

# Clustering et analyse sur un dataset de maladies cardiaques

BELABBAS-BENGRAA Joubrane

BENZIANE Amir

CORROLLER Nathan

Octobre 2025

## Résumé

Ce projet vise à explorer et analyser un jeu de données sur les maladies cardiaques afin d'identifier des structures sous-jacentes grâce à des méthodes de clustering non supervisées. Après une étude descriptive des variables numériques et catégorielles, les données ont été nettoyées et transformées.

Trois algorithmes de clustering ont été appliqués : **K-Means**, **DBSCAN** et **Gaussian Mixtures**. Chaque méthode a été testée avec différentes configurations de paramètres pour identifier la meilleure structuration des données. L'analyse des résultats met en évidence les variables les plus discriminantes pour la formation des clusters et la justification des paramètres des différents algorithmes

Ce projet a permis d'effectuer une classification des patients, en identifiant notamment ceux présentant un risque élevé et ceux présentant un risque faible de maladies cardiaques.

**Joubrane** : Partie 1, 2, 5 **Amir** : Partie 2, 4 **Nathan** : Partie 2, 3

# Table des matières

<b>1 Description des données</b>	
1.1 Données . . . . .	
1.2 Analyse . . . . .	
1.2.1 Valeurs numériques . . . . .	
1.2.2 Valeurs catégorielles . . . . .	
<b>2 Préparation des données</b>	
<b>3 K-Means</b>	
3.1 Fonctionnement de K-means . . . . .	
3.2 Sélection du nombre de clusters et évaluation de la qualité de ces clusters . . . . .	
3.2.1 La méthode du coude . . . . .	
3.2.2 Métriques utilisées . . . . .	
3.3 Prétraitement des données . . . . .	
3.3.1 Suppression des outliers et des valeurs aberrantes . . . . .	
3.3.2 Discrimination des variables en fonction de leur corrélation avec la target	
3.4 Conclusion et résultats . . . . .	
<b>4 Gaussian Mixture Model (GMM)</b>	
4.1 Prétraitement : suppression des valeurs aberrantes . . . . .	
4.2 Choix de la matrice de covariance . . . . .	
4.2.1 Choix du nombre de clusters et du type de covariance . . . . .	
4.2.2 Visualisation et interprétation des clusters . . . . .	
4.2.3 Variables discriminantes : Feature Importance (XAI) . . . . .	
4.2.4 Répartition de la maladie selon les clusters . . . . .	
4.2.5 Analyse des variables continues . . . . .	
4.2.6 Analyse des variables catégorielles . . . . .	
4.2.7 Conclusion : Interprétation globale des clusters . . . . .	
<b>5 Clustering avec DBSCAN</b>	
5.1 Présentation de l'algorithme DBSCAN . . . . .	
5.2 Recherche des hyperparamètres de DBSCAN . . . . .	
5.3 Résultats DBSCAN sur les données PCA . . . . .	
5.4 Résultats DBSCAN après projection t-SNE . . . . .	
5.5 Résultats DBSCAN après réduction Isomap . . . . .	
5.6 Synthèse comparative . . . . .	
5.7 Interprétation des clusters Isomap . . . . .	

## 6 Conclusion

# 1 Description des données

## 1.1 Données

Le jeu de données utilisé dans ce projet provient de la plateforme **Kaggle**, une source reconnue pour la qualité et la fiabilité de ses ensembles de données open-source dans le domaine de la science des données. Ce dataset a été construit à partir de données médicales réelles de patients et constitue donc une base fiable pour une étude de clustering dans le cadre d'une analyse.

Le jeu de données contient **918 observations** (patients) et **12 variables** décrivant certaines caractéristiques sur les individus. La variable cible, **HeartDisease**, indique la présence (1) ou l'absence (0) d'une maladie cardiaque diagnostiquée. Bien que cette variable ne soit pas utilisée pour le clustering, elle permettra d'évaluer la cohérence des clusters obtenus.

TABLE 1 – Description des variables du jeu de données

Variable	Description	Type / Unités
<b>Age</b>	Âge du patient	Numérique (années)
<b>Sex</b>	Sexe du patient (M : Homme, F : Femme)	Catégorielle
<b>ChestPainType</b>	Type de douleur thoracique (TA : Typical Angina, ATA : Atypical Angina, NAP : Non-Anginal Pain, ASY : Asymptomatic)	Catégorielle
<b>RestingBP</b>	Pression artérielle au repos	Numérique (mm Hg)
<b>Cholesterol</b>	Taux de cholestérol sérique	Numérique (mg/dl)
<b>FastingBS</b>	Glycémie à jeun (1 si $> 120 \text{ mg/dl}$ , 0 sinon)	Binaire
<b>RestingECG</b>	Résultats de l'électrocardiogramme au repos (Normal, ST, LVH)	Catégorielle
<b>MaxHR</b>	Fréquence cardiaque maximale atteinte	Numérique (entre 60 et 202)
<b>ExerciseAngina</b>	Angine induite par l'exercice (Y : Oui, N : Non)	Catégorielle
<b>Oldpeak</b>	Dépression du segment ST (valeur numérique)	Numérique
<b>ST_Slope</b>	Pente du segment ST lors de l'effort (Up, Flat, Down)	Catégorielle
<b>HeartDisease</b>	Présence d'une maladie cardiaque (1 : malade, 0 : sain)	Binaire (variable cible)

## 1.2 Analyse

Une première étape d'analyse exploratoire a été réalisée afin de mieux comprendre la distribution et la variabilité des données et de pouvoir détecter les valeurs aberrantes du jeu de données.

### 1.2.1 Valeurs numériques

L'analyse des variables numériques a été réalisée à l'aide de statistiques descriptives et de visualisations sous forme de boxplots, permettant d'identifier les éventuelles valeurs aberrantes.

Les cinq variables continues considérées sont : *Age*, *RestingBP*, *Cholesterol*, *MaxHR* et *Oldpeak*.

On remarque une anomalie claire pour la variable *Cholesterol*. En effet, un nombre important d'observations possèdent une valeur égale à 0, ce qui n'est pas physiologiquement possible. Ces valeurs sont donc très probablement le résultat d'une erreur de saisie ou d'un enregistrement manquant.

Mis à part ce point, aucune autre irrégularité notable n'a été détectée : les distributions de *RestingBP* et *Oldpeak* apparaissent plausibles, et aucune valeur extrême ne semble nécessiter un traitement particulier.

Cette première inspection confirme la bonne qualité générale des données numériques, à l'exception du taux de cholestérol, qui devra être corrigé ou imputé lors de l'étape de prétraitement.

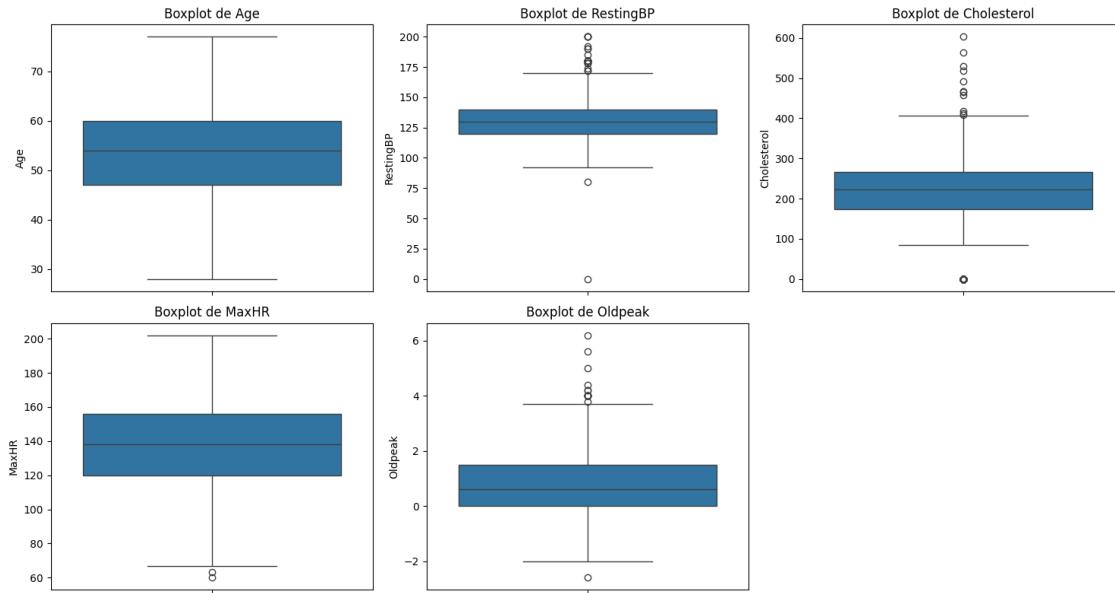


FIGURE 1 – Boxplots des principales variables numériques du jeu de données.

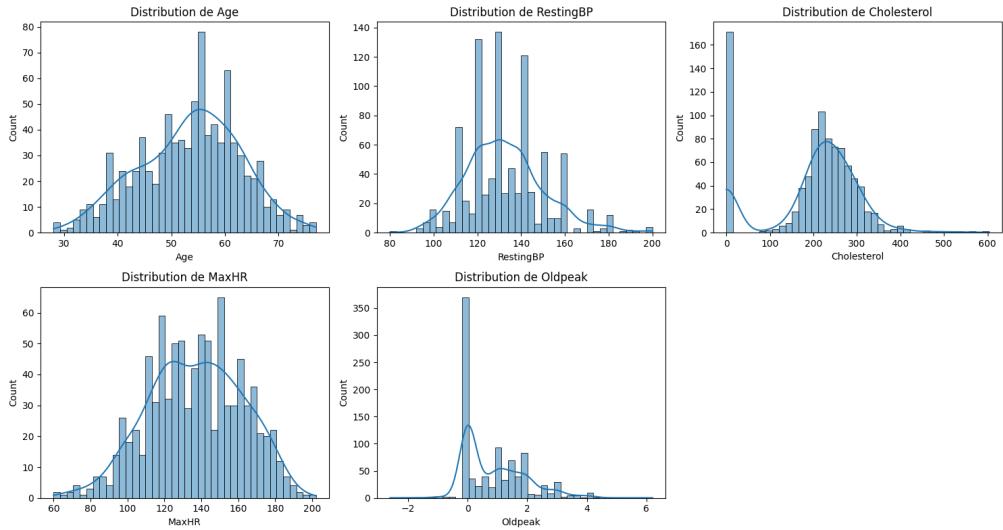


FIGURE 2 – Distribution des valeurs numériques

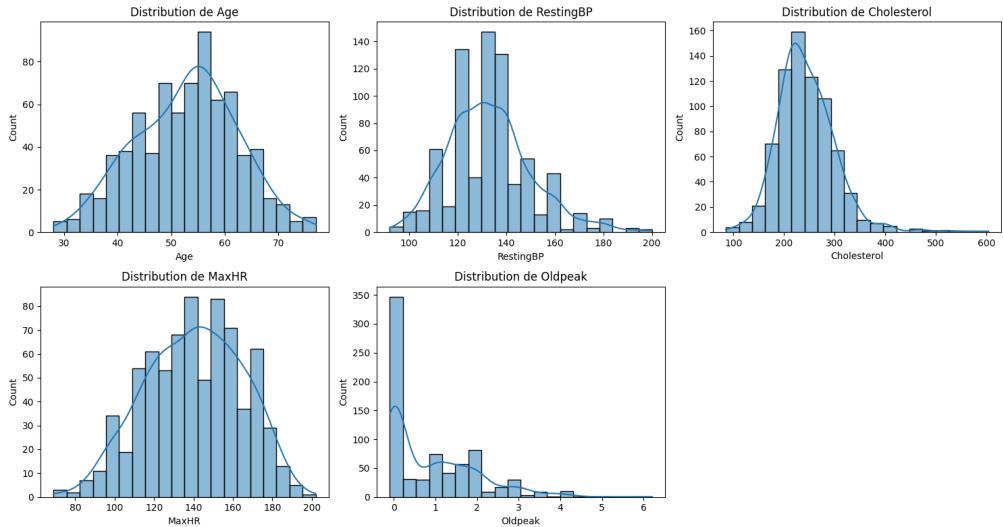


FIGURE 3 – Distribution des valeurs numériques après nettoyage.

On tombe alors à **746 observations** après ce nettoyage.

### 1.2.2 Valeurs catégorielles

L’analyse des variables catégorielles montre leur répartition globale et leur lien avec la présence d’une maladie cardiaque (*HeartDisease*).

**Sexe** Les hommes représentent 79% de l’échantillon et présentent une proportion plus élevée de maladies cardiaques (63%) que les femmes (26%).

**Type de douleur thoracique** La catégorie *ASY* est la plus fréquente (54%) et compte 79% de malades. Les autres types présentent des proportions plus faibles, indiquant que cette variable est discriminante.

**Électrocardiogramme au repos** Les ECG sont majoritairement *Normaux* (60%). Les patients avec *ST* ont une proportion de malades plus élevée (66%), suggérant un lien avec la pathologie.

**Angine à l'effort** 40% des patients présentent une angine (*Y*), parmi lesquels 85% sont malades, contre 35% sans angine. Cette variable est fortement discriminante.

**Pente du segment ST** Les pentes *Flat* (50%) et *Down* (7%) sont associées à plus de 77% de malades, tandis que *Up* concerne surtout les patients sains (80%).

**Glycémie à jeun** 23% des patients ont un taux élevé, parmi lesquels 79% sont malades, contre 48% pour les autres.

Les variables les plus discriminantes pour la présence d'une maladie cardiaque sont *ChestPainType*, *ExerciseAngina* et *ST\_Slope*, ce qui peut influencer la formation des clusters.

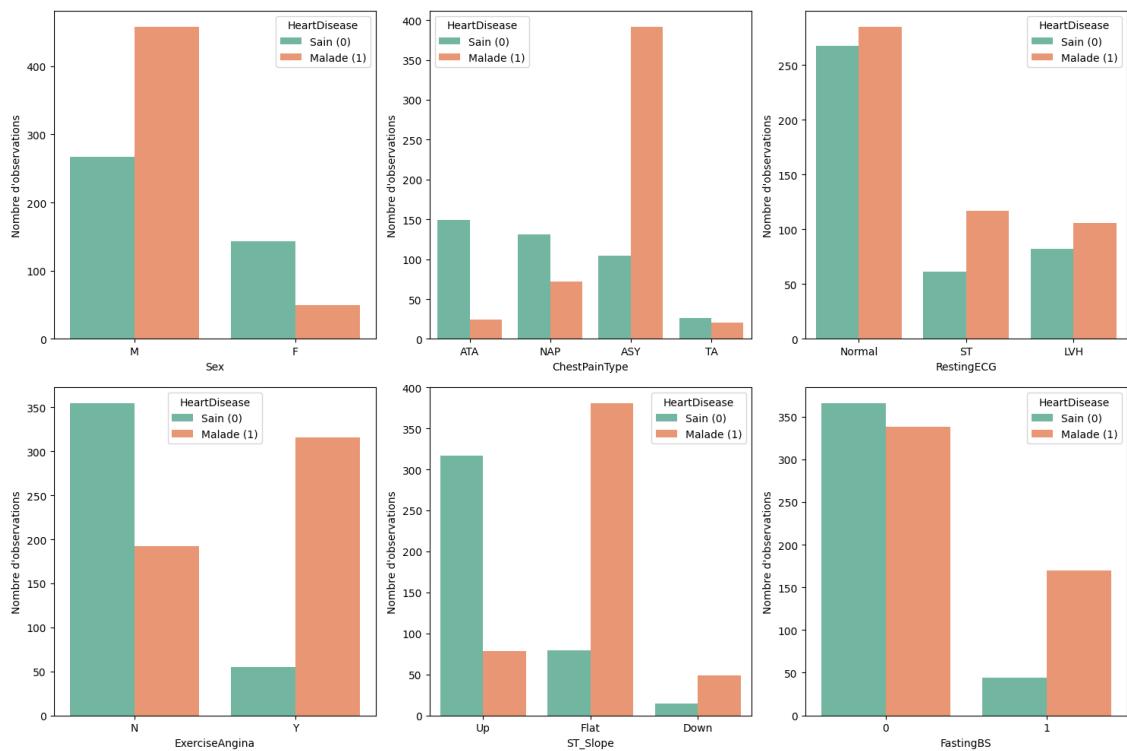


FIGURE 4 – Répartition des principales variables catégorielles selon la présence d'une maladie cardiaque.

## 2 Préparation des données

Avant d'appliquer les méthodes de clustering, les données ont été préparées de la manière suivante :

**Encodage des variables catégorielles** Les variables catégorielles (*Sex*, *ChestPainType*, *RestingECG*, *ExerciseAngina*, *ST\_Slope*) ont été transformées en variables numériques à l'aide de `get_dummies`, ce qui a permis de passer à un total de **21 colonnes** après encodage. Cette transformation est nécessaire pour que les algorithmes de clustering puissent utiliser correctement ces informations plutôt qu'un One Hot Encoder.

**Mise à l'échelle des variables numériques** Les variables continues (*Age*, *RestingBP*, *Cholesterol*, *MaxHR*, *Oldpeak*, *FastingBS*) ont été standardisées ( $\text{mean}=0$ ,  $\text{std}=1$ ) à l'aide d'un `StandardScaler`. Le scaling est particulièrement important pour les algorithmes de clustering :

- **K-Means** utilise les distances euclidiennes : si les variables sont sur des échelles différentes, certaines auront un poids disproportionné et biaiseront la formation des clusters.
- **DBSCAN** se base sur les distances et densités : des variables non-scalées peuvent rendre certains points artificiellement isolés ou regroupés.
- **Gaussian Mixtures** modélise les distributions : la mise à l'échelle aide à ce que chaque dimension contribue de façon équilibrée à la probabilité des clusters.

**Réduction de dimension** Une analyse en composantes principales (PCA) a été appliquée sur l'ensemble des variables (après encodage et scaling) afin de réduire la dimensionnalité tout en conservant **90% de la variance** du dataset. Cette étape a permis de passer de 21 colonnes à **9 composantes principales** que nous utiliserons pour le clustering. La PCA aide à limiter le bruit et les redondances (corrélations) dans les données, tout en améliorant la vitesse et la stabilité des algorithmes de clustering.

Après ces transformations, le dataset est prêt pour l'application des différents algorithmes, avec des variables toutes sur une échelle comparable et une dimensionnalité réduite pour une meilleure interprétation des clusters.

### 3 K-Means

Dans un premier temps, nous avons étudié la méthode K-Means. Pour ce faire, nous avons suivi la méthodologie suivante :

- préparation et nettoyage du jeu de données en réalisant une suppression des valeurs aberrantes
- normalisation des variables quantitatives et encodage des variables catégorielles
- sélection des variables les plus pertinentes via les tests de corrélation Spearman et Chi<sup>2</sup>
- application du clustering K-Means sur les données transformées

- visualisation des résultats obtenus par Kmean avec une réduction de dimension en 2D via la méthode PCA

L'évaluation du modèle a été réalisée à l'aide de plusieurs métriques internes (Inertie, Silhouette, Calinski-Harabasz, et Davies-Bouldin). Afin d'avoir une visualisation plus parlante, nous l'avons complétée par une visualisation des clusters après réduction de dimension par PCA. À chaque étape des exécutions et des applications intermédiaires de la méthode K-Means ont été réalisées afin d'avoir une évolution tout au long de l'implémentation de features pour le traitement des données.

### **3.1 Fonctionnement de K-means**

L'algorithme K-Means est une méthode de clustering non supervisée. Elle vise à regrouper les données en  $k$  ensembles distincts (clusters). Le principe repose sur la minimisation de la distance entre les points et le centre de leur groupe (centroïde). Dans un premier temps, l'algorithme choisit aléatoirement  $k$  centroïdes, puis attribue chaque point au cluster dont le centroïde est le plus proche. Les centroïdes sont par la suite mis à jour en calculant la moyenne des points qui leur sont associés.

L'idée est donc de créer des groupes compacts et bien séparés en se basant sur la distance euclidienne entre un point et le centroïde le plus proche. Comme dit plus haut, K-Means présente quand même certaines limites : il part du principe que les clusters sont sphériques et de taille comparable. En plus de ceci, il est sensible à la normalisation des données ainsi qu'à la présence d'outliers.

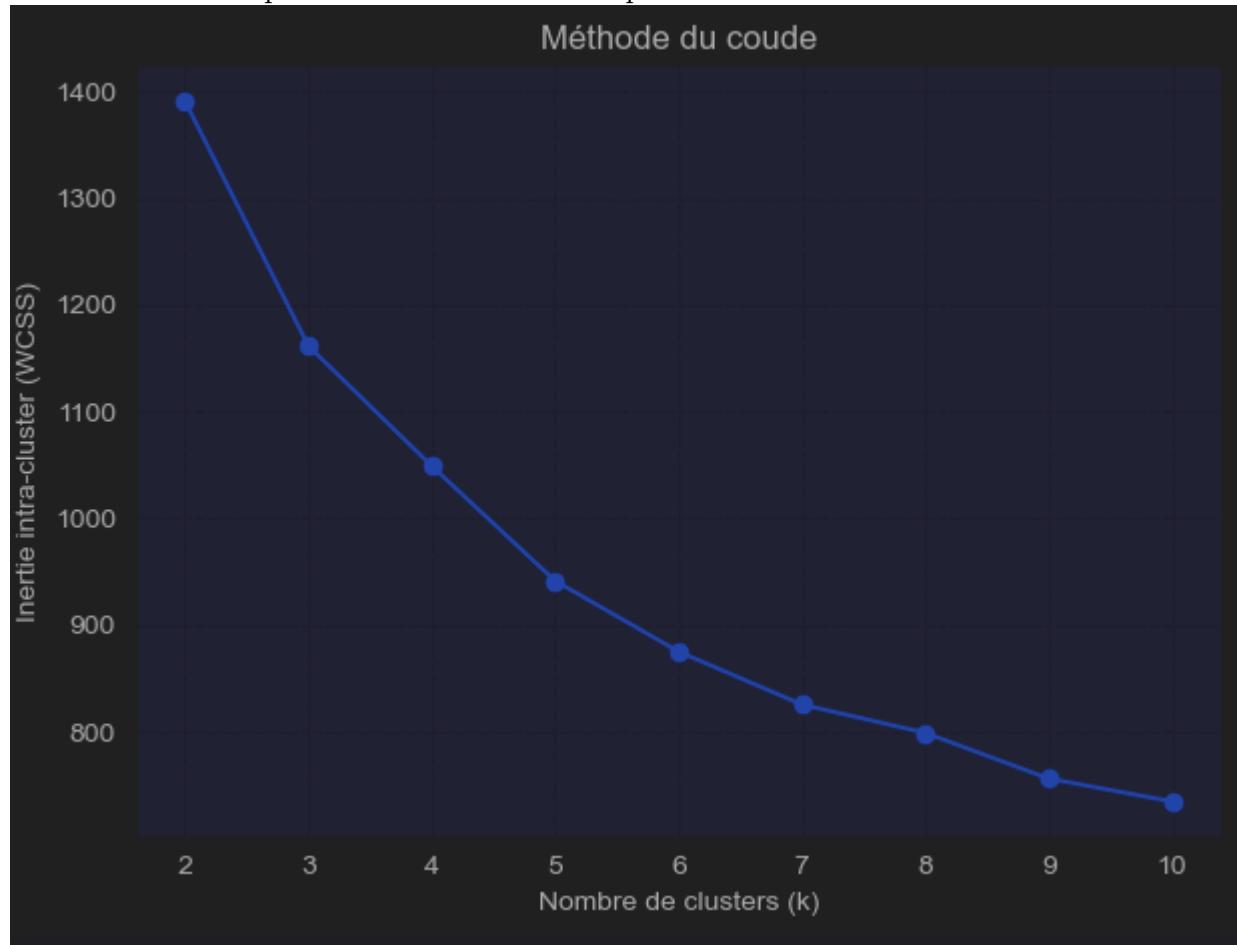
### **3.2 Sélection du nombre de clusters et évaluation de la qualité de ces clusters**

Comme expliqué précédemment, une des particularités de la méthode K-means est qu'il faut choisir en amont un nombre de clusters  $K$ . Afin de choisir un nombre de clusters approprié au problème et de ne pas sous-partitionner ou surpartitionner, nous avons utilisé plusieurs métriques d'évaluations ainsi que la méthode du coude (Elbow Method).

#### **3.2.1 La méthode du coude**

La méthode du coude est une technique utilisée pour déterminer le nombre optimal de clusters ( $K$ ) dans l'algorithme K-Means. Elle repose sur une métrique appelée « inertie ». L'inertie correspond à la somme des distances au carré entre les points et leur centroïde. Pour utiliser cette métrique, on exécute l'algorithme pour plusieurs valeurs de  $K$  différentes (dans notre cas de 2 à 11) et on trace l'évolution de l'inertie en fonction du nombre de clusters  $K$ . Une particularité de l'inertie est qu'elle diminue naturellement lorsque  $K$  augmente, puisque plus de clusters permettent de mieux regrouper les points. À partir d'un certain seuil, la

réduction de l'inertie devient minime. Cette baisse minime entraîne le fait qu'il y a alors une cassure dans la courbe qui se forme (le coude). Ce point marque le meilleur compromis entre la précision du regroupement et la simplicité du modèle. La valeur K correspondant à ce coude est alors considérée comme la plus pertinente. Vous pourrez retrouver ci-dessous la dernière courbe correspondant à cette méthode que nous avons obtenu.



### 3.2.2 Métriques utilisées

Afin de compléter la méthode du coude, et de ne pas juste se fier à l'inertie, nous avons décidé d'utiliser plusieurs autres métriques qui sont les suivantes :

Métrique	Description	Objectif / Interprétation
<b>Inertie</b>	Somme des distances au carré entre chaque point et le centroïde de son cluster.	Mesure la <b>dispersion interne</b> : plus elle est faible, mieux les points sont regroupés.
<b>Silhouette Score</b>	Moyenne du rapport entre la distance intra-cluster et la distance au cluster le plus proche.	Évalue la <b>cohésion et la séparation</b> : plus il est proche de 1, meilleurs sont les clusters.
<b>Indice de Calkinski–Harabasz</b>	Rapport entre la variance inter-cluster et la variance intra-cluster.	Valeur élevée = <b>clusters denses et bien séparés</b> .
<b>Indice de Davies–Bouldin</b>	Moyenne des rapports entre la dispersion intra-cluster et la distance inter-cluster.	Valeur faible = <b>clusters distincts et bien formés</b> .

TABLE 2 – Résumé des métriques d'évaluation utilisées pour le clustering K-Means.

### 3.3 Prétraitement des données

La méthode K-Means étant sensible à l'échelle des variables et au type des variables (continue ou catégorie), cela nous oblige à réaliser plusieurs prétraitements des données si nous voulons obtenir des résultats cohérents.

#### 3.3.1 Suppression des outliers et des valeurs aberrantes

La première étape de prétraitement des données que nous avons réalisé a donc été d'enlever les valeurs continues qui n'avaient pas forcément de sens, on peut retrouver dans notre dataset deux colonnes « RestingBP » ainsi que « Cholesterol » qui sont des variables continues mais qui ont pour certaine ligne pour valeur 0. Médicalement ça ne fait pas sens, nous les avons donc enlevé manuellement. On pourra d'ailleurs remarquer que ces valeurs à 0 provoquaient de fausse relation entre point et cluster. Pour ce qui est de la suite afin d'avoir du recul sur la pertinence de la méthode utilisée nous avons fait le choix d'implémenter deux méthodes. Ces deux méthodes sont les suivantes :

Méthode	Description
<b>IQR (Interquartile Range)</b>	Cette méthode repose sur les quartiles (valeur statistique qui divise un ensemble de données triées en quatre parties égales) statistiques. Cette méthode considère comme valeurs aberrantes tous les points situés en dehors de l'intervalle défini par $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ , où $IQR = Q3 - Q1$ . Elle est simple à mettre en œuvre et adaptée aux variables numériques continues.
<b>Isolation Forest</b>	Méthode d'apprentissage non supervisée basée sur des arbres de décision. Elle isole les observations anormales en construisant des partitions aléatoires : plus un point est isolé rapidement, plus il est probable qu'il s'agisse d'un outlier.

TABLE 3 – Méthodes utilisées pour la détection et la suppression des valeurs aberrantes

En exécutant le code correspondant dans le notebook, libre à vous de choisir quelle méthode vous voulez utiliser, les deux méthodes ont été plutôt pertinentes pour le nettoyage de données.

### 3.3.2 Discrimination des variables en fonction de leur corrélation avec la target

N'obtenant pas de résultats suffisamment concluants avec nos premières expérimentations de clustering, nous avons choisi de réaliser une phase de discrimination des variables afin d'identifier celles qui avaient le plus d'influence sur la variable cible (*HeartDisease*). Pour cela, nous avons appliqué deux approches distinctes selon la nature des variables :

- Pour les **variables quantitatives**, nous avons utilisé le **coefficient de corrélation de Spearman**. Cette méthode permet de mesurer la force et la direction de la relation entre chaque variable numérique et la target. Nous avons fixé un seuil de corrélation stricte à 0,3 afin de ne conserver que les variables ayant un lien significatif avec la présence d'une maladie cardiaque.
- Pour les **variables catégorielles**, nous avons réalisé un test **Chi<sup>2</sup>**. Cette méthode permet de vérifier l'existence d'une dépendance statistique entre chaque variable catégorielle et la target. Les variables dont la *p-value* était inférieure à 0,05 ont été retenues comme significatives.

Cette étape avait pour objectif de réduire la dimensionnalité des données. De cette manière, la méthode K-means peut se concentrer davantage sur les variables et les corrélations qui ont le plus de sens afin d'avoir un clustering plus précis. Les résultats que nous avons obtenus sont les suivants :

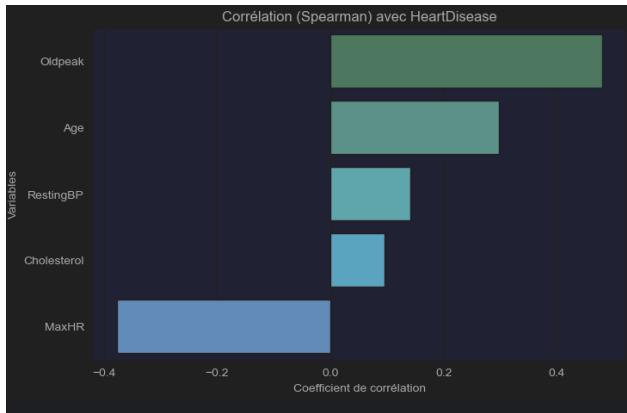


FIGURE 5 – Graphique montrant la corrélation des variables continues avec la Target (méthode spearman)

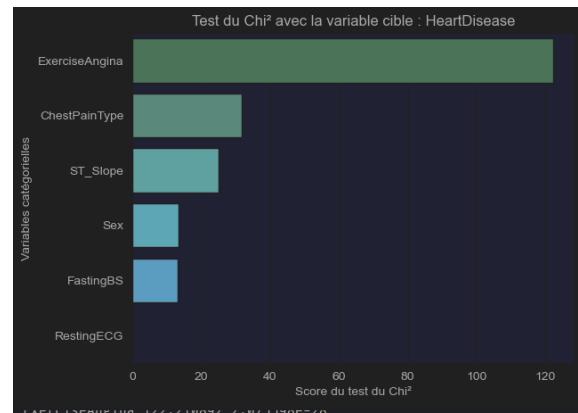


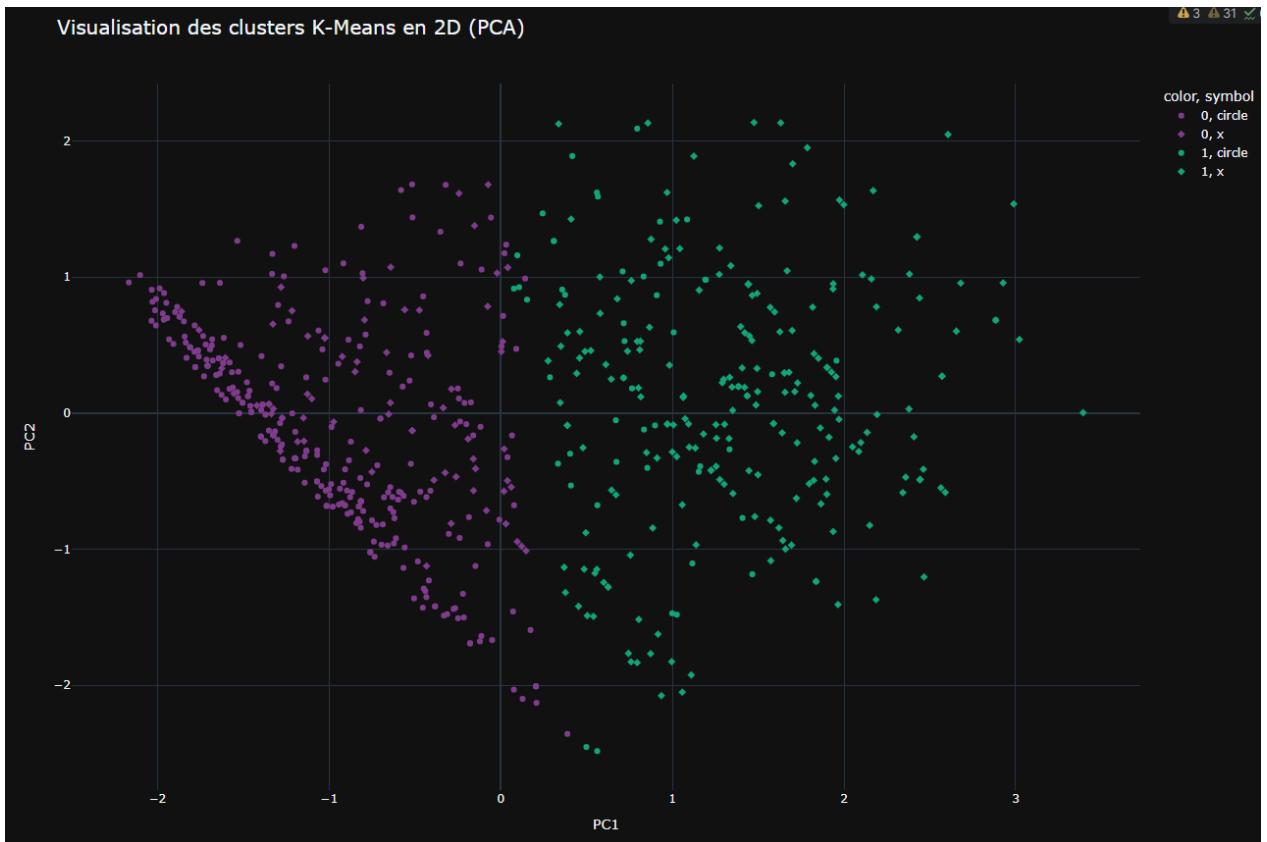
FIGURE 6 – Graphique montrant la corrélation des variables catégorielles avec la Target (méthode Chi<sup>2</sup>)

Avec un filtrage à 0,3 pour les variables quantitatives et à 0,01 pour les variables catégorielles, les variables dépassant leur seuil de corrélation sont les suivantes :

- **Variables quantitatives sélectionnées :**
  - Oldpeak
  - MaxHR
- **Variables catégorielles sélectionnées :**
  - ExerciseAngina
  - ChestPainType
  - ST\_Slope
  - Sex
  - FastingBS

### 3.4 Conclusion et résultats

Après toutes ces étapes de prétraitement des données et de sélection de variables, voici le résultat final de clustering obtenu :



La projection 2D obtenue par la PCA suite à cette dernière application de la méthode K-means met en valeur deux clusters assez distincts. Cela laisse penser que l'algorithme a détecté une structure binaire au sein du jeu de données. Cette séparation, pourtant ne correspond pas explicitement à la variable cible (HeartDisease) : les individus malades et non malades sont répartis dans les deux clusters.

Cluster	<b>HeartDisease = 0</b>	<b>HeartDisease = 1</b>
<b>0</b>	296	70
<b>1</b>	52	210

TABLE 4 – Répartition des individus selon les clusters et la présence de maladie cardiaque

On observe que le cluster 0 contient majoritairement des individus sans maladie cardiaque (296 vs 70), tandis que le cluster 1 regroupe principalement des individus avec maladie cardiaque (210 vs 52). Cela peut nous laisser penser que K-Means a réussi à capturer une séparation partielle entre les deux groupes, bien que la séparation ne soit pas parfaite.

Il est donc probable que K-Means ait surtout capté des différences générales de profil physiologique plutôt qu'une distinction directe liée à l'état de santé.

Cette observation nous montre les limites de la méthode K-means dans ce contexte. Il n'est pas exclu que notre approche ou nos choix de variables puissent influencer ce résultat.

D'autres techniques de clustering, comme DBSCAN ou les modèles de mélange gaussien (GMM), pourraient offrir une meilleure prise en compte de la complexité des données et

révéler des structures plus pertinentes vis-à-vis de la présence d'une maladie cardiaque. Ces pistes seront explorées dans la suite du projet.

En conclusion, la méthode K-Means appliquée à notre dataset cardiaque permet de dégager une structure interne cohérente dans les données, mais elle ne permet pas de reproduire la séparation entre malades et non-malades. L'analyse des métriques (Silhouette, Calinski-Harabasz, Davies-Bouldin) nous montre bien que le modèle est globalement stable avec deux clusters. Cependant l'interprétation médicale que nous pouvons en faire reste assez floue.

## 4 Gaussian Mixture Model (GMM)

Après K-Means, nous considérons le modèle de mélange gaussien (GMM). Contrairement à K-Means qui impose des clusters sphériques de taille similaire, le GMM repose sur une approche probabiliste permettant de modéliser des clusters de formes plus variées et éventuellement chevauchants. Chaque cluster est représenté par une distribution gaussienne définie par une moyenne  $\mu$  et une matrice de covariance  $\Sigma$ .

### 4.1 Prétraitement : suppression des valeurs aberrantes

Nous avons appliqué **Isolation Forest** afin de retirer les observations atypiques, ce qui a conduit à supprimer **118 points** et à passer de **746 à 628 individus**, permettant d'obtenir des clusters plus nets et plus cohérents.

### 4.2 Choix de la matrice de covariance

Le GMM propose quatre configurations selon la manière dont la covariance est estimée :

- **full** : une covariance complète par cluster (flexible mais risque de sur-ajustement).
- **tied** : une seule covariance partagée (bon compromis stabilité / expressivité).
- **diag** : seulement les variances individuelles (rapide mais peu descriptif).
- **spherical** : variance identique dans toutes les directions (clusters trop rigides).

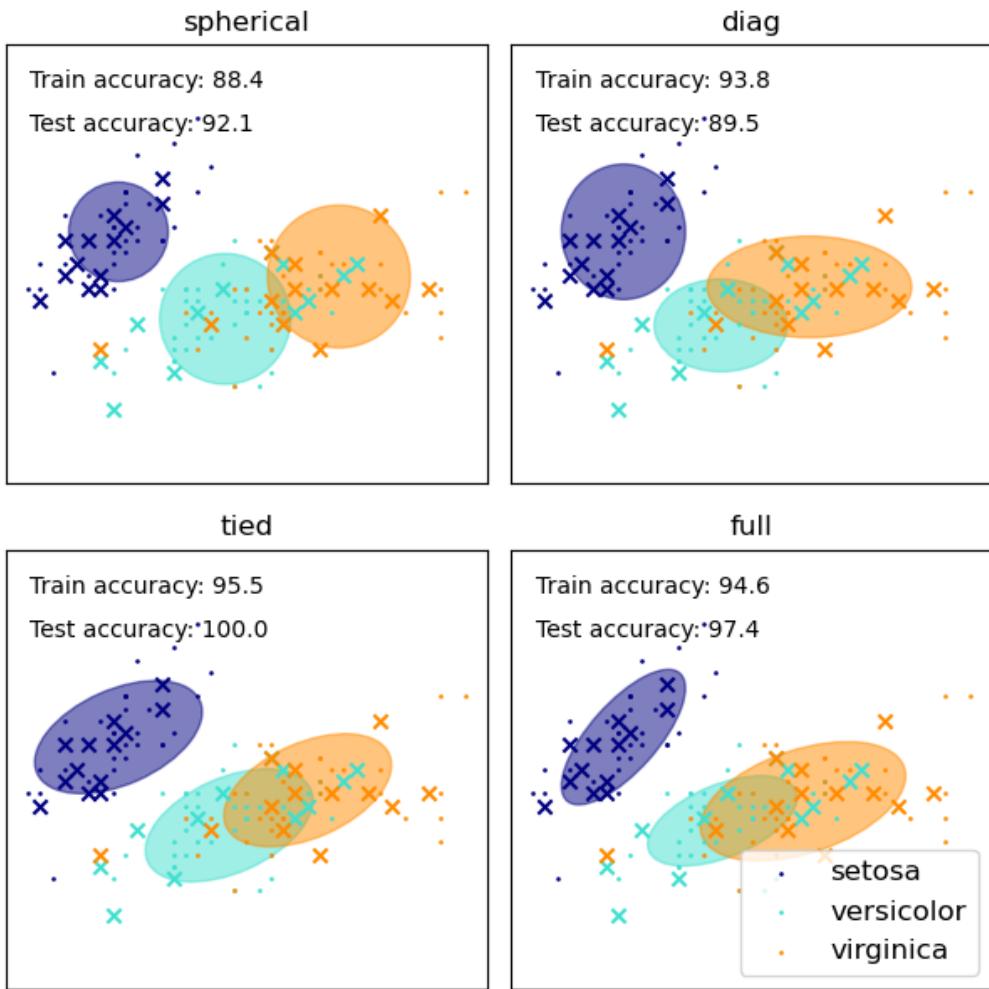


FIGURE 7 – Illustration des quatre types de matrices de covariance.

#### 4.2.1 Choix du nombre de clusters et du type de covariance

Nous avons comparé les variantes du GMM à l'aide de deux indicateurs :

- **Silhouette** : qualité de séparation entre les clusters.
- **Log-vraisemblance moyenne (ad hoc)** : qualité d'ajustement du modèle.

Le score de silhouette est maximal pour  $K = 2$ , ce qui indique une séparation nette entre deux groupes. Le score ad hoc augmente légèrement lorsque  $K$  augmente, mais sans amélioration significative, ce qui suggère qu'un nombre plus élevé de clusters apporterait surtout du **sur-ajustement**.

La covariance **full** obtient les meilleurs scores pour  $K = 2$ , tout en conservant une structure claire et interprétable.

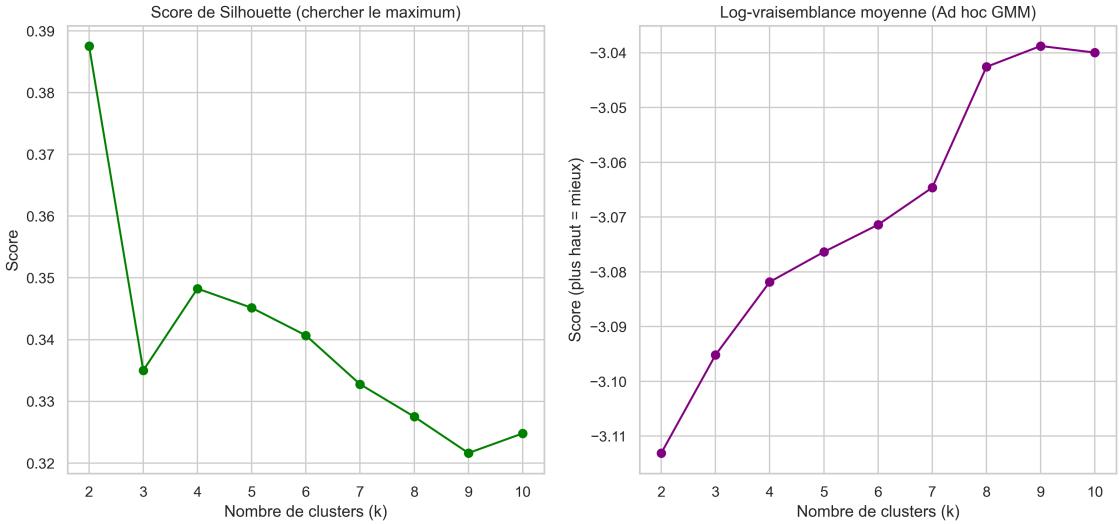


FIGURE 8 – Évolution du score de silhouette (gauche) et du score ad hoc (droite) en fonction de  $K$ . Le compromis optimal est atteint pour  $K = 2$ .

**Modèle retenu : GMM full,  $K = 2$  clusters.**

#### 4.2.2 Visualisation et interprétation des clusters

Après avoir retenu un modèle GMM avec `covariance_type = full` et  $K = 2$ , nous visualisons maintenant la structure des clusters obtenus. Étant donné que les données sont multidimensionnelles, nous utilisons deux méthodes de réduction de dimension pour les représenter en 2D :

- **PCA** : préserve la variance globale et la structure linéaire des données.
- **t-SNE** : met en évidence les proximités locales entre individus.

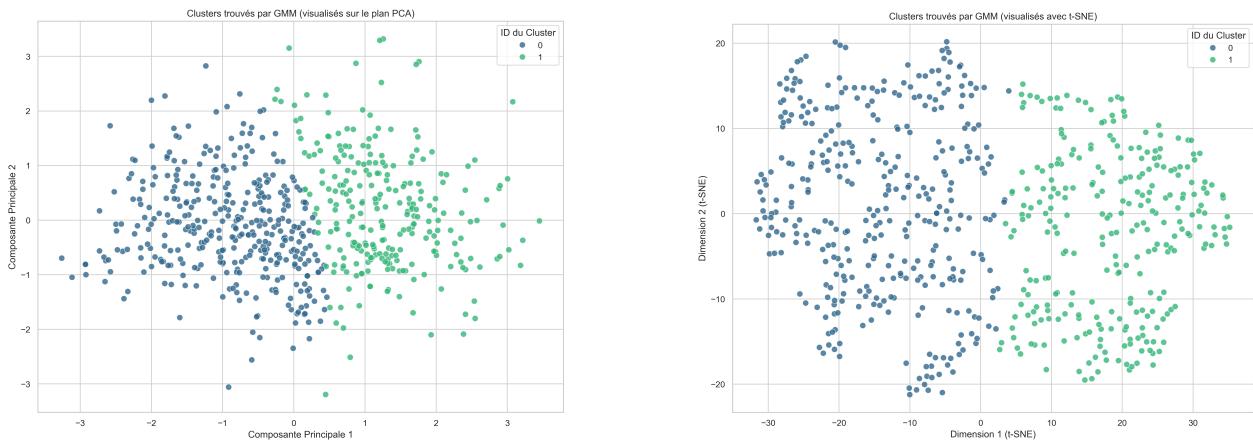


FIGURE 9 – Visualisation des clusters GMM (**full**,  $K = 2$ ) via PCA (gauche) et t-SNE (droite).

Les deux représentations montrent une séparation nette entre les deux groupes. La version **t-SNE** met davantage en évidence la distance visuelle entre les clusters en amplifiant les différences locales, ce qui confirme l'existence de deux sous-populations cohérentes.

Cependant, la **PCA** propose une visualisation plus stable et directement interprétable, car elle conserve la structure globale des données. Pour la suite de l'analyse, nous nous appuierons donc principalement sur la représentation PCA, qui permet une lecture plus fiable des tendances générales.

Cluster	Nombre d'individus
0	366
1	262

TABLE 5 – Effectif de chaque cluster.

**Répartition des effectifs.** La répartition des individus entre les deux clusters est équilibrée, mais avec une légère majorité dans le **Cluster 0** (366 individus) par rapport au **Cluster 1** (262 individus).

#### 4.2.3 Variables discriminantes : Feature Importance (XAI)

Afin d'expliquer la formation des deux clusters, nous avons entraîné un arbre de décision interprétable (faible profondeur), permettant d'identifier les variables ayant le plus contribué à leur séparation. Cette démarche s'inscrit dans une logique d'explicabilité des modèles (XAI) vue en cours.

Pour valider cette première analyse, nous avons également appliqué une méthode agnostique par permutation, consistant à perturber aléatoirement les valeurs d'une variable à la fois. Si la performance du modèle diminue fortement, alors la variable est considérée comme importante. Cette approche, indépendante du modèle, renforce la robustesse de l'interprétation.

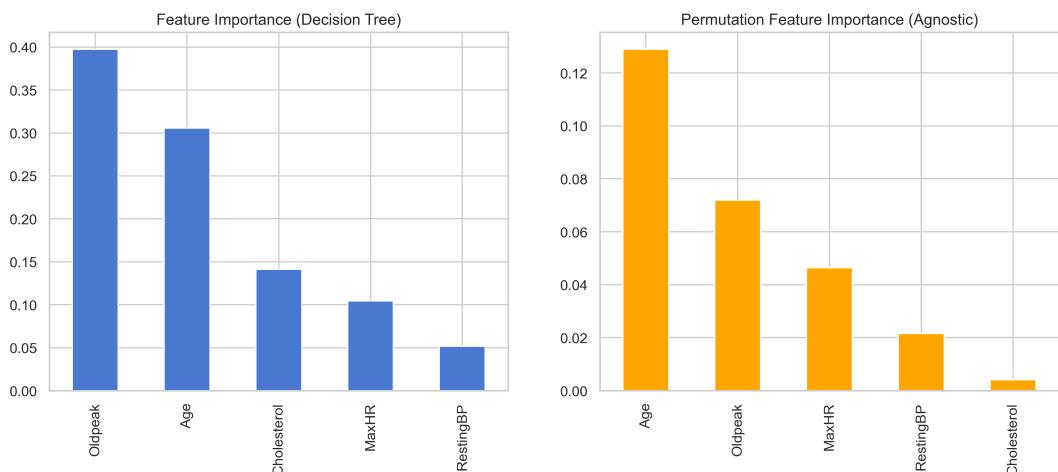


FIGURE 10 – Importance des variables selon l'arbre de décision (gauche) et l'analyse par permutation (droite).

Les deux méthodes convergent vers les mêmes variables discriminantes :

- **Oldpeak** (variation du signal mesurée lors de l'effort),
- **Âge**,
- **MaxHR** (fréquence cardiaque maximale mesurée),
- **Cholestérol** (rôle secondaire),
- **RestingBP** (impact plus faible).

Même si le **Cholestérol** n'est pas déterminant dans la séparation automatique, on observe des valeurs légèrement plus élevées dans le Cluster 1, ce qui reste cohérent avec les profils identifiés.

#### 4.2.4 Répartition de la maladie selon les clusters

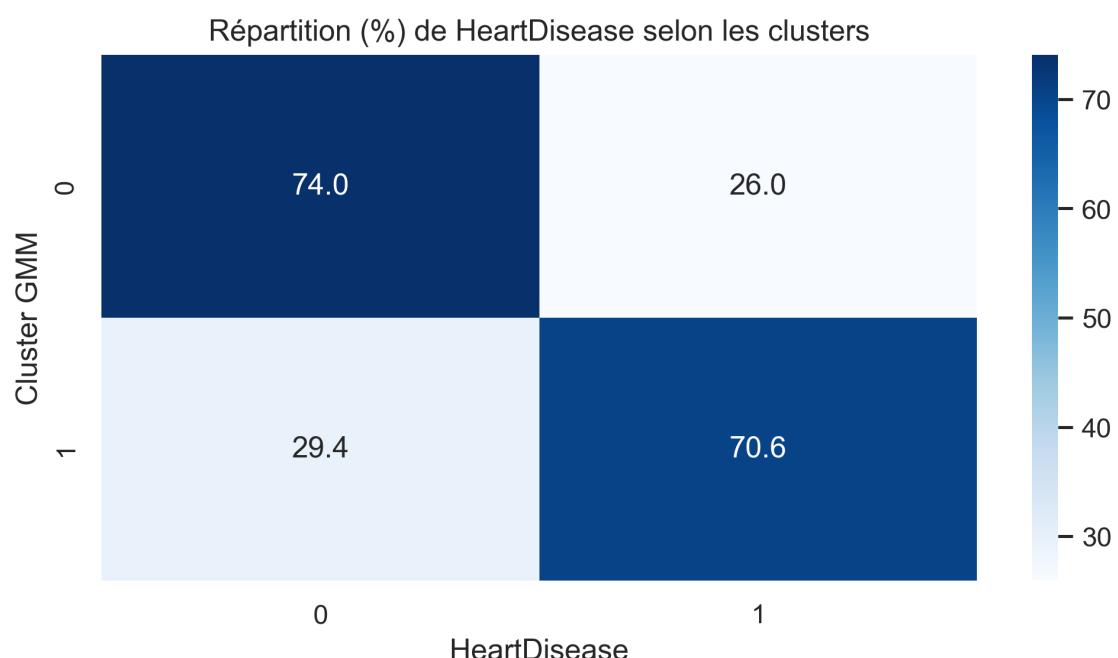


FIGURE 11 – Proportions des individus atteints selon les clusters.

On observe une différence claire entre les deux clusters :

- **Cluster 0** : environ 26% d'individus atteints — groupe **majoritairement non atteint**.
- **Cluster 1** : environ 71% d'individus atteints — groupe **où la maladie est fréquente**.

Ainsi, le **Cluster 0** peut être interprété comme un groupe à **faible occurrence**, tandis que le **Cluster 1** regroupe des individus présentant **un risque plus élevé**. Cette distinction servira de base pour analyser plus finement les caractéristiques associées à chaque cluster.

#### 4.2.5 Analyse des variables continues

Les variables continues permettent de caractériser plus précisément les deux clusters. Les différences se jouent principalement sur : **Oldpeak**, **Âge**, et **MaxHR**, puis dans une moindre mesure **Cholestérol** et **RestingBP**.

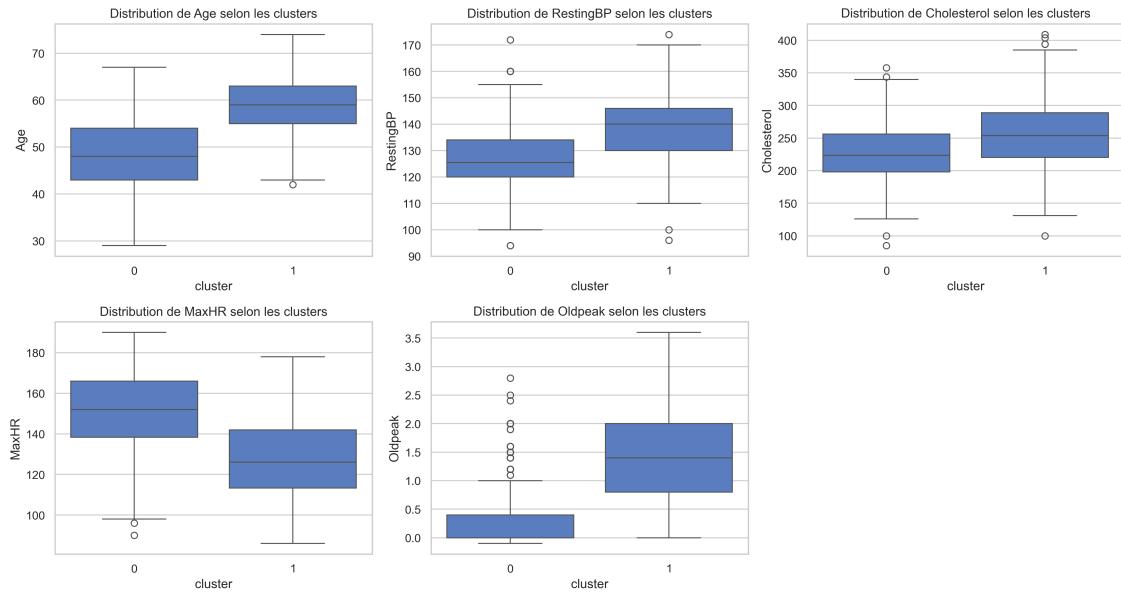


FIGURE 12 – Distribution des variables continues selon les clusters (boxplots).

**Cluster 0 :** Les individus sont en moyenne **plus jeunes**, avec une **fréquence cardiaque maximale plus élevée** et des valeurs **Oldpeak faibles**. Cela indique une **meilleure capacité de réponse à l'effort** et des indicateurs physiologiques globalement favorables.

**Cluster 1 :** Ce groupe est **plus âgé**, présente une **fréquence cardiaque maximale plus basse** et des valeurs **Oldpeak plus élevées**. Ces éléments traduisent une **réponse à l'effort plus limitée** ainsi qu'une **variation plus marquée des indicateurs mesurés lors de l'effort**.

Dans l'ensemble, la séparation entre les clusters repose donc surtout sur **la capacité de réponse à l'effort et l'âge**, ce qui correspond à la hiérarchie d'importance observée précédemment dans l'analyse XAI.

#### 4.2.6 Analyse des variables catégorielles

Les variables catégorielles viennent compléter l'analyse en montrant des différences nettes dans les profils des deux clusters.

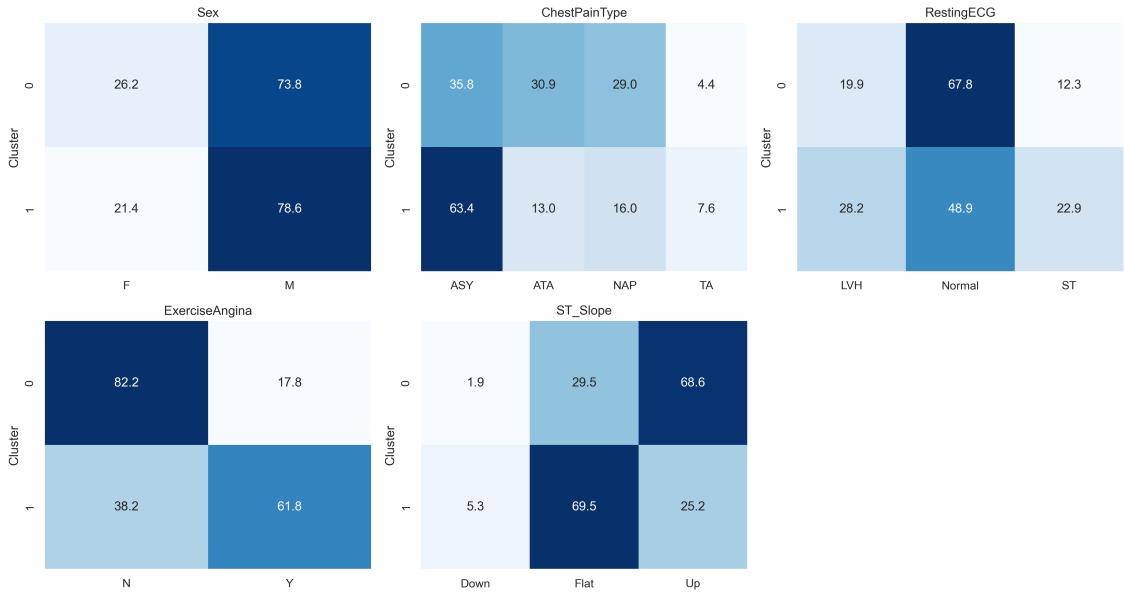


FIGURE 13 – Proportions des principales variables catégorielles selon les clusters.

**Cluster 0 :** Les individus présentent plus souvent des **douleurs peu spécifiques** et une pente **ST\_Slope ascendante**. Ces éléments sont cohérents avec une **bonne réponse générale à l'effort**. De plus, l'**angine d'effort est moins fréquente** dans ce groupe.

**Cluster 1 :** On observe une proportion plus élevée de **douleurs typiques lors de l'effort** et une **pente ST\_Slope plate**, ainsi qu'une **angine d'effort plus fréquente**. Ces caractéristiques traduisent une **réaction à l'effort plus limitée**.

Dans l'ensemble, les variables catégorielles confirment la distinction déjà observée avec les variables continues : le **Cluster 0** correspond à un **profil plus favorable**, tandis que le **Cluster 1** regroupe des individus présentant **des signes plus marqués lors de l'effort**.

#### 4.2.7 Conclusion : Interprétation globale des clusters

L'analyse conjointe des variables continues et catégorielles met en évidence deux profils bien distincts dans les données.

##### Cluster 0 : Profil « réponse à l'effort préservée »

Les individus de ce cluster sont en moyenne plus jeunes et présentent une fréquence cardiaque maximale plus élevée, associée à des valeurs d'Oldpeak faibles. Leur réponse à l'effort est plus stable et homogène, ce qui se traduit par une **faible proportion d'individus atteints** dans ce groupe.

##### Cluster 1 : Profil « réponse à l'effort limitée »

Ce cluster regroupe des individus plus âgés, avec une fréquence cardiaque maximale plus basse et des valeurs d'Oldpeak plus élevées. La réaction à l'effort est plus contrastée et moins

efficace, ce qui se reflète par une **proportion nettement plus élevée d'individus atteints**.

## 5 Clustering avec DBSCAN

### 5.1 Présentation de l'algorithme DBSCAN

L'algorithme **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) est une méthode de clustering non supervisée reposant sur la **densité des points** dans l'espace des données. Contrairement à **K-Means**, il ne nécessite pas de spécifier le nombre de clusters à l'avance et peut identifier des groupes de formes arbitraires ainsi que des observations isolées considérées comme du *bruit*.

Deux paramètres principaux guident le fonctionnement de DBSCAN :

- **eps** : la distance maximale entre deux points pour être considérés comme voisins ;
- **min\_samples** : le nombre minimum de voisins requis pour qu'un point soit considéré comme appartenant à un cluster dense.

### 5.2 Recherche des hyperparamètres de DBSCAN

La recherche des hyperparamètres de DBSCAN s'appuie sur la méthode des *k-distances*. Pour chaque point, la distance à son *k*-ième plus proche voisin est calculée, puis les distances triées sont tracées. Le point d'inflexion de la courbe indique la valeur optimale de  $\varepsilon$  (*eps*), correspondant à une transition entre zones denses et zones clairsemées.

Une **grid search** a été effectuée afin de minimiser le bruit et maximiser le score de silhouette. Une fois les valeurs candidates obtenues, une **ajustement manuel** a été réalisé pour équilibrer la taille des clusters et assurer une séparation cohérente entre groupes.

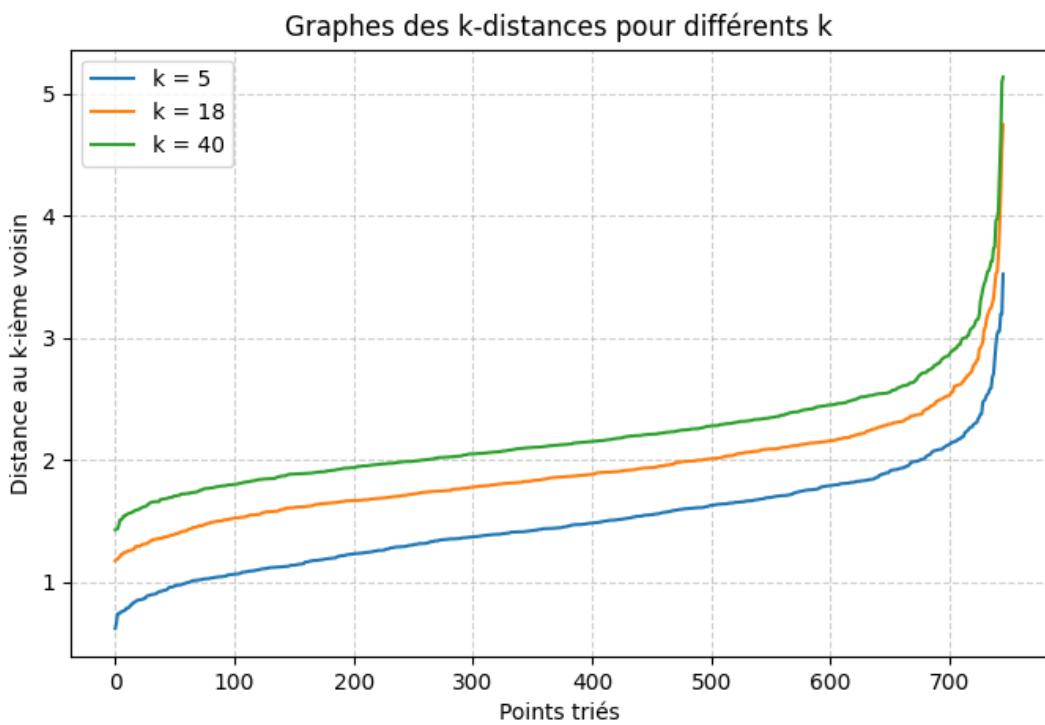


FIGURE 14 – Courbes des k-distances sur les données PCA ( $k = 5$ ,  $k = 18$ ,  $k = 40$ ).

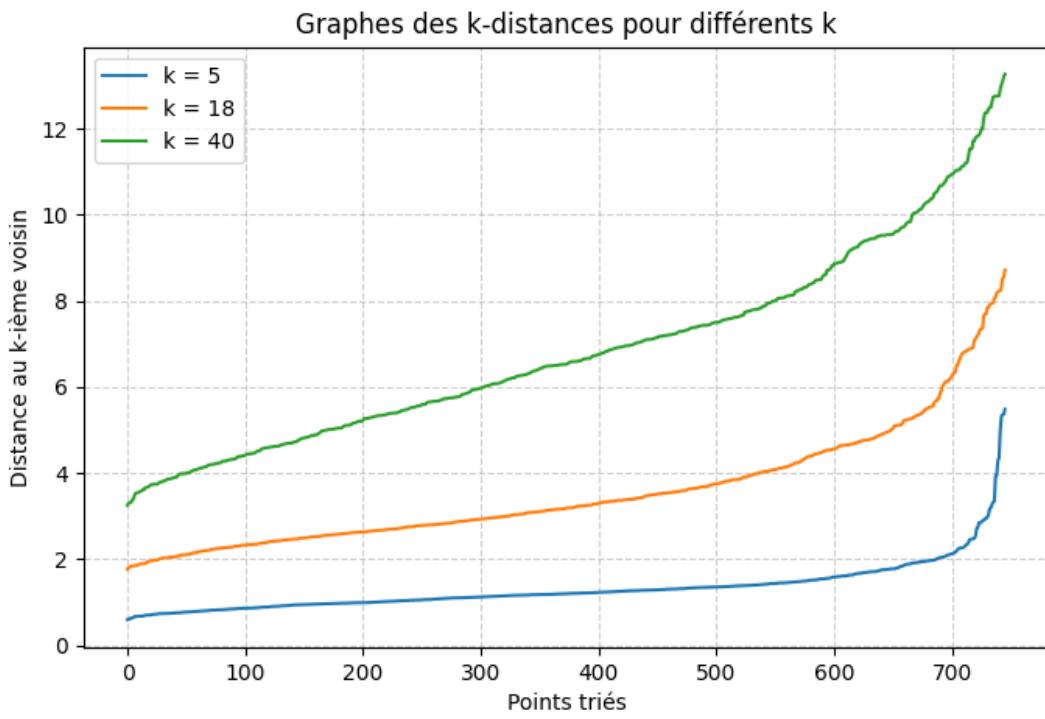


FIGURE 15 – Courbes des k-distances sur les données t-SNE ( $k = 5$ ,  $k = 18$ ,  $k = 40$ ).

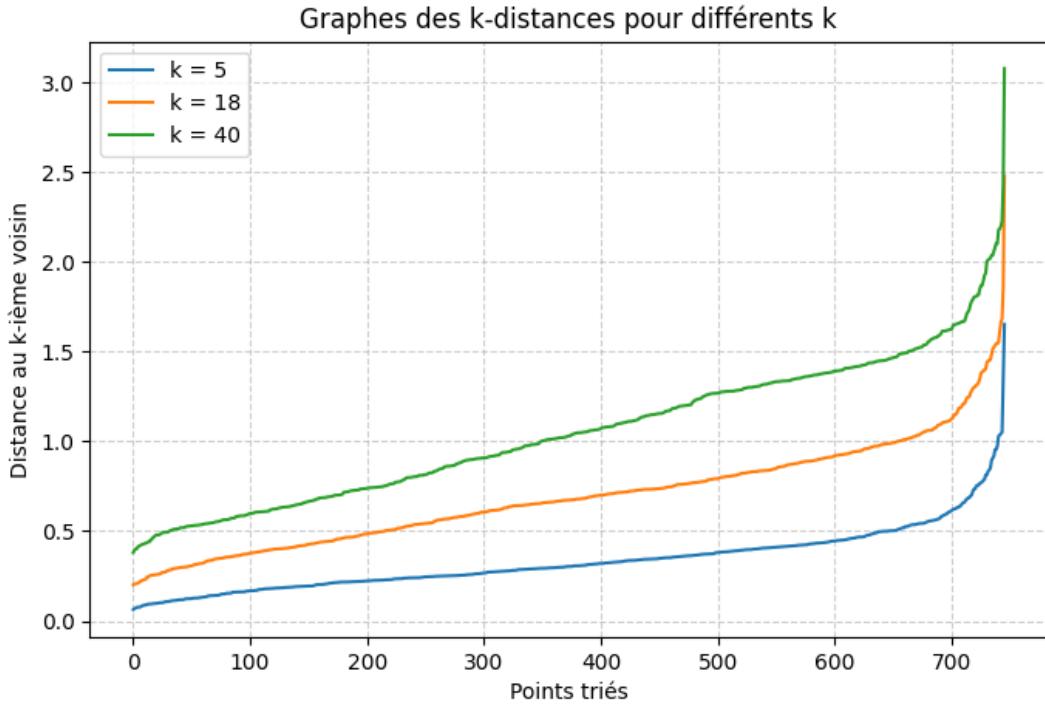


FIGURE 16 – Courbes des k-distances sur les données Isomap ( $k = 5, k = 18, k = 40$ ).

### 5.3 Résultats DBSCAN sur les données PCA

Avec les paramètres ( $\text{eps} = 1.78, \text{min\_samples} = 38$ ), DBSCAN a identifié deux clusters principaux, mais une forte proportion de points considérés comme du bruit.

Cluster_Label	Total_Points	Nb_Malades	Taux_Maladie
-1	201	104	51.74%
0	268	221	82.46%
1	277	31	11.19%

TABLE 6 – Résultats DBSCAN sur PCA ( $\text{eps}=1.78, \text{min\_samples}=38$ ).

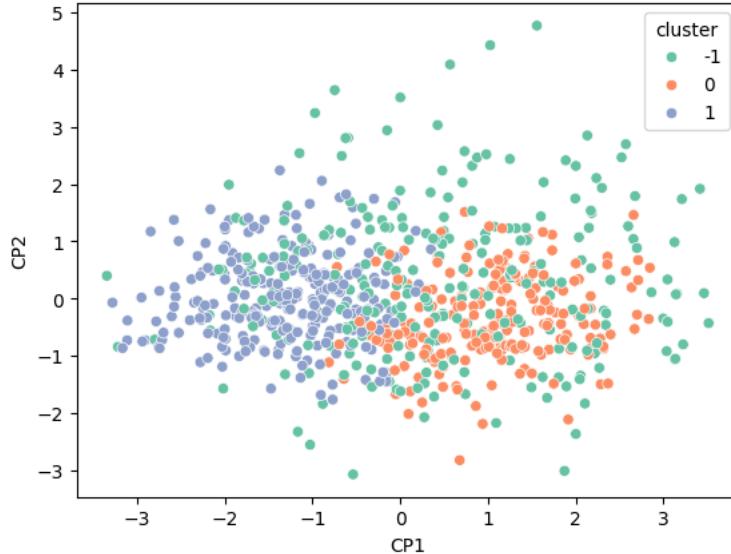


FIGURE 17 – Clusters DBSCAN sur les données réduites par PCA.

Ces résultats traduisent une **densité non homogène** dans l'espace PCA. La réduction linéaire de PCA capture la variance globale mais perd des relations non linéaires entre variables cliniques. DBSCAN, fondé sur la densité uniforme, ne parvient donc pas à segmenter correctement l'espace, expliquant la présence d'un grand nombre de points isolés (près de 29%).

#### 5.4 Résultats DBSCAN après projection t-SNE

Pour pallier ces limites, la méthode **t-SNE** a été appliquée sur les données PCA afin de mieux préserver les relations locales entre individus. Les paramètres ( $eps = 7.94$ ,  $min\_samples = 42$ ) ont permis d'obtenir des résultats nettement meilleurs.

Cluster_Label	Total_Points	Nb_Malades	Taux_Maladie
-1	22	8	36.36%
0	341	40	11.73%
1	383	308	80.42%

TABLE 7 – Résultats DBSCAN après t-SNE ( $eps=7.94$ ,  $min\_samples=42$ ).

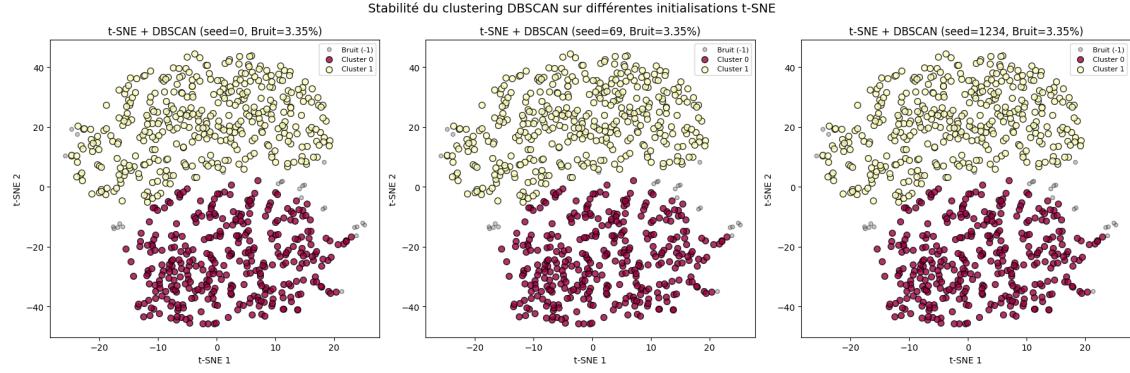


FIGURE 18 – Clusters DBSCAN sur les données projetées par t-SNE.

On observe une nette amélioration : la proportion de bruit chute à 2,55%, et deux clusters bien distincts apparaissent, contenant respectivement 12% et 81% de patients malades. Le t-SNE met en évidence des structures non linéaires non captées par PCA, rendant l'espace de représentation plus adapté au clustering par densité.

Cependant, t-SNE est un algorithme **stochastique** : ses résultats peuvent légèrement varier selon l'initialisation. Des exécutions multiples ont montré une stabilité globale des clusters, confirmant la robustesse du découpage obtenu.

## 5.5 Résultats DBSCAN après réduction Isomap

Enfin, nous avons appliqué **Isomap**, une méthode de réduction de dimension non linéaire basée sur la géodésie des graphes de voisinage. Contrairement à t-SNE, Isomap préserve les **distances globales et locales** en suivant la structure du *manifold* sous-jacent, ce qui en fait une alternative efficace et plus déterministe.

L'application de DBSCAN sur l'espace Isomap a permis d'obtenir une segmentation encore plus fine des patients, avec quatre clusters significatifs.

Cluster_Label	Total_Points	Nb_Malades	Taux_Maladie
-1	73	25	34.25%
0	266	31	11.65%
1	188	153	81.38%
2	152	132	86.84%
3	67	15	22.39%

TABLE 8 – Résultats DBSCAN sur les données réduites par Isomap.

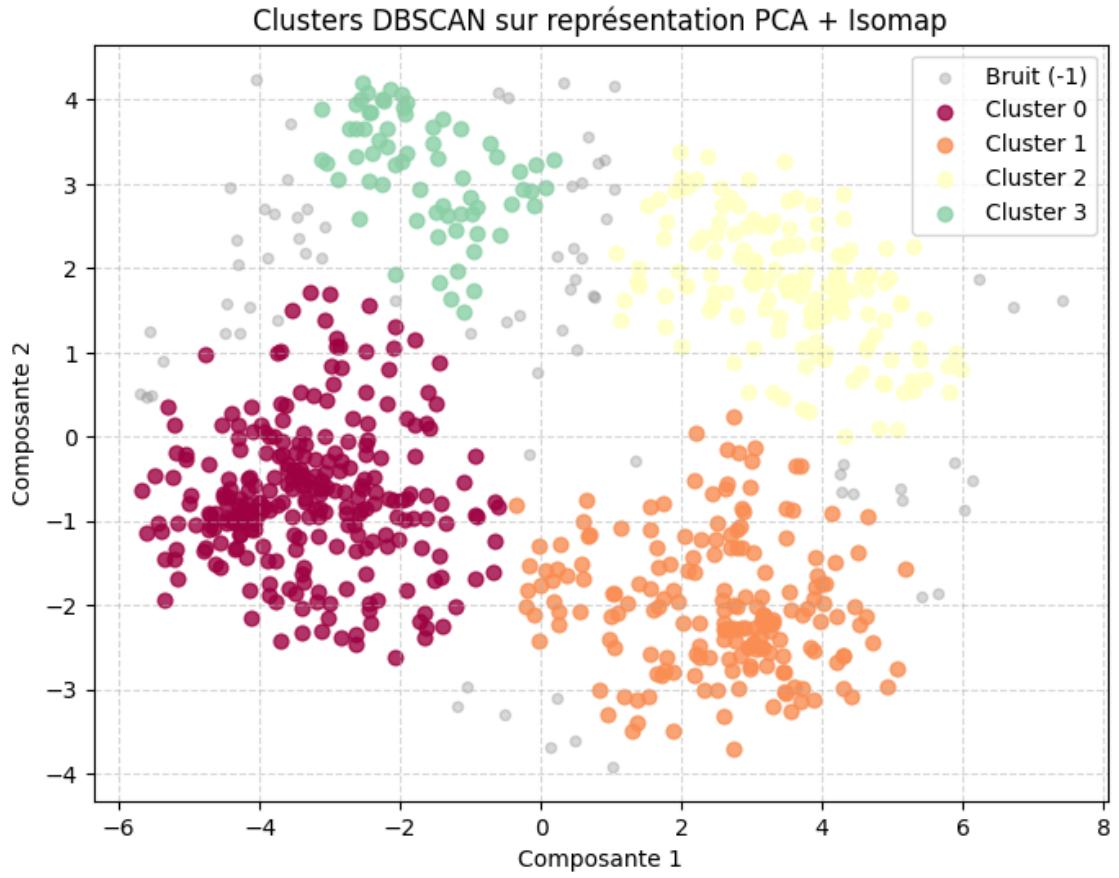


FIGURE 19 – Clusters DBSCAN après réduction Isomap.

Ces résultats montrent que **DBSCAN couplé à Isomap** parvient à isoler **quatre groupes cliniquement cohérents**, dont deux majoritairement composés de patients à risque et deux majoritairement sains. Cette granularité plus fine traduit la capacité d’Isomap à représenter la structure non linéaire globale du jeu de données, tout en conservant la continuité des voisinages.

L’approche Isomap offre ici un compromis intéressant entre la **finesse locale de t-SNE** et la **stabilité globale de PCA**, tout en restant **non stochastique**. Cela explique la qualité et la cohérence des clusters obtenus, particulièrement adaptée au comportement de DBSCAN, qui exploite directement la densité locale.

## 5.6 Synthèse comparative

Méthode	Nb de Clusters	Taux de Bruit	Observations principales
PCA	2	29%	Segmentation grossière, beaucoup de bruit
t-SNE	2	2.5%	Très bonne séparation, mais instabilité possible
Isomap	4	10%	Bonne stabilité, segmentation fine et cohérente

TABLE 9 – Comparaison des résultats DBSCAN selon la méthode de réduction de dimension.

En conclusion, l'association **PCA** + **Isomap** + **DBSCAN** se révèle être le meilleur compromis entre performance, robustesse et interprétabilité, permettant de distinguer efficacement plusieurs profils de patients cardiaques.

## 5.7 Interprétation des clusters Isomap

L'analyse des clusters obtenus avec **Isomap** + **DBSCAN** révèle quatre groupes cliniquement cohérents : deux groupes majoritairement sains et deux groupes majoritairement malades. Les figures 20 et 21 présentent respectivement la répartition des principales variables numériques et catégorielles au sein de ces clusters.

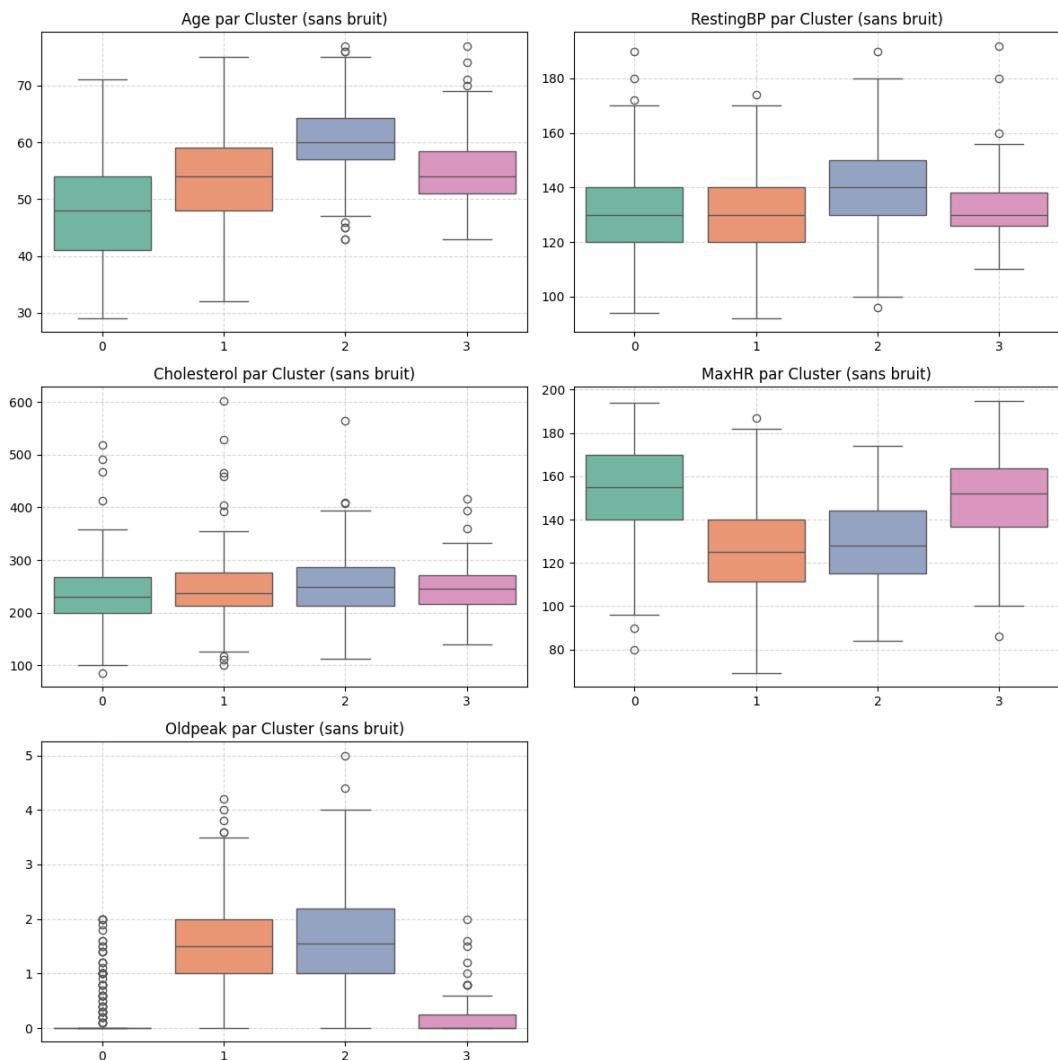


FIGURE 20 – Répartition des variables numériques selon les clusters Isomap.

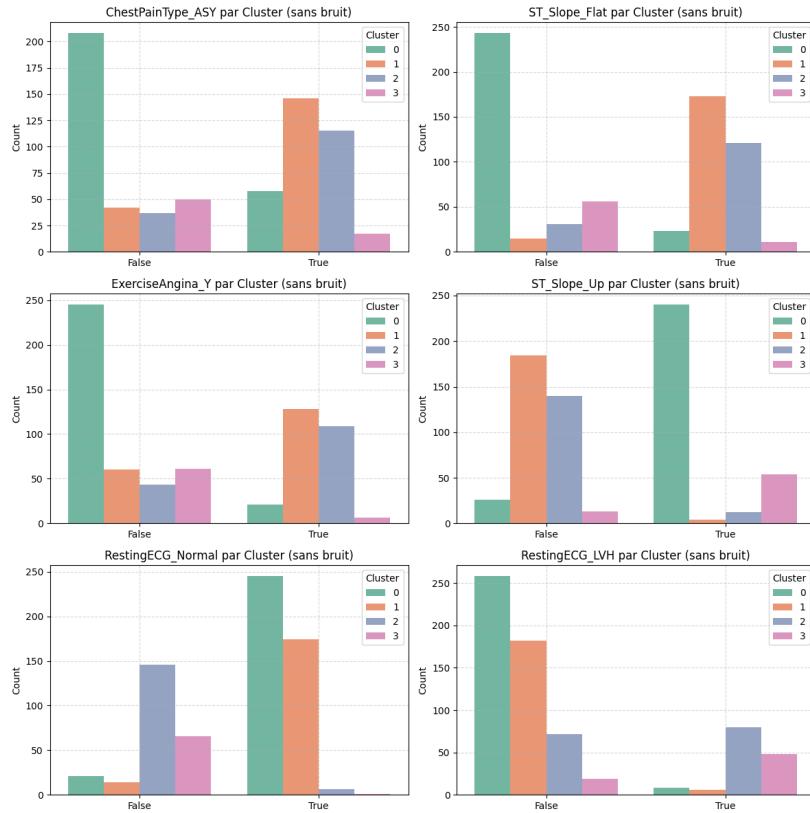


FIGURE 21 – Distribution des principales variables catégorielles discriminantes selon les clusters Isomap.

**Analyse des variables numériques.** Les groupes sains (clusters 0 et 3) présentent un **Oldpeak** nettement plus faible et une **fréquence cardiaque maximale (MaxHR)** plus élevée que les groupes malades (clusters 1 et 2), traduisant une meilleure condition cardiaque. Le cluster 3 regroupe des individus un peu plus âgés et légèrement moins performants que le cluster 0. Parmi les patients malades, le cluster 2 contient les individus les plus âgés et les plus atteints, tandis que le cluster 1 présente des profils similaires mais plus jeunes.

**Analyse des variables catégorielles.** Les patients sains ont majoritairement une **pente ST ascendante (ST\_Slope\_Up)** et peu d'**angine d'effort**, contrairement aux malades qui présentent souvent une pente plate (**ST\_Slope\_Flat**). Le **Resting ECG** est normal pour le cluster 0, anormal (LVH) pour le 3, normal pour le 1 et pathologique pour le 2.

**Synthèse.** L'approche **Isomap + DBSCAN** distingue clairement les patients sains et malades, tout en révélant des sous-groupes cohérents selon l'âge et les caractéristiques ECG.

## 6 Conclusion

Ce projet a permis d'explorer différentes approches de **clustering** appliquées à un jeu de données clinique sur les maladies cardiaques, dans l'objectif d'identifier des profils de patients présentant des risques cardiovasculaires distincts.

Après un travail de préparation des données comprenant le nettoyage, la transformation

des variables et la réduction de dimension par **PCA**, **t-SNE** et **Isomap**, plusieurs algorithmes ont été testés : **K-Means**, **DBSCAN** et **Gaussian Mixture Models (GMM)** :

- **K-Means** a permis une première partition simple et interprétable ;
- **GMM** a introduit une modélisation probabiliste plus souple, capable de capturer des distributions chevauchantes entre patients ;
- **DBSCAN**, combiné à des techniques non linéaires de réduction de dimension, s'est révélé pertinente pour détecter des groupes denses.

Les différentes analyses ont mis en évidence plusieurs variables discriminantes dans la séparation des patients : **Oldpeak** (dépression du segment ST), **MaxHR** (fréquence cardiaque maximale), **ST\_Slope** (pente du segment ST), **ExerciseAngina** et **RestingECG**. Ces indicateurs cliniques se sont avérés décisifs pour distinguer les profils à faible risque et à haut risque.

Ce projet démontre qu'une approche de clustering bien paramétrée peut mettre en évidence des sous-populations cliniquement cohérentes sans supervision explicite. Au-delà des résultats obtenus, il illustre l'intérêt des **méthodes non supervisées en santé publique** pour la détection précoce de profils à risque, l'aide au diagnostic et la personnalisation de la prévention cardiovasculaire.

Des pistes d'amélioration pourraient inclure l'intégration d'autres méthodes de réduction de dimension (UMAP), ou encore l'application d'approches semi-supervisées afin d'affiner la frontière entre patients sains et malades.