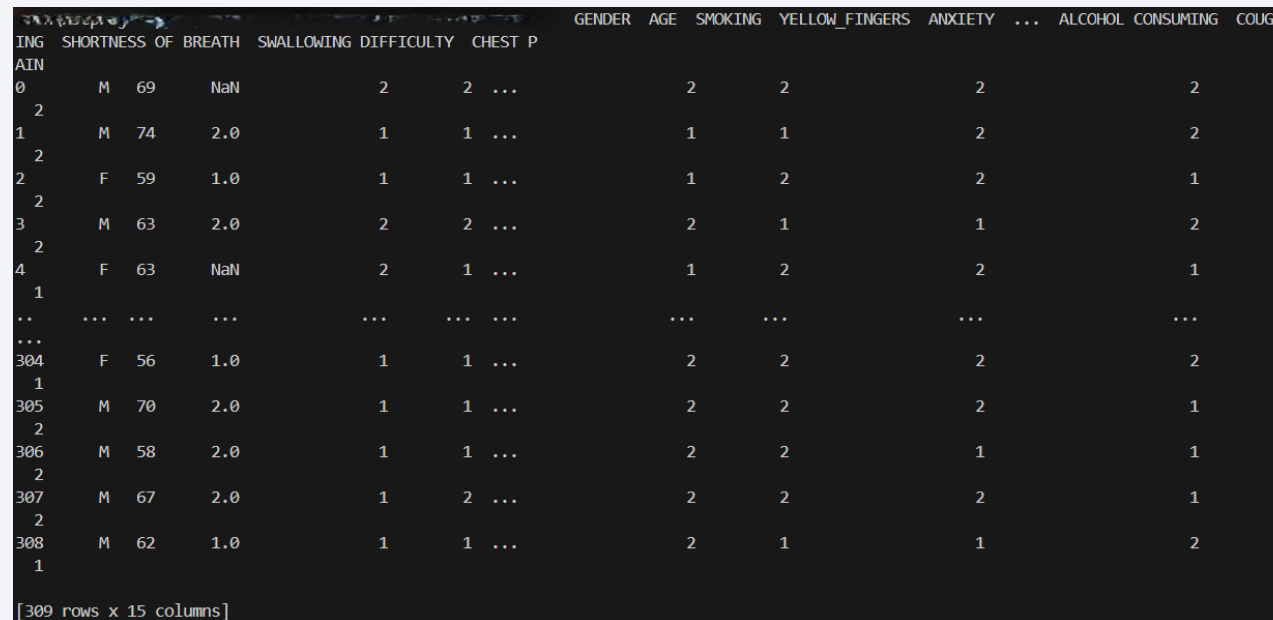
---

## Executive Summary

- The data preparation phase
  - Pretreatment
  - Transformations
- Feature extraction
- Data Mining
- Graphical representation of the proposed system



# Data preparation phase



The screenshot shows a Jupyter Notebook interface with a data preview. The table has 309 rows and 15 columns. The columns are: INDEX, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, CHEST PAIN, GENDER, AGE, SMOKING, YELLOW\_FINGERS, ANXIETY, ..., ALCOHOL CONSUMING, COUGH. The data is displayed in a dark-themed interface. The first few rows are as follows:

INDEX	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	...	ALCOHOL CONSUMING	COUGH
0	M 69	NaN	2	2	...	2	2	2		2	
1	M 74	2.0	1	1	...	1	1	2		2	
2	F 59	1.0	1	1	...	1	2	2		1	
3	M 63	2.0	2	2	...	2	1	1		2	
4	F 63	NaN	2	1	...	1	2	2		1	
...	...	...	...	...	...	...	...	...		...	
304	F 56	1.0	1	1	...	2	2	2		2	
305	M 70	2.0	1	1	...	2	2	2		1	
306	M 58	2.0	1	1	...	2	2	1		1	
307	M 67	2.0	1	2	...	2	2	2		1	
308	M 62	1.0	1	1	...	2	1	1		2	

[309 rows x 15 columns]

- Number of observations: 309
- Number of features: 15
- There are 3 missing values. Missing values replaced with column means.

# Data preparation phase

- Categorical variables encoded successfully.

- Data normalized

- Data preprocessing completed.

- Most correlated variable pairs:

• ANXIETY      YELLOW\_FINGERS      0.565  
829

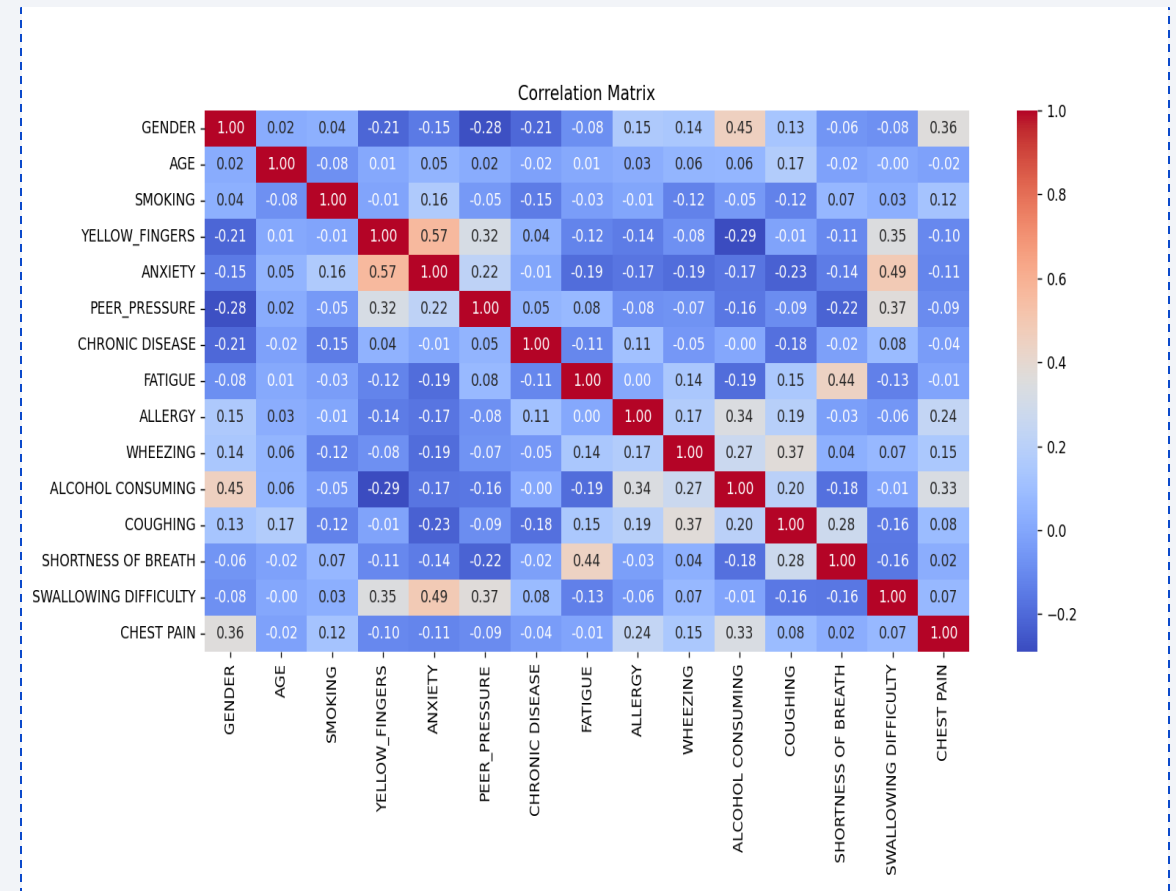
- Individuals with higher levels of anxiety are more likely to have yellow fingers, possibly due to increased smoking habits associated with anxiety.

• SWALLOWING  
DIFFICULTY      ANXIETY      0.489403

- individuals experiencing anxiety may also experience difficulty in swallowing

• ALCOHOL  
CONSUMING      GENDER      0.454268

- There may be differences in alcohol consumption patterns between genders, with one gender being more likely to consume alcohol than the other.



# Feature extraction

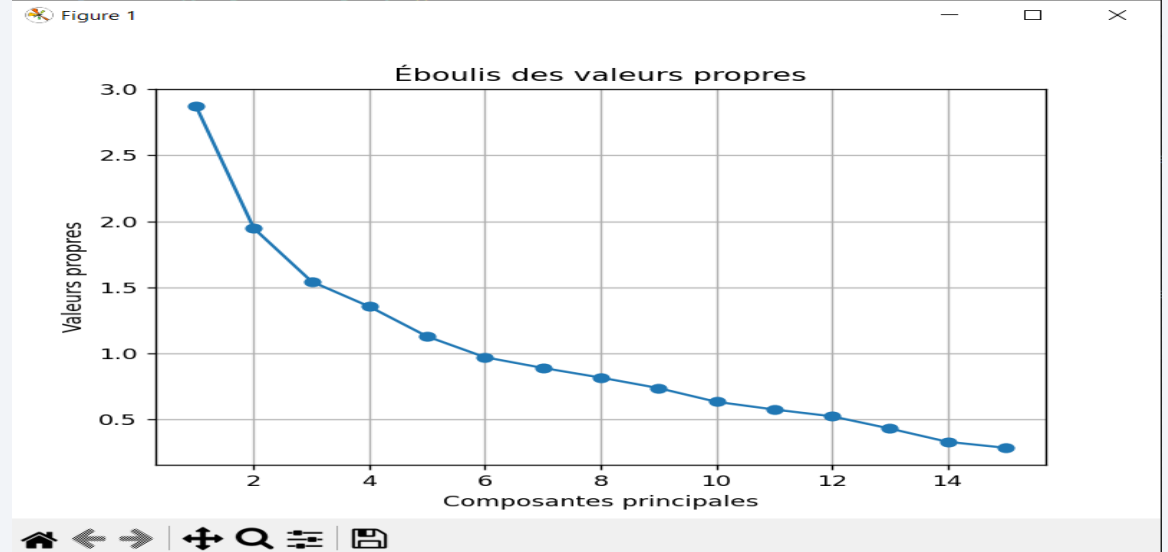
- We obtained 15 eigenvalues, where the first eigenvalue is 2.871, the second is 1.947, and so on. The higher the eigenvalue, the more variance in the data the corresponding principal component captures.
- Cumulative variance explained by the first nine principal components:  $19.08\% + 12.94\% + 10.26\% + 9.00\% + 7.50\% + 6.45\% + 5.92\% + 5.43\% + 4.90\% = 81.48\%$
- Therefore, approximately 9 principal components are needed to explain at least 80% of the total variance in the data.

Valeurs propres:

```
[2.87102374 1.94700442 1.54418309 1.35445336 1.1286086 0.97145805  
0.89059047 0.8176818 0.73806281 0.63445738 0.57620854 0.52468084  
0.43208774 0.33163979 0.28656066]
```

Pourcentage d'inertie expliqué par chaque composante principale:

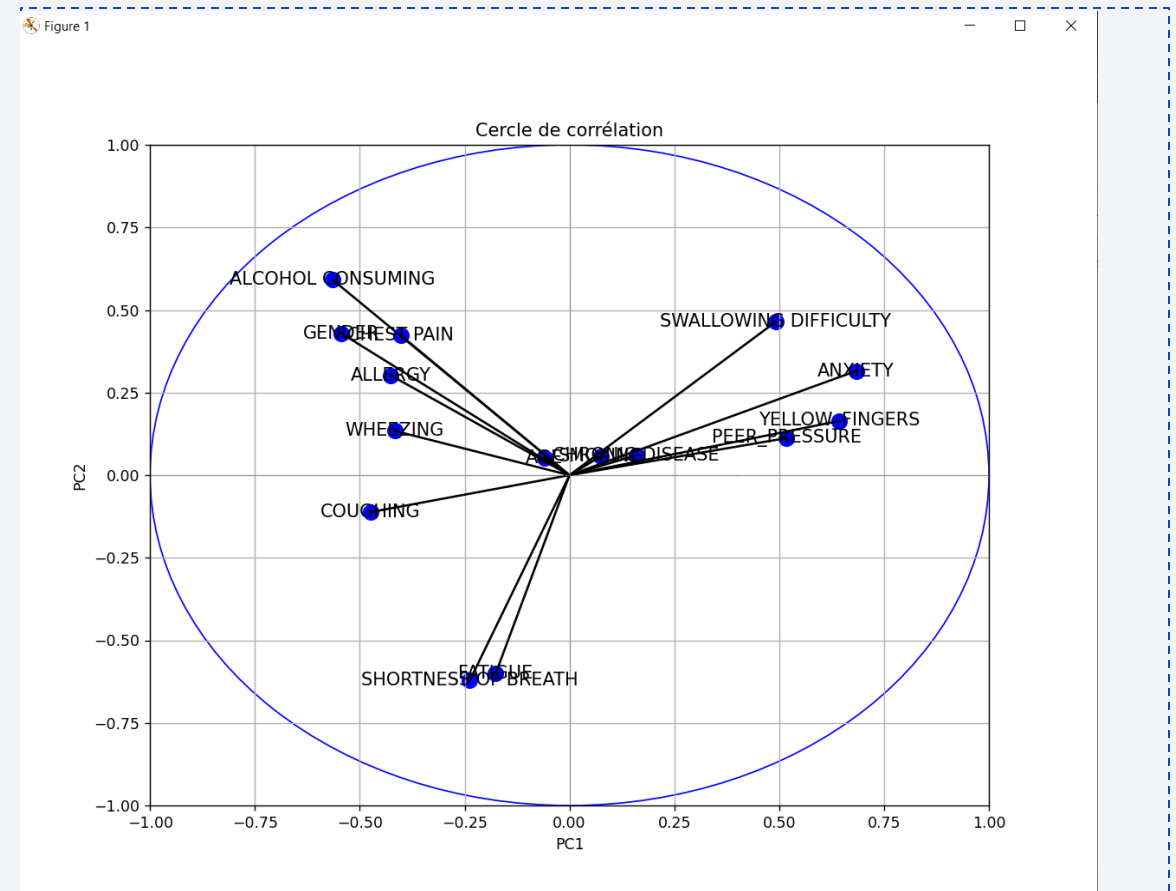
```
[0.19078216 0.12938023 0.10261238 0.09000467 0.07499708 0.06455428  
0.05918055 0.05433571 0.04904495 0.04216027 0.03828959 0.03486552  
0.02871263 0.02203777 0.01904222]
```





# Feature extraction

- Variable Importance: The length of the vector indicates the importance of the variable in explaining the variability in the dataset.
  - Alcohol consuming
  - Fatigue
  - Shortness breath
  - Swallowing difficulty
  - Anxiety
  - Yellow Fingers



# Feature extraction

---

- Variable Clustering: Variables with vectors pointing in similar directions or clustering together on the plot are highly correlated with each other.
  - Alcohol consuming, Chest pain, Gender, Age
  - Shortness breath, Swallowing difficulty
  - Anxiety, Yellow Fingers, Peer pressure, Chronic Disease, Smoking

# Data Mining

---

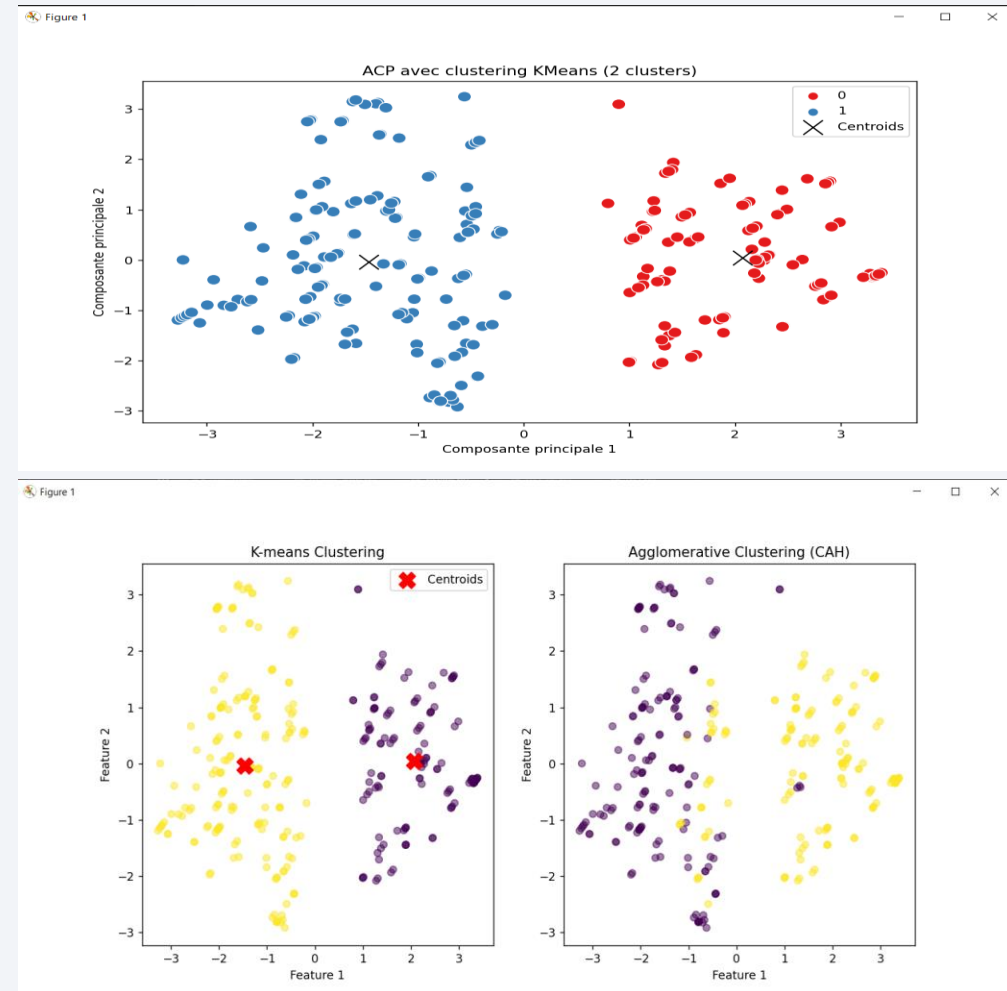
- Silhouette Score for K-means: 0.13584
- Silhouette Score for Hierarchical Agglomerative Clustering: 0.12591
- Best Cluster Assignments is K-means

```
Cluster 1: [ 2.07463889  0.04534674 -0.11646451 -0.0199275  0.01307457  0.08365864  
-0.0392228  0.01252392  0.04924946  0.00255981  0.03534754 -0.04315434  
0.00398751  0.03013861 -0.01453791  0.19663351]  
Inertie associée : 3988.119273775089  
Cluster 2: [-1.46714794 -0.03206841  0.08236164  0.01409237 -0.00924611 -0.05916191  
0.02773767 -0.00885669 -0.03482835 -0.00181026 -0.02499716  0.03051799  
-0.0028199 -0.02131349  0.01028095 -0.13905575]  
Inertie associée : 3988.119273775089
```

- inertia associated with both clusters is the same, indicating similar compactness or dispersion of data points around their centroids.

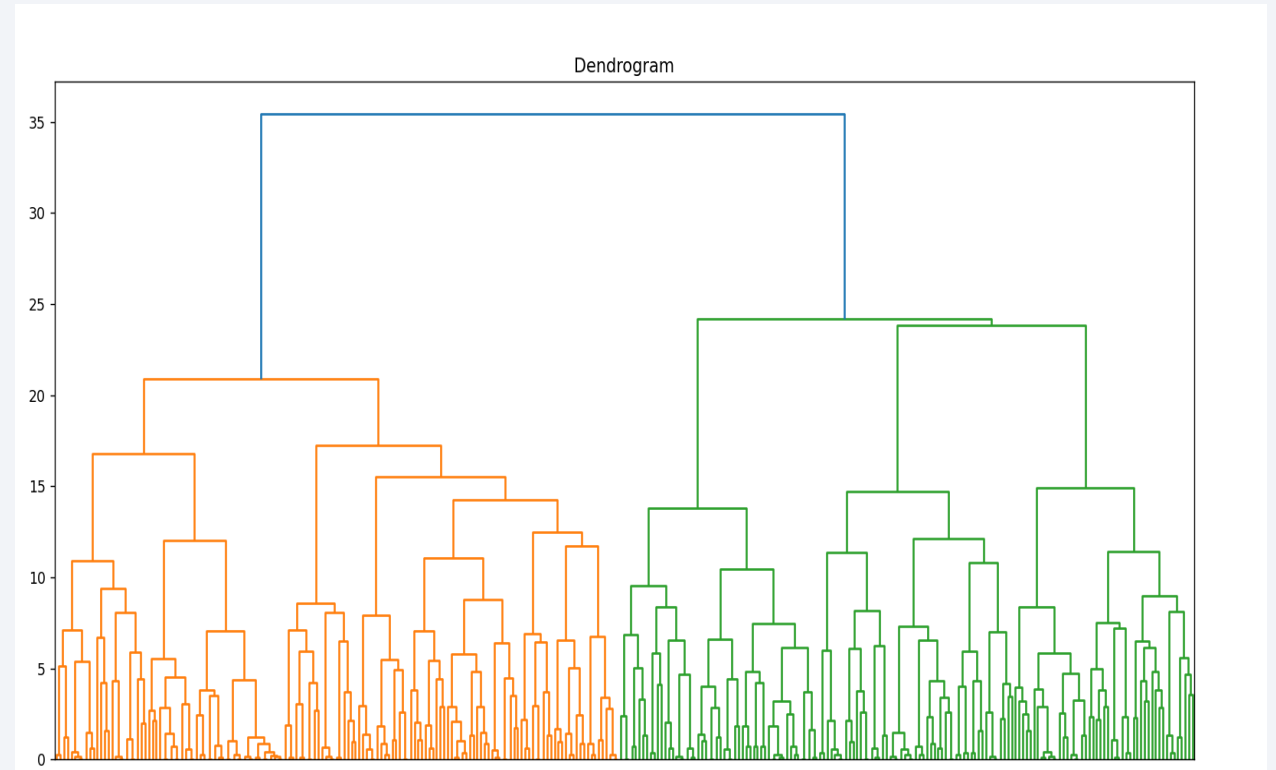
# Data Mining

- Variables are grouped into two distinct clusters based on their characteristics in the space of principal components. These characteristics are obtained from Principal Component Analysis (PCA), which reduces the dimensionality of the data while preserving its variability.
- Centroids: The centroids represent the centers of gravity of each cluster in the space of principal components. Each centroid is an average of the characteristics of the individuals belonging to that cluster.
- Proximity of individuals to centroids: Individuals belonging to a cluster are typically closer to its centroid than to others. This means that the similarity between individuals within the same cluster is higher than that between individuals from different clusters.



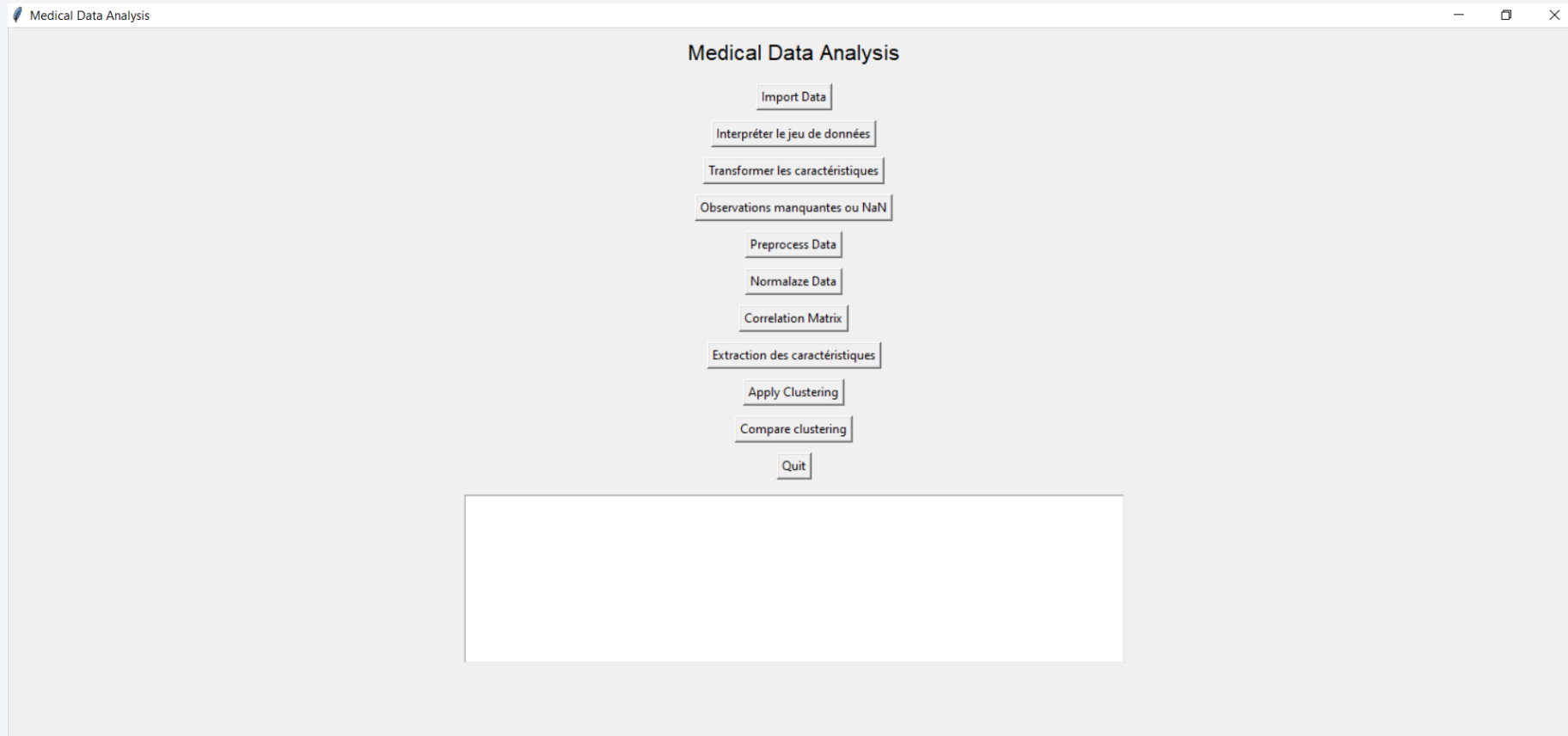
# Data Mining

- Horizontal Axis: Represents individuals or clusters.
- Vertical Axis: Indicates the distance or similarity between individuals or clusters.
- Branches: Connect variables or clusters, showing how they are grouped at different stages of the clustering process.
- Branch Heights: The height at which two variables or clusters are merged indicates their similarity or distance. Higher heights represent greater distances between individuals or clusters.



# Graphical representation

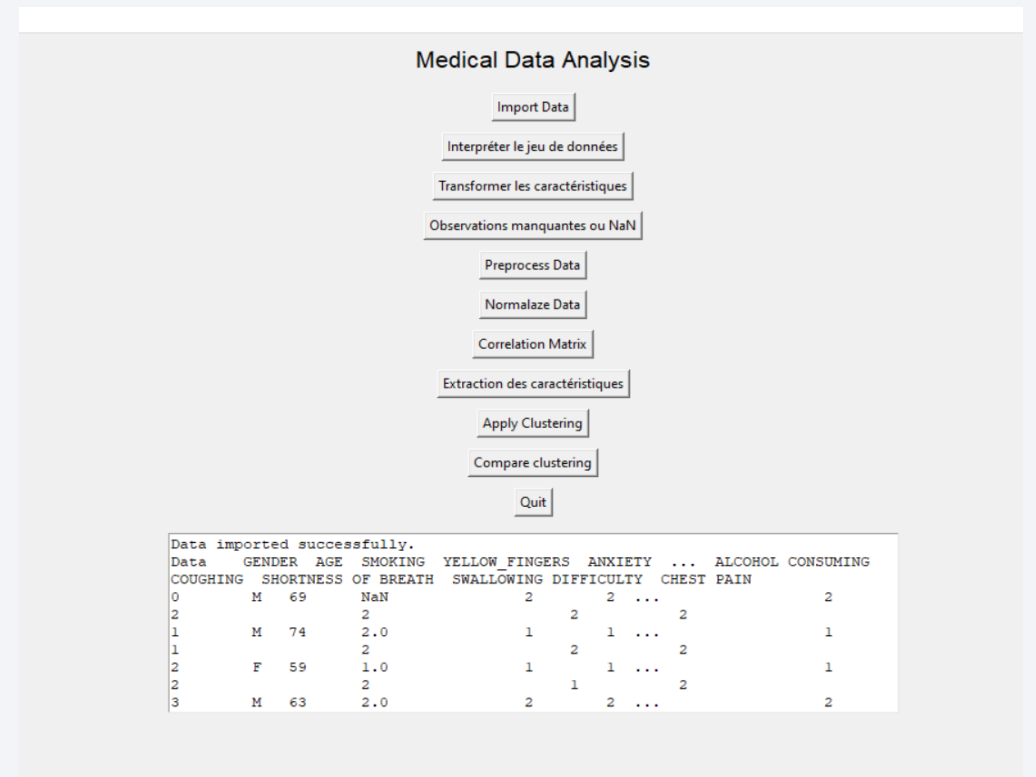
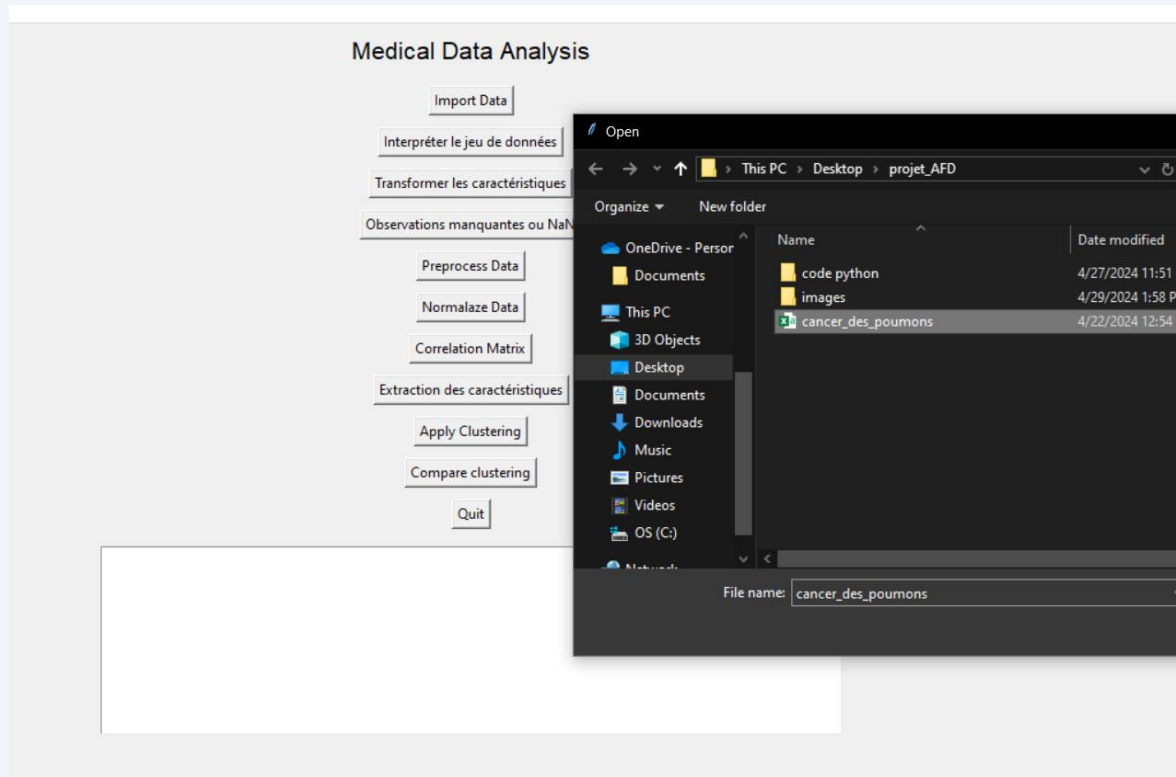
---





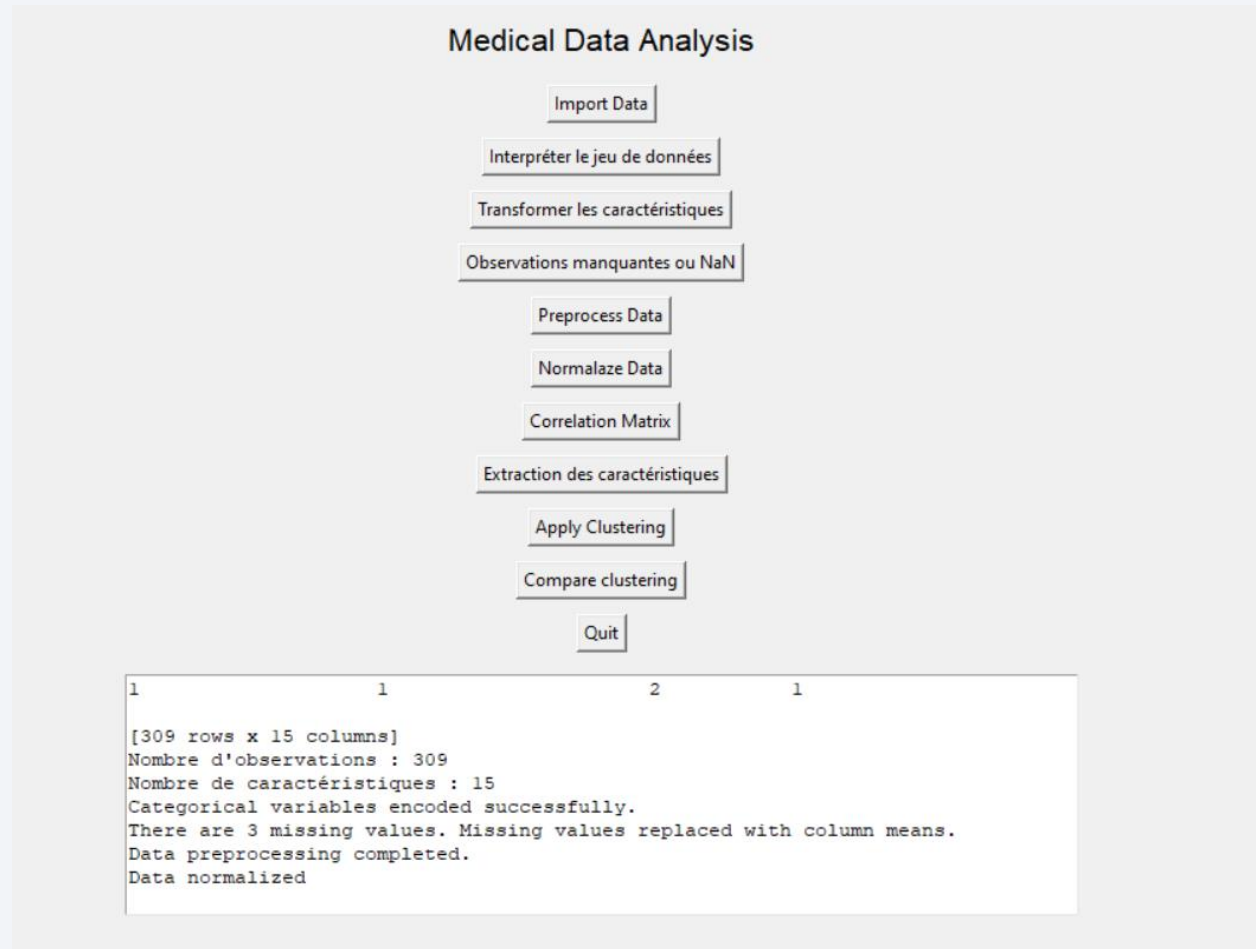
# Graphical representation

- Importing Data



# Graphical representation

---



# Conclusion

---

- Anxiety Reduction:
  - Given that anxiety is a significant risk factor for lung cancer, efforts should prioritize reducing anxiety levels among individuals.
  - It's essential to dispel the misconception that smoking alleviates anxiety. On the contrary, smoking has been shown to increase the risk of chronic diseases and exacerbate anxiety symptoms.
- Smoking Cessation:
  - Smoking cessation programs should be promoted and made widely available to the population. These programs can provide support and resources to help individuals quit smoking and reduce their risk of developing lung cancer.

# Conclusion

---

- Alcohol Consumption Reduction:
  - There is a correlation between alcohol consumption, particularly among men, and an increased risk of lung cancer. Therefore, interventions aimed at reducing alcohol consumption are crucial.
  - Educational campaigns and support services should be implemented to raise awareness about the risks of alcohol consumption and encourage moderation, especially among men.
- Early Detection Awareness:
  - Symptoms such as shortness of breath and swallowing difficulty should not be overlooked, as they could be early indicators of lung cancer.
  - Encouraging individuals to seek medical attention promptly if they experience these symptoms can lead to earlier detection and improved treatment outcomes.

# Conclusion

---

- Public Health Initiatives:
  - Public health initiatives should focus on raising awareness about the link between anxiety, smoking, alcohol consumption, and lung cancer risk.
  - Targeted campaigns should emphasize the importance of lifestyle modifications, such as smoking cessation and moderation of alcohol intake, in reducing the risk of lung cancer.

Thank you!

