



FACULTÉ INFORMATIQUE

MASTER 1 BIOINFORMATIQUE  
RAPPORT PROJET

---

## Fouille de Données : "Méthodes de Clustering"

---

***Etudiants :***

MOUALHI SARAH

181831031232

YAHIAOUI MOHAMED ANIS

171731077246

***Enseignant :***

Mme. BABA ALI

# Table des matières

<b>1</b>	<b>Introduction :</b>	<b>1</b>
<b>2</b>	<b>Pré-processing</b>	<b>2</b>
2.1	Le preprocesing :	3
2.2	Etapes du pre-processing :	3
2.2.1	Data cleaning :	3
2.2.2	Intégration des données :	3
2.2.3	Transformation des données :	4
2.2.4	Réduction des données :	4
2.3	Introduction apprentissage non supervisé :	4
<b>3</b>	<b>Clustering :</b>	<b>5</b>
3.1	Techniques de Clustering :	7
3.1.1	Méthodes de partitionnement :	7
3.1.2	Méthodes hiérarchiques :	8
3.1.3	Méthodes basées sur la densité :	9
<b>4</b>	<b>Implementation :</b>	<b>10</b>
4.1	Bibliothèques utilisées :	10
4.2	Preprocessing :	11
4.3	Méthode d'Elbow :	13
4.4	K-Means :	14
4.5	K-Medoids :	15
4.6	Agnes :	15
4.7	Diana :	17
4.8	DBSCAN :	18
4.9	Comparaison entre les algorithmes :	19
<b>5</b>	<b>Conclusion :</b>	<b>21</b>

# Table des figures

4.1	Bibliothèques utilisées . . . . .	10
4.2	interface application . . . . .	11
4.3	Dataset avant Pre-processing . . . . .	11
4.4	Resultat Pre-processing . . . . .	12
4.5	Choix methode clustering . . . . .	12
4.6	Courbe d'elbow . . . . .	13
4.7	K-Means scatter plot & interclass intra class . . . . .	14
4.8	K-Medoids scatter plot & intra inter classe . . . . .	15
4.9	Agnes dendogram & inter et intraclass . . . . .	16
4.10	Diana dendogram & inter et intraclass . . . . .	17
4.11	DBSCAN implementation . . . . .	18
4.12	Bouton comparer methodes . . . . .	19
4.13	Histogramme de comparaison . . . . .	19

# Chapitre 1

## Introduction :

Le data mining, ou La fouille de données, constitue le cœur d'un processus d'extraction des connaissances à partir d'un large volume de données. Son spectre d'applications s'élargit de plus en plus, mais il est relativement récent dans le domaine de l'éducation.

Il y existe de différentes techniques de data mining, dont la plus connue ; clustering qui occupe une place centrale. Le clustering, ou regroupement, est une méthode d'analyse des données qui vise à diviser un ensemble de données en groupes homogènes ou similaires. Il travaille directement sur des données non étiquetées. En l'absence d'étiquettes pour orienter le processus d'apprentissage, ces étiquettes doivent être "découvertes". Dans le cadre de ce projet de module sur le clustering, notre objectif est d'implémenter et de comparer différents algorithmes de clustering. Nous nous concentrerons sur les algorithmes de clustering les plus répandus et influents, tels que le K-means, le K-medoids, l'Agnes (Agrégation Hiérarchique Ascendante) et le Diana (Agrégation Hiérarchique Descendante) : K-Means, K-Medoids, Agnes, Diana et DBSCAN et les testons sur différents datasets comme breast cancer, diabetes etc. Ces algorithmes sont des outils essentiels pour explorer et organiser des données non étiquetées, permettant ainsi de révéler des structures et de prendre des décisions éclairées. Cette démarche nous permettra de mettre en lumière les particularités de chaque algorithme et de comprendre comment ils peuvent être appliqués dans divers contextes.

Dans un premier lieu, nous implémentons les algorithmes de clustering en développant les solutions en utilisant le langage python. Nous allons mettre en œuvre et visualiser différents algorithmes de clustering tels que le K-means, le K-medoids, l'Agnes et le Diana en utilisant des scatter plots. Ces visualisations nous permettront de mieux comprendre la manière dont chaque algorithme partitionne les données en groupes distincts. De plus, nous allons calculer les mesures d'interclasse et d'intraclasse pour évaluer la qualité des clusters formés par chaque algorithme. Quant à l'interface de test sera implémenté à l'aide du framework Streamlit qui permet la création d'applications web pour la data science, où l'utilisateur aura la main de choisir le dataset et la méthode de clustering souhaitées.

Enfin, nous allons comparer les performances des différents algorithmes. Nous allons examiner des critères tels que la compacité des clusters, la séparation entre les clusters et la stabilité des résultats. En analysant ces aspects, nous pourrions déterminer les forces et les faiblesses de chaque algorithme et identifier les cas d'utilisation où un algorithme particulier se distingue.

# Chapitre 2

## Pré-processing

### 2.1 Le preprocesing :

Le prétraitement des données a une importance cruciale dans tout projet d'apprentissage automatique. En effet, lors de l'acquisition des données, des erreurs humaines ou techniques peuvent survenir, ce qui peut corrompre notre jeu de données et introduire des biais lors de l'entraînement. Parmi ces erreurs, on trouve des informations incomplètes, des valeurs manquantes ou incorrectes, ainsi que des bruits parasites dus à l'acquisition des données. Par conséquent, il est souvent indispensable de mettre en place une stratégie de prétraitement des données, également appelée prétraitement des données qui permet de garantir la qualité et la pertinence des données utilisées en transformant des données brutes dans un format compréhensible ,efficace et exploitable pour le traitement .Les données originales sont souvent incomplètes, incohérentes ou dépourvues de certains comportements ou tendances, et elles peuvent contenir de nombreuses erreurs . Le prétraitement des données est une méthode éprouvée visant à résoudre ces problèmes et à préparer ces données brutes pour un traitement ultérieur et pour une exploitation approfondie afin de construire un modèle plus performant.

### 2.2 Etapes du pre-processing :

Les données passent par une série d'étapes pendant le prétraitement :

#### 2.2.1 Data cleaning :

Le nettoyage des données est l'étape la plus importante du prétraitement, car il garantit que vos données sont prêtes à être utilisées pour vos besoins ultérieurs ,les données peuvent comporter de nombreuses parties non pertinentes et manquantes. Pour gérer cette partie, un nettoyage des données est effectué. Cela implique la gestion des données manquantes, des données bruitées, etc , les données sont nettoyées par des processus tels que le remplissage des valeurs manquantes, le lissage des données bruyantes ou la résolution des incohérences dans les données en les remplaçant par la valeur moyenne, par la médiane, ou par exemple par la modalité la plus fréquente dans le cas de variables catégorielles .(autrement appelé le mode) .

#### 2.2.2 Intégration des données :

L'intégration des données est le processus de fusion de données provenant de plusieurs sources ou systèmes en un seul ensemble. Cela implique l'intégration des données et la résolution des conflits de valeurs des données dispersées dans différents systèmes de représentation ou de mesure, chacun ayant sa propre structure, son propre format ou sa propre sémantique. L'intégration des données permet d'acquérir une compréhension globale des informations, ce qui facilite l'analyse, l'établissement de rapports et la prise de décision.

### 2.2.3 Transformation des données :

Le processus de transformation ou de modification des données de leur format ou structure d'origine en un format adapté à l'analyse, à la modélisation ou à d'autres opérations ultérieures est appelé transformation des données. Pour créer de nouvelles caractéristiques, modifier l'échelle ou la distribution, ou réorganiser la structure des données. Dans cette étape les données sont normalisées, agrégées et généralisées.

### 2.2.4 Réduction des données :

La réduction des données est un processus dans lequel la taille ou la dimensionnalité d'un ensemble de données est réduite , cette étape consiste à obtenir une représentation réduite des données en termes de volume tout en préservant l'intégrité des données d'origine ; ses caractéristiques importantes et en minimisant la perte d'informations.

## 2.3 Introduction apprentissage non supervisé :

L'apprentissage automatique est une sous-catégorie de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistique pour donner aux ordinateurs la capacité d'apprendre à partir des données, autrement dit améliorer leurs performances à résoudre des tâches sans être explicitement programmés.

On trouve différents types de machine learning dont l'apprentissage supervisé qui est une approche qui consiste à faire apprendre une machine à travers des exemples d'entrées qui sont labellisés avec la sortie souhaitée afin que l'algorithme puisse prédire les valeurs des sorties en fonction des entrées .

Dans l'autre côté on trouve l'apprentissage non supervisé et on parle d'apprentissage non supervisé si les données ne sont pas étiquetées, on dispose donc de données d'entrée dont on ne connaît pas la sortie associée. L'algorithme doit apprendre tout seul à trouver le point commun entre les données d'entrée et les regrouper dans des clusters ou des classes à travers des algorithmes de regroupement (clustering). Chaque élément du groupe doit avoir des caractéristiques proches de celles des éléments du même groupe mais des caractéristiques relativement éloignées de celles des autres groupes

## Chapitre 3

### Clustering :



La classification est une tâche couramment utilisée dans la vie quotidienne. Elle permet d'expliquer de nouveaux phénomènes en les comparant à des concepts et phénomènes connus, en cherchant à identifier les caractéristiques les plus distinctives.

C'est un domaine de recherche actif qui joue un rôle central dans l'analyse de grandes quantités de données, ce qui justifie l'intérêt qui lui est accordé. Il se situe à l'intersection de plusieurs disciplines, telles que la statistique, l'apprentissage automatique, le data mining, etc.

Le principal défi auquel les méthodes de classification tentent de répondre est : comment associer une classe à un objet ?

Les techniques de classification sont essentiellement réparties en deux approches : la classification supervisée et la classification non supervisée, également appelée clustering.

Dans l'approche supervisée, on cherche à construire un modèle de classification à partir d'un ensemble de données étiquetées, c'est-à-dire que la classe de chaque objet est connue à l'avance. On parle alors d'apprentissage supervisé, et cet ensemble de données est souvent appelé ensemble d'apprentissage. Cela suppose une connaissance préalable des classes, notamment sur la base de laquelle deux objets sont regroupés ou séparés. Le modèle obtenu doit être capable de prédire la classe la plus probable pour de nouvelles entrées (ou objets), afin de préserver la cohérence avec la structure initiale des classes. Cette cohérence est évaluée à l'aide d'une mesure de qualité des classes. Cependant, cette cohérence peut être remise en question si le nombre de nouvelles entrées à classer dépasse largement celui de l'ensemble d'apprentissage, en raison des nouvelles informations apportées. C'est pourquoi il est nécessaire de mettre à jour la classification initiale.

En revanche, dans une approche non supervisée, l'ensemble d'apprentissage n'est pas étiqueté. Le problème de classification devient donc plus complexe car aucune information préalable sur les classes n'est disponible. L'objectif est de détecter des objets similaires en fonction des variables (attributs) qui les décrivent, afin de les regrouper tout en ignorant certains détails de similarité ou de dissimilarité. Ce regroupement des objets repose sur une mesure appelée mesure de similarité ou distance. La qualité du résultat obtenu est évaluée en fonction du degré d'homogénéité intra-classes et d'hétérogénéité inter-classes.

L'approche non supervisée est également connue sous le nom de clustering, segmentation et parfois regroupement. Les groupes obtenus sont appelés classes ou clusters.

La modélisation des données par le clustering puise ses fondements dans les mathématiques, les statistiques et l'analyse numérique. Le data mining ajoute à cette méthode de modélisation la complexité des ensembles de données volumineux avec de nombreux attributs ou variables de types différents. Cela nécessite une puissance de calcul plus élevée et des algorithmes plus performants.

Le clustering est souvent la première tâche à effectuer pour construire des groupes sur lesquels on applique ensuite des tâches de classification ou d'estimation.

### 3.1 Techniques de Clustering :

Le regroupement permet d'effectuer des analyses de surface des données non structurées. La distance la plus courte, la représentation graphique et la densité des points de données sont quelques-uns des facteurs qui influencent la formation des clusters. En déterminant le degré de similitude entre les éléments à l'aide d'une métrique appelée mesure de similitude, les objets sont regroupés en clusters.

Pour cela de nombreuses méthodes de regroupement ont été développées, chacune d'entre elles utilisant un principe d'induction différent. la méthode utilisée pour définir les classes, les algorithmes peuvent être classifiés de façon générale en trois catégories principales :

#### 3.1.1 Méthodes de partitionnement :

Son principe général est de démarrer à partir d'un seul cluster qui est partitionné d'une manière itérative en effectuant une redistribution des objets ou en essayant d'identifier les clusters comme étant des régions très peuplées jusqu'à la rencontre d'un critère d'arrêt. L'objectif de ces méthodes est de diviser de manière optimale l'ensemble des objets en un nombre fixe de groupes. Les clusters identifiés ont généralement une forme sphérique.

Les algorithmes de partitionnement les plus utilisés sont K-means et K-medoids .

#### K-Means :

Le regroupement K-Means est un type d'apprentissage non supervisé, utilisé lorsque vous disposez de données non étiquetées (c'est-à-dire des données sans catégories ou groupes définis). L'objectif de cet algorithme est de trouver des groupes dans les données, le nombre de groupes étant représenté par la variable K. L'algorithme travaille de manière itérative pour affecter chaque point de données à l'un des K groupes sur la base des caractéristiques fournies. Les points de données sont regroupés en fonction de la similarité des caractéristiques. Les résultats de l'algorithme de regroupement des K-moyennes sont les suivants : les centroïdes des K groupes, qui peuvent être utilisés pour étiqueter de nouvelles données. Les étiquettes pour les données d'apprentissage (chaque point de données est assigné à un seul groupe).k-means vise donc à minimiser la variance intra classe .La qualité de la solution trouvée dépend fortement du choix de la valeur de k de départ.

#### K-Medoids :

K-Medoids est un algorithme de clustering qui partitionne un ensemble de données donné en k clusters, où chaque observation appartient au cluster avec le médoid le plus proche. Le médoid peut être défini comme l'observation qui minimise la dissimilarité moyenne entre elle-même et toutes les autres observations dans le même cluster. L'algorithme K-Medoids est une variation de l'algorithme K-Means qui utilise des médoids au lieu de centroïdes pour définir les clusters. K-Medoids peut être utilisé de différentes manières pour le clustering des données et peut être utilisé dans diverses applications pour

obtenir des informations et prendre des décisions éclairées.

- **L'intraclass** : est une mesure de la similarité entre les différentes observations qui appartiennent au même cluster. Plus précisément, l'intraclass est défini comme la distance moyenne entre chaque point et le médoid correspondant dans le même cluster. Le médoid est l'observation qui minimise la dissimilarité moyenne entre elle-même et toutes les autres observations dans le même cluster. Cette mesure permet de quantifier la cohésion du cluster, en évaluant à quel point les observations du même cluster sont similaires les unes aux autres.
- **L'interclass** : mesure la distance moyenne entre chaque point et le médoid correspondant dans tous les autres clusters. Cette mesure permet de quantifier la séparation entre les différents clusters, en évaluant à quel point les observations d'un cluster sont différentes de celles des autres clusters.

### 3.1.2 Méthodes hiérarchiques :

Ces méthodes construisent les clusters en partitionnant récursivement les instances de manière descendante ou ascendante jusqu'à la satisfaction d'un critère d'arrêt. à la différence des méthodes de partitionnement présentées précédemment, les méthodes hiérarchiques produisent une séquence de partitions imbriquées appelée dendrogramme. Ces méthodes hiérarchiques peuvent être subdivisées comme suit :

- **Regroupement hiérarchique agglomératif** : Chaque objet représente initialement un cluster qui lui est propre. Les clusters sont ensuite fusionnés successivement jusqu'à l'obtention de la structure de cluster souhaitée.
- **Regroupement hiérarchique divisé** : Tous les objets appartiennent initialement à un seul cluster. Ensuite, le cluster est divisé en sous-clusters, qui sont successivement divisés en leurs propres sous-clusters. Ce processus se poursuit jusqu'à ce que l'on obtienne la structure de cluster souhaitée.

Le résultat des méthodes hiérarchiques est un dendrogramme, qui représente le regroupement imbriqué des objets et les niveaux de similarité auxquels les regroupements changent. Un regroupement (clustering) des objets de données est obtenu en coupant le dendrogramme au niveau de similarité souhaité.

La fusion ou la division des clusters est effectuée en fonction d'une mesure de similarité, choisie de manière à optimiser un critère (tel que la somme des carrés). Les méthodes de regroupement hiérarchique peuvent être subdivisées en fonction de la manière dont la mesure de similarité est calculée.

#### Points faibles et forts de Agnes et Diana :

- **Poins forts** :  
L'un des principaux avantages de l'algorithme AGNES est qu'il est plus simple à

comprendre et à mettre en œuvre que l'algorithme DIANA. De plus, il est plus robuste aux valeurs aberrantes que l'algorithme DIANA.

Un autre avantage de l'algorithme AGNES est qu'il peut être utilisé pour des ensembles de données de grande taille, y compris des données multidimensionnelles. Quant à DIANA, il est plus rapide que l'algorithme AGNES pour les grands ensembles de données. De plus, il est adapté aux données non numériques ou catégorielles.

Un autre avantage de l'algorithme DIANA est qu'il est plus robuste aux choix de la mesure de similarité et de la méthode d'agrégation que l'algorithme AGNES.

- **Points faibles :**

l'algorithme AGNES présente des inconvénients tels que sa lenteur pour les grands ensembles de données et son inadaptation aux données non numériques ou catégorielles. Il est également sensible aux choix de la mesure de similarité et de la méthode d'agrégation.

Quant à Diana, L'un des principaux inconvénients de l'algorithme DIANA est sa complexité. Il peut être très lent pour les ensembles de données de grande taille. De plus, il ne fonctionne pas bien avec des données multidimensionnelles.

Un autre inconvénient de l'algorithme DIANA est qu'il peut être sensible aux valeurs aberrantes.

### 3.1.3 Méthodes basées sur la densité :

Ce type de clustering se base sur l'utilisation de la densité à la place de la distance. On dit qu'un point est dense si le nombre de ses voisins dépasse un certain seuil. Un point est voisin d'un autre point s'il est à une distance inférieure à une valeur fixée. L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un exemple des algorithmes à base de densité.

#### DBSCAN :

C'est un algorithme de clustering non paramétrique utilisé pour regrouper des points de données dans des groupes basés sur leur densité. L'objectif principal de DBSCAN est de trouver des zones denses de points dans des données à  $n$  dimensions, séparées par des zones moins denses. Il utilise deux paramètres : la distance epsilon et le nombre minimum de points MinPts devant se trouver dans un rayon epsilon pour que ces points soient considérés comme un cluster. Les 91 paramètres d'entrées sont donc une estimation de la densité de points des clusters. L'idée de base de l'algorithme est ensuite, pour un point donné, de récupérer son epsilon -voisinage et de vérifier qu'il contient bien MinPts points ou plus. Ce point est alors considéré comme faisant partie d'un cluster. On parcourt ensuite l'epsilon-voisinage de proche en proche afin de trouver l'ensemble des points du cluster.

# Chapitre 4

## Implementation :

### 4.1 Bibliothèques utilisées :

```
app.py > ...
1  #pip install kneed
2  import streamlit as st
3  from PIL import Image
4  import pandas as pd
5  import seaborn as sns
6  import numpy as np
7  import matplotlib.pyplot as plt
8  from sklearn.preprocessing import LabelEncoder
9  from sklearn.preprocessing import StandardScaler
10 from sklearn.cluster import KMeans
11 #pip install scikit-learn-extra
12 from sklearn_extra.cluster import KMedoids
13 from kneed import KneeLocator
14 from sklearn.metrics import pairwise_distances , silhouette_score
15 from sklearn.cluster import AgglomerativeClustering
16 import scipy.cluster.hierarchy as sch
17 from scipy.spatial.distance import pdist
18 from scipy.spatial.distance import squareform
19 from scipy.cluster.hierarchy import linkage, dendrogram
20 from sklearn.decomposition import PCA
21 from sklearn.cluster import DBSCAN
22
```

FIG. 4.1 : Bibliothèques utilisées

### 4.2 Preprocessing :

Selectionner ou uploader une dataset

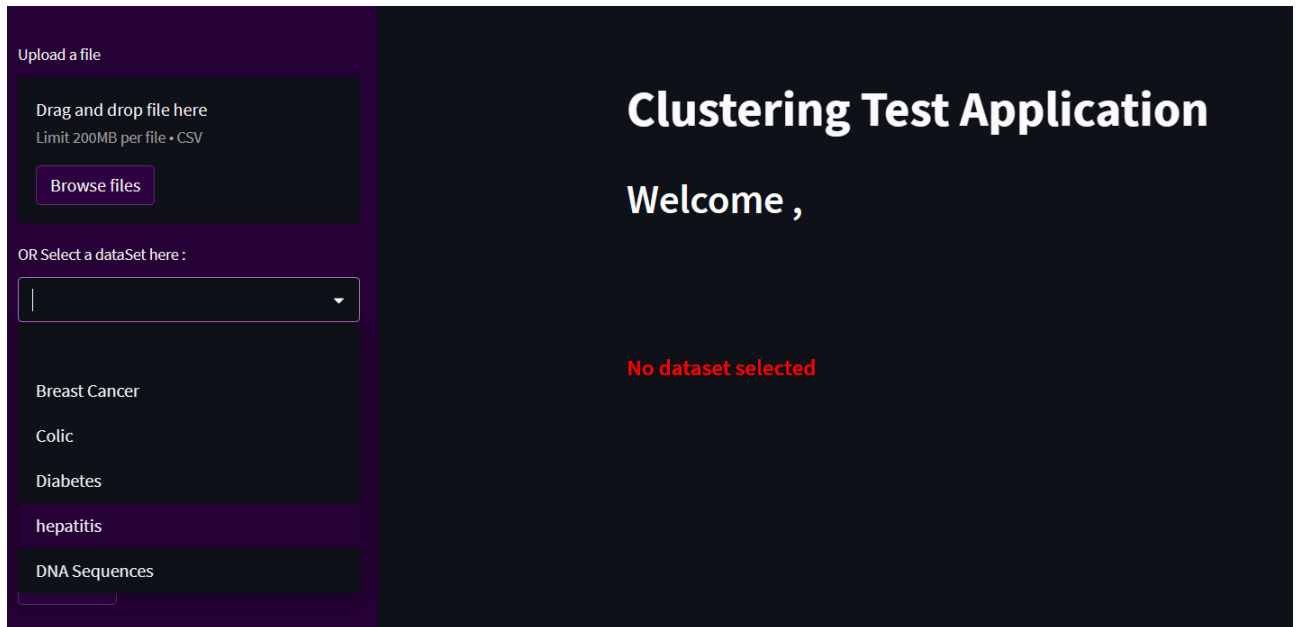


FIG. 4.2 : interface application

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	
1	1	85	66	29	0	26.6	0.351	
2	8	183	64	0	0	23.3	0.672	
3	1	89	66	23	94	28.1	0.167	
4	0	137	40	35	168	43.1	2.288	
5	5	116	74	0	0	25.6	0.201	
6	3	78	50	32	88	31	0.248	
7	10	115	0	0	0	35.3	0.134	
8	2	197	70	45	543	30.5	0.158	
9	8	125	96	0	0	0	0.232	

FIG. 4.3 : Dataset avant Pre-processing

Pre-Processing phase :

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
0	0.6399	0.8483	0.1496	0.9073	-0.6929	0.204	0.4685
1	-0.8449	-1.1234	-0.1605	0.5309	-0.6929	-0.6844	-0.3651
2	1.2339	1.9437	-0.2639	-1.2882	-0.6929	-1.1033	0.6044
3	-0.8449	-0.9982	-0.1605	0.1545	0.1233	-0.494	-0.9208
4	-1.1419	0.5041	-1.5047	0.9073	0.7658	1.4097	5.4849
5	0.343	-0.1532	0.253	-1.2882	-0.6929	-0.8113	-0.8181
6	-0.251	-1.3425	-0.9877	0.7191	0.0712	-0.126	-0.6761
7	1.8278	-0.1845	-3.5726	-1.2882	-0.6929	0.4198	-1.0204
8	-0.5479	2.3819	0.0462	1.5346	4.0219	-0.1894	-0.9479
9	1.2339	0.1285	1.3904	-1.2882	-0.6929	-4.0605	-0.7245

FIG. 4.4 : Resultat Pre-processing

Upload a file

Drag and drop file here  
Limit 200MB per file • CSV

Browse files

OR Select a dataSet here :

Diabetes ▼

Choose a Clustering method

☒ None

☐ K-Means

☐ K-Medoids

☐ Agnes

☐ Diana

☐ DBScan

FIG. 4.5 : Choix methode clustering

### 4.3 Méthode d'Elbow :

L'un des problèmes les plus complexes auxquels nous sommes confrontés lorsque nous essayons de segmenter des clients ou des produits est le choix du nombre idéal de segments. Il s'agit d'un paramètre clé pour de nombreux algorithmes de regroupement tels que K means, K medoids et le regroupement agglomératif.

Il s'agit de la méthode la plus répandue pour déterminer le nombre optimal de clusters. La méthode est basée sur le calcul de la somme des carrés des erreurs à l'intérieur des clusters (WSS) pour différents nombres de clusters (k) et sur la sélection du k pour lequel le changement dans la WSS commence à diminuer.

L'idée derrière la méthode elbow est que la variation expliquée change rapidement pour un petit nombre de grappes et qu'elle ralentit ensuite, ce qui conduit à la formation d'un coude dans la courbe. Le point de coude est le nombre de clusters que nous pouvons utiliser pour notre algorithme de regroupement.

On appliquant la methode d'elbow sur K-Means on obtient le nombre optimal de cluster qui est 4 clusters dans notre cas .

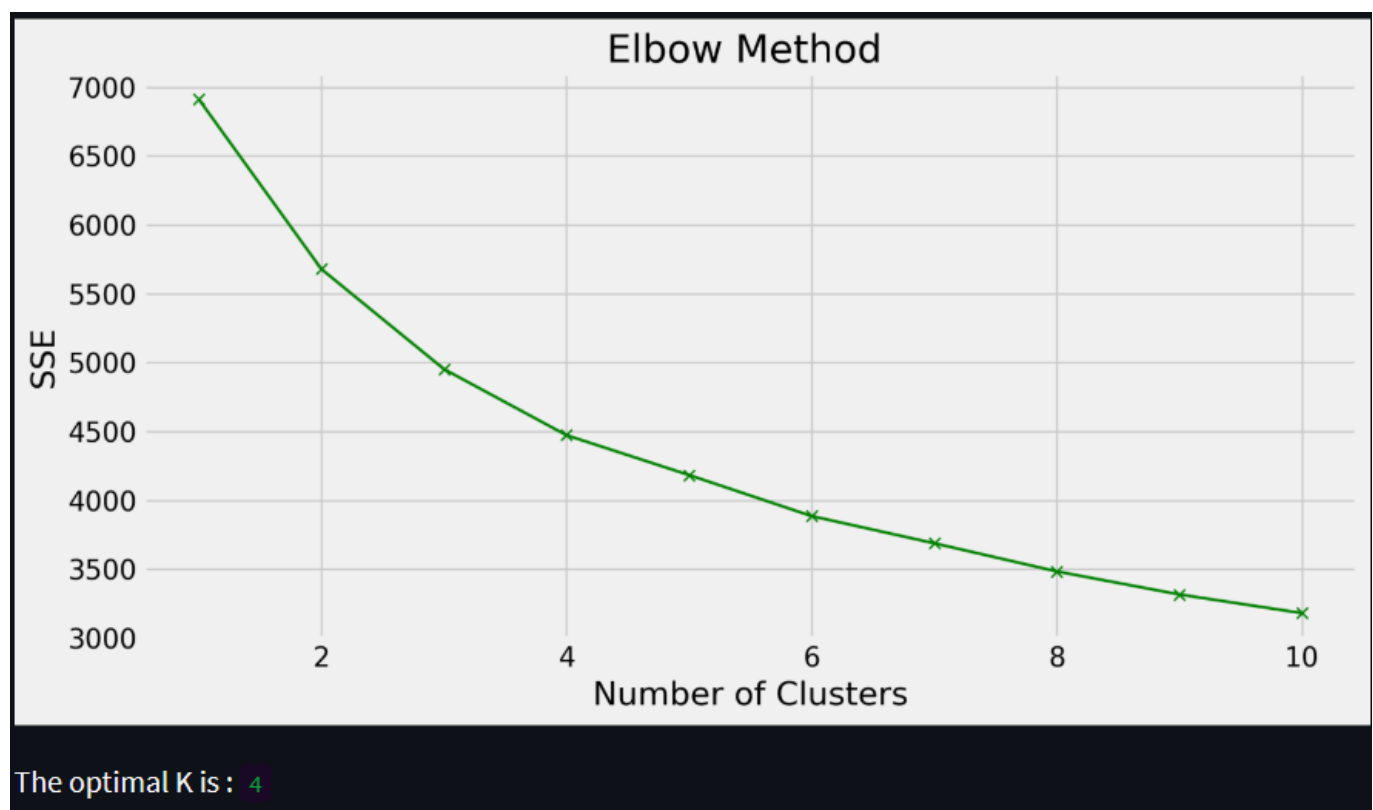


FIG. 4.6 : Courbe d'elbow



## 4.4 K-Means :

Voici les résultats obtenus pour la méthode K-means lors de son implémentation en visualisant un graphique de dispersion (scatter plot) ainsi les calculs les mesures d'inter-classe et d'intraclasse kmeans en utilisant nombre de clusters k optimal obtenu avec la methode d'elbow :

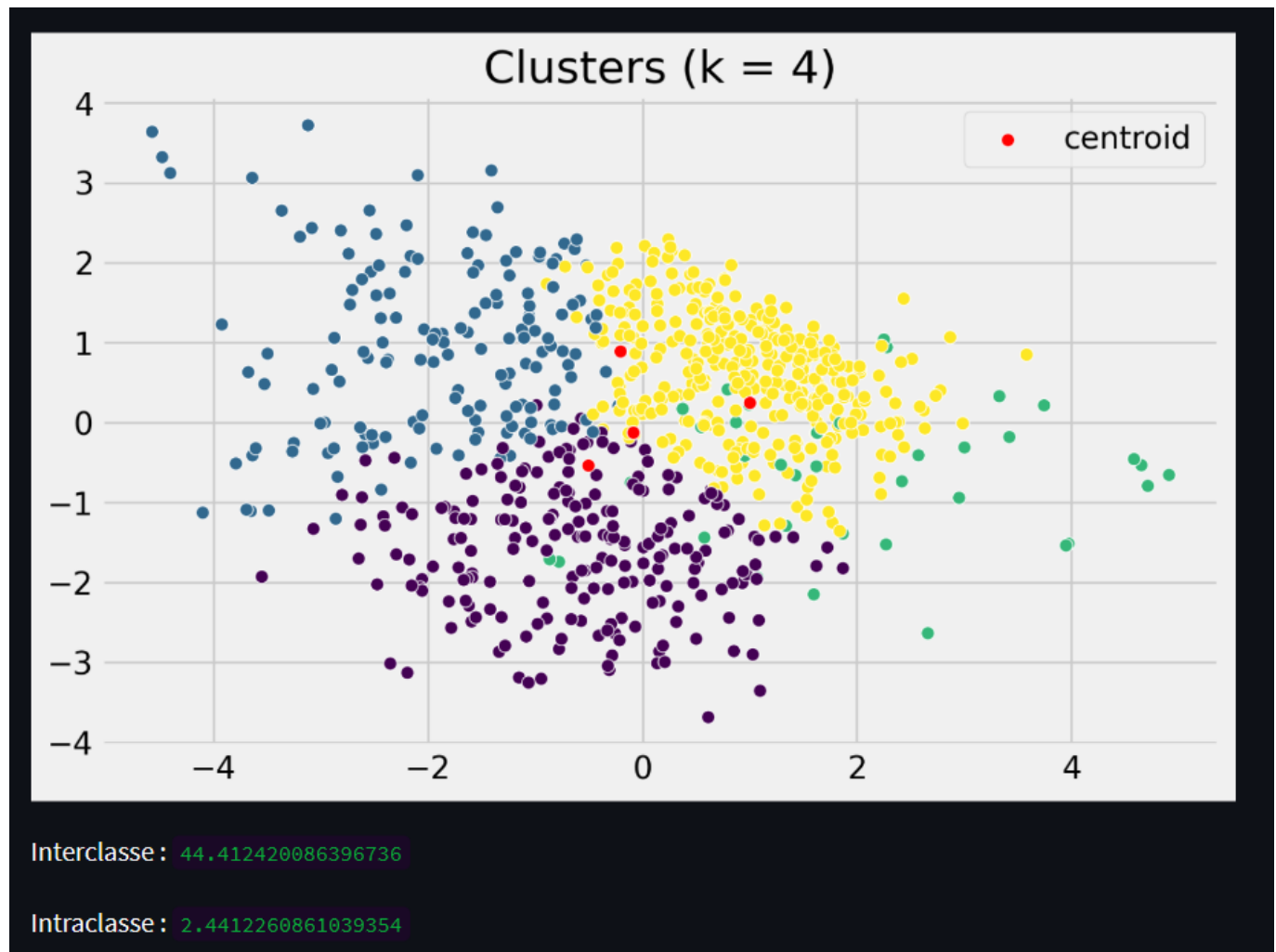


FIG. 4.7 : K-Means scatter plot & interclass intra class

**Interclasse : 44.4124280**

**Intraclasse : 2.44122668**

## 4.5 K-Medoids :

Voici les résultats obtenus pour la méthode K-Medoids en visualisant un graphique de dispersion (scatter plot) et donnant les calculs les mesures d'interclasse et d'intraclasse en utilisant nombre de clusters k optimal obtenu avec la methode d'elbow :

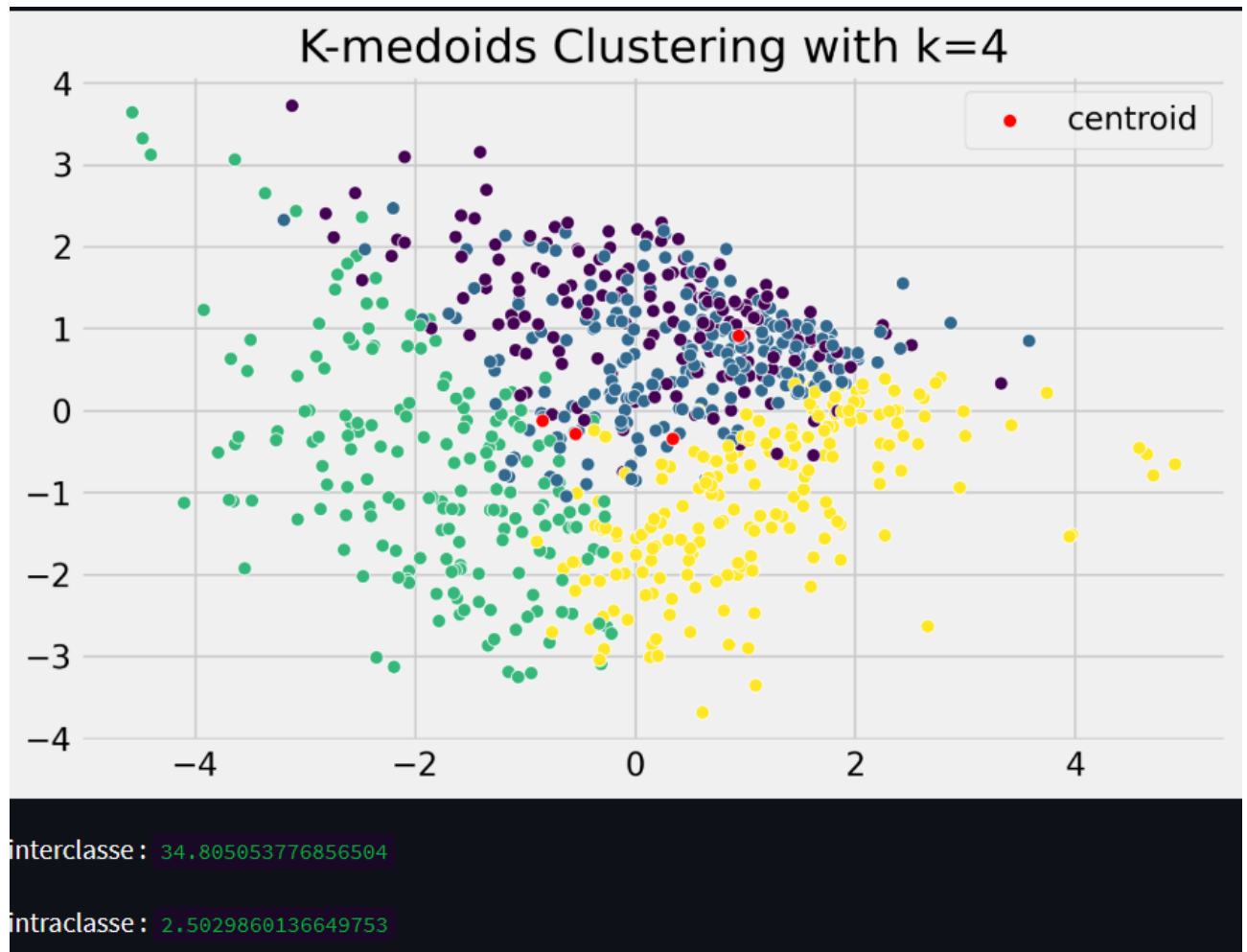


FIG. 4.8 : K-Medoids scatter plot & intra inter classe

**Interclasse : 34.88585377**

**Intraclasse : 2.502986013**

## 4.6 Agnes :

Voici les résultats obtenus pour la méthode Agnes en affichant son dendogram et donnant les calculs les mesures d'interclasse et d'intraclasse en utilisant meme nombre de clusters k :

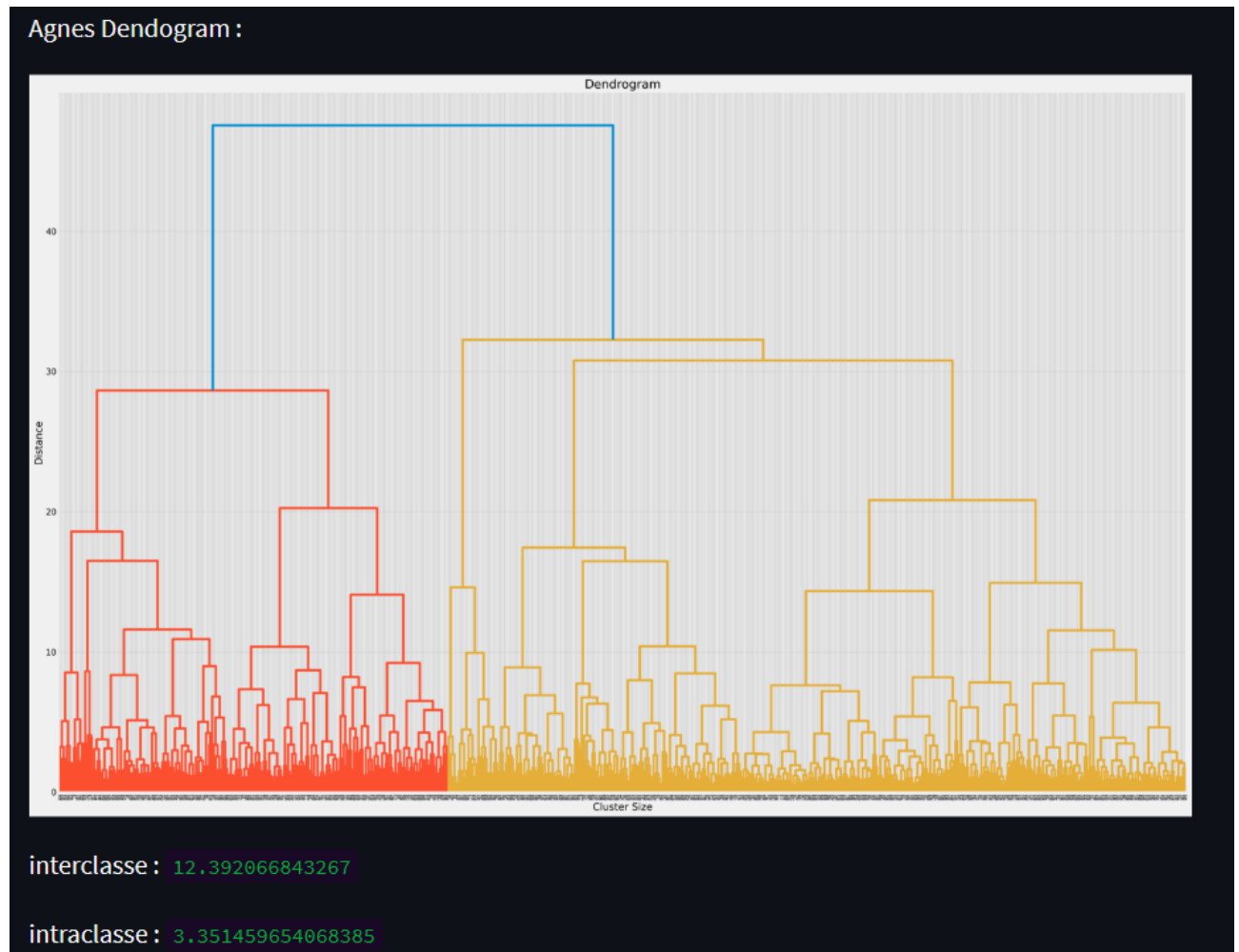


FIG. 4.9 : Agnes dendrogram & inter et intraclasse

**Interclasse : 12.392066**

**Intraclasse : 3.351459**

### 4.7 Diana :

Voici les résultats obtenus pour la méthode Diana en affichant son dendrogram et donnant les calculs les mesures d'interclasse et d'intraclasse en utilisant meme nombre de clusters  $k$  :



FIG. 4.10 : Diana dendrogram & inter et intraclasse

**Interclasse : 11.39484372**

**Intraclasse : 4.5984532**

### 4.8 DBSCAN :

Cette visualisation vous permettra d'observer comment le nombre de clusters varie en fonction des valeurs de "minPts" et "epsilon". Sur l'axe des abscisses, la valeur de "epsilon" et sur l'axe des ordonnées, on représente la valeur des "minPts".

Pour chaque exécution, on peut enregistrer le nombre de clusters trouvés par DBSCAN. Le graphisme suivant montre ces résultats en utilisant des couleurs ou des niveaux de bleus pour indiquer le nombre de clusters. L'échelle de couleurs où les couleurs plus claires représentent un plus petit nombre de clusters et les couleurs plus foncées représentent un plus grand nombre de clusters.

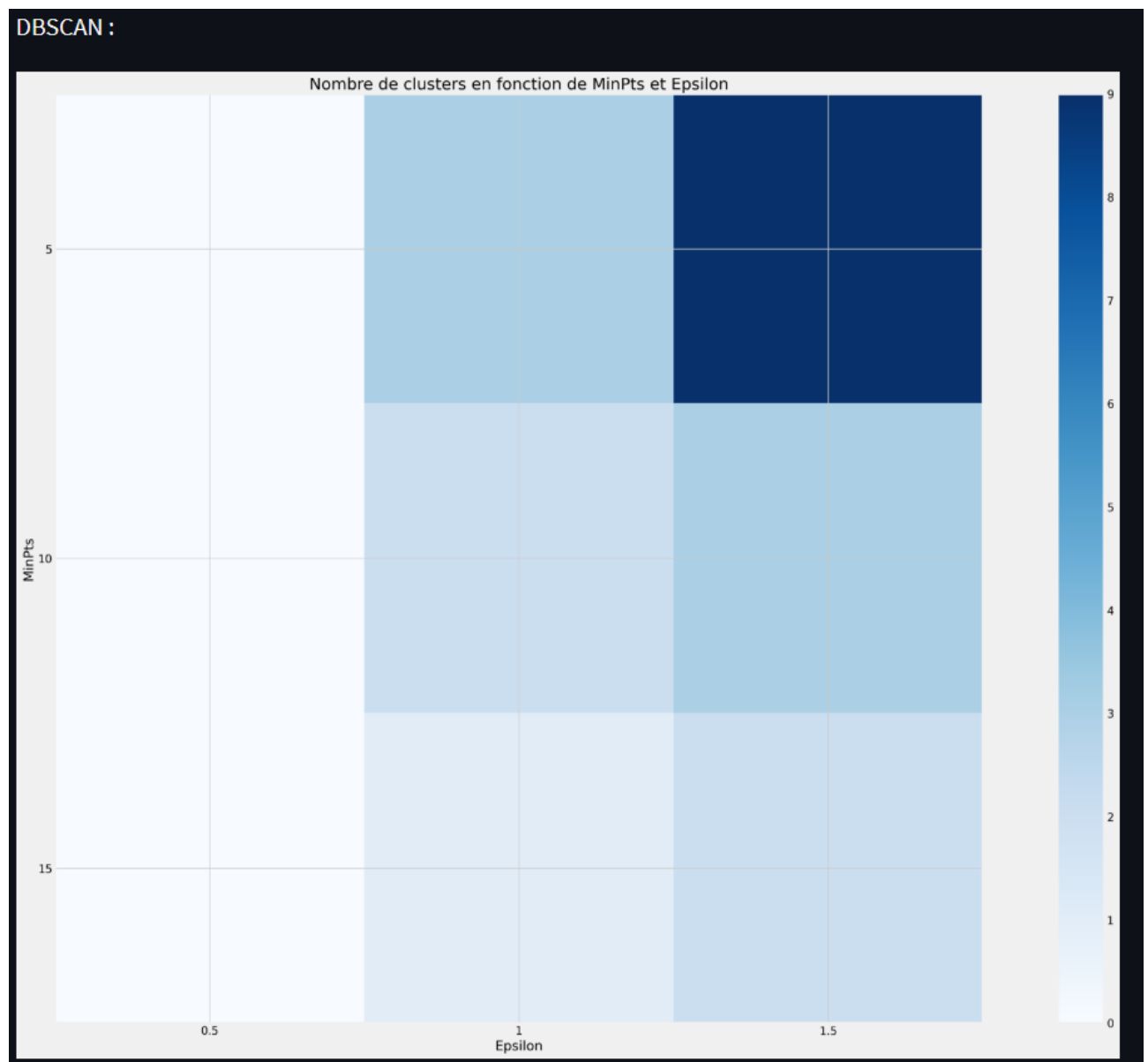
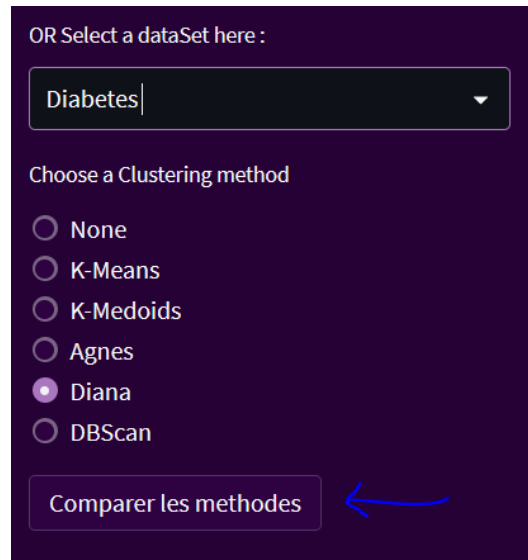


FIG. 4.11 : DBSCAN implementation

## 4.9 Comparaison entre les algorithmes :

On aura sur l'interface un bouton dont il affiche l'histogramme de comparaison entre les méthodes de clustering .



OR Select a dataSet here :

Diabetes

Choose a Clustering method

☐ None

☐ K-Means

☐ K-Medoids

☐ Agnes

☒ Diana

☐ DBScan

Comparer les methodes

A blue arrow points to the 'Comparer les methodes' button.

FIG. 4.12 : Bouton comparer methodes

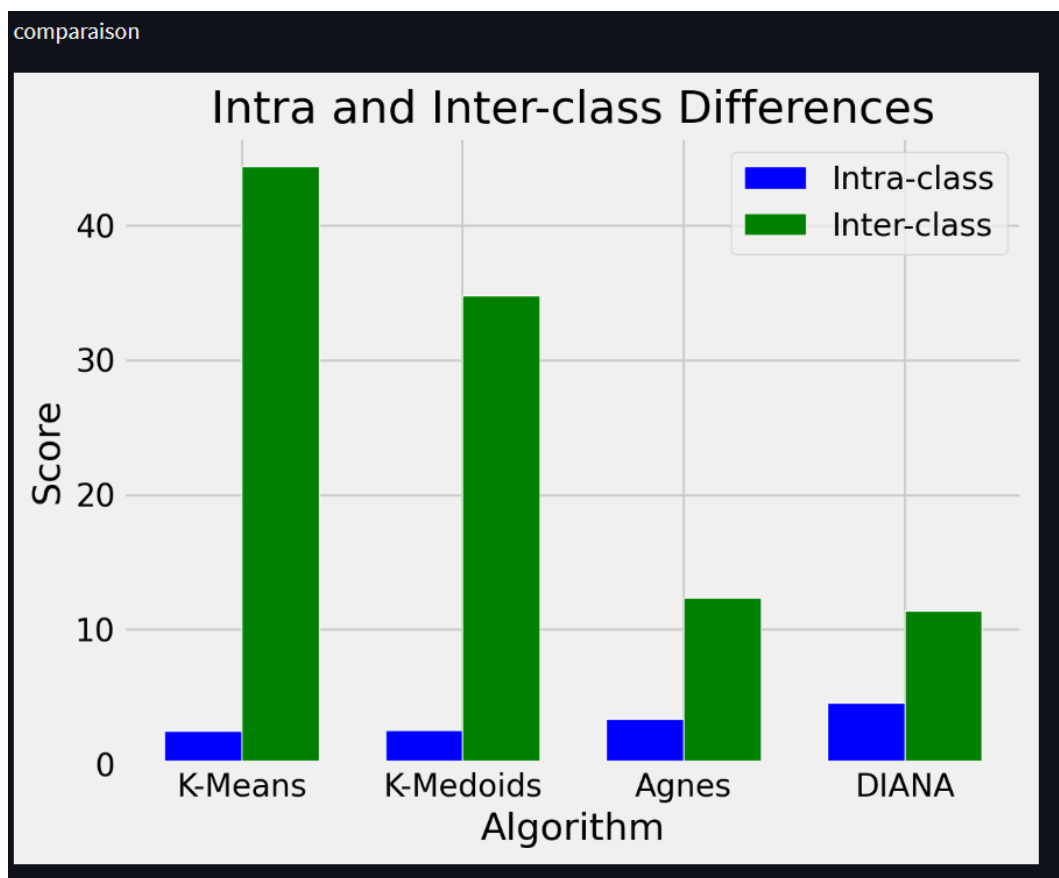


FIG. 4.13 : Histogramme de comparaison

L'histogramme des comapraison de K-means, K-medoids, Agnes et Diana montre L'analyse de chacun de ses algorithmes qui a été porté sur les histogrammes des mesures Intra et Inter pour chaque méthode.

D'après les résultats on constate que K-means et K-medoids ont démontré leur pertinence , les valeurs d'interclasse et intraclasse de chacun des deux indiquent une séparation relativement élevée entre les clusters et une compacité des données au sein des clusters. K-means a réussi à former des clusters bien distincts avec une faible dispersion interne. En revanche , la méthode K-medoids présente une valeur d'interclasse légèrement inférieure au environ de 34, tandis que la valeur d'intraclasse est d'environ 2.5. Donc on constate une certaine similarité entre les clusters et une dispersion interne légèrement plus élevée que celle observée avec K-means.

En ce qui concerne les méthodes Agnes et Diana, ils ont montré une certaine résistance, nous remarquons des valeurs d'interclasse plus faibles, avec 12 pour Agnes et 11 pour Diana, et des valeurs d'intraclasse relativement plus élevées, avec 3.35 pour Agnes et 4.44 pour Diana. Ces résultats suggèrent que les deux méthodes ont rencontré des difficultés dans la formation de clusters bien séparés et compacts. La dispersion interne des données au sein des clusters est plus importante, ce qui peut indiquer une plus grande variabilité des données au sein de chaque groupe.

En ce qui concerne DBSCAN , IL diffère des autres méthodes de clustering car il identifie les clusters en fonction de la densité des points. Cela lui permet de détecter des clusters de forme arbitraire et de gérer des données présentant des densités variables. Cependant, DBSCAN peut être sensible aux paramètres de distance et de densité, ce qui peut avoir un impact sur les résultats obtenus. Donc , la méthode DBSCAN peut ne pas être adéquate pour notre ensemble de données, ce qui suggère ses limites dans ce contexte particulier.

L'analyse des mesures d'interclasse et d'intraclasse ainsi que des histogrammes correspondants nous permet de comparer les performances des différentes méthodes de clustering. Les résultats indiquent que K-means et K-medoids ont réussi à former des clusters distincts et compacts, tandis qu'Agnes et Diana ont rencontré des difficultés dans la séparation et la compacité des clusters. Ces observations soulignent l'importance de choisir judicieusement la méthode de clustering en fonction des objectifs de l'analyse et des caractéristiques spécifiques des données.

# Chapitre 5

## Conclusion :

En conclusion, ce projet a exploré l'utilisation de différents algorithmes de clustering tels que K-means, K-medoids, Agnes, Diana et DBSCAN. Chaque algorithme a ses propres caractéristiques et performances, ce qui offre aux praticiens une variété d'options pour l'analyse et l'interprétation des données.

L'algorithme K-means s'est révélé efficace pour regrouper les données en clusters cohérents, mais il peut être sensible à l'initialisation des centroides et peut produire des résultats différents à chaque exécution. D'autre part, l'algorithme K-medoids a montré une meilleure stabilité des résultats en utilisant des médoides au lieu de centroides, mais il peut être plus coûteux en termes de calcul.

Les algorithmes Agnes et Diana se sont avérés utiles pour la détection de structures hiérarchiques dans les données, permettant une exploration des regroupements à différents niveaux d'agrégation. Cependant, ils peuvent être sensibles aux distances utilisées pour mesurer la similarité entre les points.

Quant à Dbscan ,comparé aux autres algorithmes, DBSCAN présente certaines limitations. Il peut avoir du mal à gérer les données de haute dimensionnalité, car la distance euclidienne utilisée peut perdre de sa pertinence dans des espaces de grande dimension. De plus, DBSCAN peut être sensible aux valeurs aberrantes et à la présence de bruit excessif.

Il est important de noter que le choix de l'algorithme de clustering dépendra des objectifs spécifiques du projet, des caractéristiques des données et des contraintes du domaine d'application. Afin de prendre une décision, il est crucial de mener une analyse des performances de chaque algorithme, en se basant sur des mesures telles que les scores intra-classes et inter-classes. Ces évaluations permettent d'obtenir des indications précieuses, facilitant ainsi le choix de l'algorithme le plus approprié pour répondre aux besoins spécifiques de l'étude en question.

En conclusion,de manière générale ,il n'est pas possible de déterminer quel algorithme de clustering est meilleur que les autres .Chaque algorithme a ses propres forces et faiblesses, et ce qui peut être considéré comme le meilleur choix dans un contexte particulier peut ne pas l'être dans un autre. par contre chacun convient le mieux à une tâche spécifique de clustering donnée.Par conséquent, on conclue qu'aucune méthode ne peut être considérée comme "meilleure" . L'algorithme optimal doit être choisi en fonction des pro-



priétés uniques de la dataset qu'on dispose et des objectifs de l'analyse.