Université des Sciences et de la Technologie Houari Boumediene
Faculté d'Informatique

TP Fouille de données

Rapport TP : Les algorithmes de clustering non supervisés

Fait par :

Nom et prénom : ABDELMALEK BENMEZIANE

Matricule : 171731046778

Spécialité : M1 BIOINFO

Section : A

Contents

List of Figures

Chapter 1

Introduction générale

1.1 Introduction générale

Le data mining désigne le processus d'analyse de volumes massifs de données et du Big Data sous différents angles afin d'identifier des relations entre les data et de les transformer en informations exploitables. Ce dispositif rentre dans le cadre de la Business Intelligence et a pour but d'aider les entreprises à résoudre des problèmes, à atténuer des risques et à identifier et saisir de nouvelles opportunités business.

En français, ce processus porte différents noms :

- Exploration de données.
- Fouille de données.
- Forage de données.
- Ou encore extraction de connaissances à partir de données.

Le data mining n'est pas un concept récent. Déjà au XVIIème siècle, les individus cherchaient des solutions pour analyser les données et identifier des caractéristiques communes.

Chapter 2

Le preprocessing

2.1 Le preprocessing et ses techniques

2.1.1 Définition de preprocessing

Le preprocessing des données dans le machine learning (ML) est une étape cruciale qui permet d'améliorer la qualité des données afin de promouvoir l'extraction d'informations significatives à partir des données. Le preprocessing des données dans Machine Learning fait référence à la technique de préparation (nettoyage et organisation) des données brutes pour les rendre adaptées à la construction et à la formation de modèles Machine Learning.

2.1.2 Les techniques du preprocessing

L'ouverture du dataset (open dataset)

Il faut ouvrir la dataset voulu pour pouvoir appliquer le preprocessing.

Le nettoyage du dataset (Data Cleaning)

- Convertir les attributs catégoriels en numérique (Encodage des données catégorielles)

  Les données catégorielles font référence aux informations qui ont des catégories spécifiques dans l'ensemble de données.

Les modèles d'apprentissage automatique sont principalement basés sur des équations mathématiques. Ainsi, vous pouvez intuitivement comprendre que le fait de conserver les données catégorielles dans l'équation causera certains problèmes puisque vous n'auriez besoin que de nombres dans les équations.

- Identifier et traiter les valeurs manquantes (missing values)
  Dans le preprocessing des données, il est essentiel d'identifier et de gérer correctement les valeurs manquantes, fondamentalement, il existe deux façons de gérer les données manquantes :
  Suppression d'une ligne particulière : Dans cette méthode, vous supprimez une ligne spécifique qui a une valeur nulle pour une caractéristique ou une colonne particulière où plus de 75% des valeurs sont manquantes. Cependant, cette méthode n'est pas efficace à 100%.
  Calcul de la moyenne : Cette méthode est utile pour les fonctionnalités contenant des données numériques telles que l'âge, le salaire, l'année, etc. Ici, vous pouvez calculer la moyenne, la médiane ou le mode d'une colonne.

La normalisation du dataset (Data Normalization)

La normalisation des données est une technique utilisée dans l'exploration de données pour transformer les valeurs d'un ensemble de données en une échelle commune. Ceci est important car de nombreux algorithmes d'apprentissage automatique sont sensibles à l'échelle des caractéristiques d'entrée et peuvent produire de meilleurs résultats lorsque les données sont normalisées.

2.2 Définition de l'apprentissage non supervisé

L'apprentissage non supervisé est une branche du machine learning, caractérisée par l'analyse et le regroupement de données non-étiquetées. Pour cela, ces algorithmes apprennent à trouver des schémas ou des groupes dans

les données, avec très peu d'intervention humaine. En termes mathématiques, l'apprentissage non supervisé implique l'observation de plusieures occurrences d'un vecteur X and l'apprentissage de la probabilité de distribution p(X) pour ces occurrences.

Chapter 3

Les techniques et caractéristiques

3.1 Les algorithmes de clustering non supervisés

3.2 K-Means

K-Means est l'un des algorithmes de clustering les plus répandus. Il permet d'analyser une dataset caractérisées par un ensemble de descripteurs, afin de regrouper les données "similaires" en terme de "distance" en groupes (ou clusters).

3.3 K-Medoids

K-Medoids est un algorithme de clustering ressemblant à la technique de clustering K-Means, il diffère principalement de l'algorithme K-Means par la manière dont il sélectionne les centres des clusters.

3.4 AGNES

AGNES (Agglomerative Nesting) est l'un des algorithmes de clustering hiérarchique les plus populaires utilisés dans l'exploration de données.
L'algorithme AGNES utilise une approche "ascendante" pour le clustering hiérarchique. L'algorithme forme des clusters singleton de chacun des points de données. Il les regroupe ensuite de bas en haut dans la structure ar-

borescente (appelée dendrogramme) jusqu'à ce que tous les points similaires forment un seul cluster (représenté par la racine du dendrogramme).

3.5 DIANA

DIANA est également connu sous le nom d'algorithme de clustering DIvisie ANAlysis. Il s'agit de la forme d'approche descendante du clustering hiérarchique où tous les points de données sont initialement affectés à un seul cluster. De plus, les clusters sont divisés en deux clusters les moins similaires.

3.6 DBSCAN

DBSCAN signifie Density-Based Spatial Clustering of Applications with Noise, est un algorithme qui regroupe des points de données « densément groupés » dans un seul cluster. DBSCAN ne requiert que deux paramètres : epsilon et minPoints. Epsilon est le rayon du cercle à créer autour de chaque point de données pour vérifier la densité et minPoints est le nombre minimum de points de données requis à l'intérieur de ce cercle pour que ce point de données soit classé comme point central.

Chapter 4

Les resultats obtenus

4.1 L'interface principale

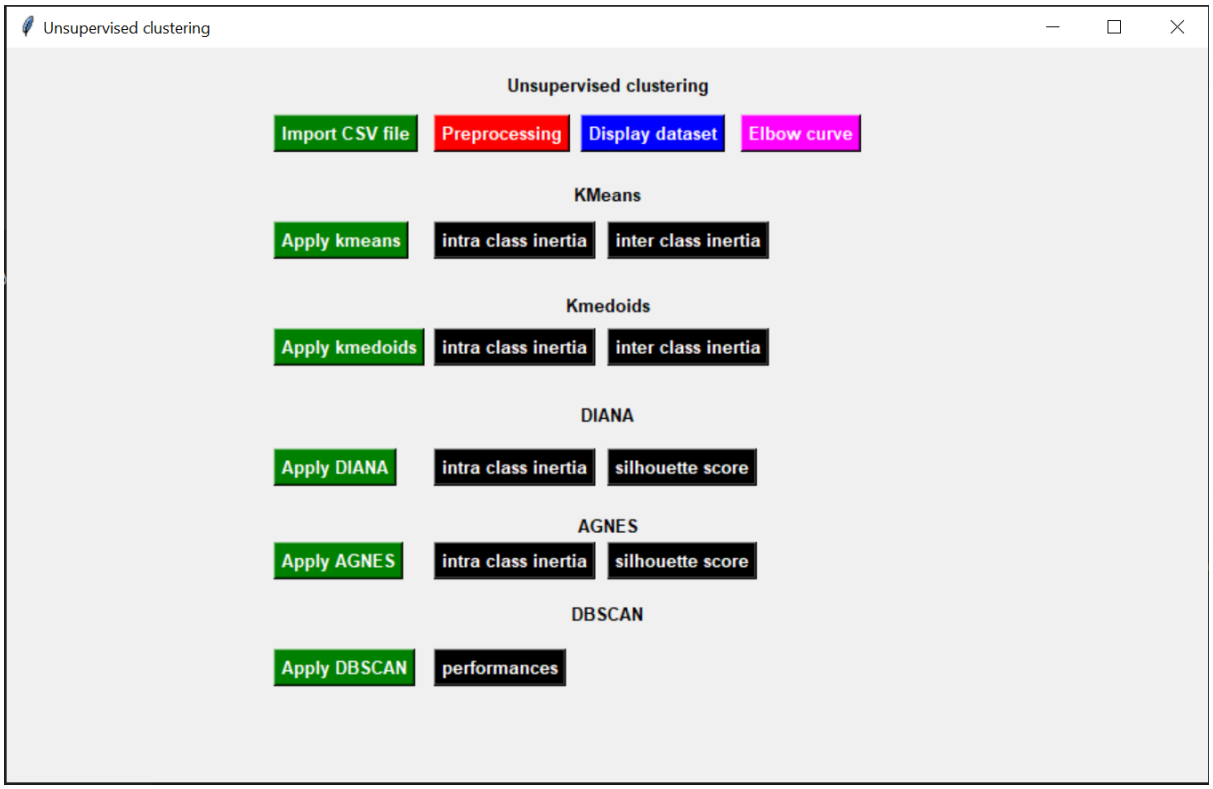

Figure 4.1: Interface

4.2 L'importation d'un dataset

Figure 4.2: Importer dataset

Figure 4.3: Le dataset diabetes

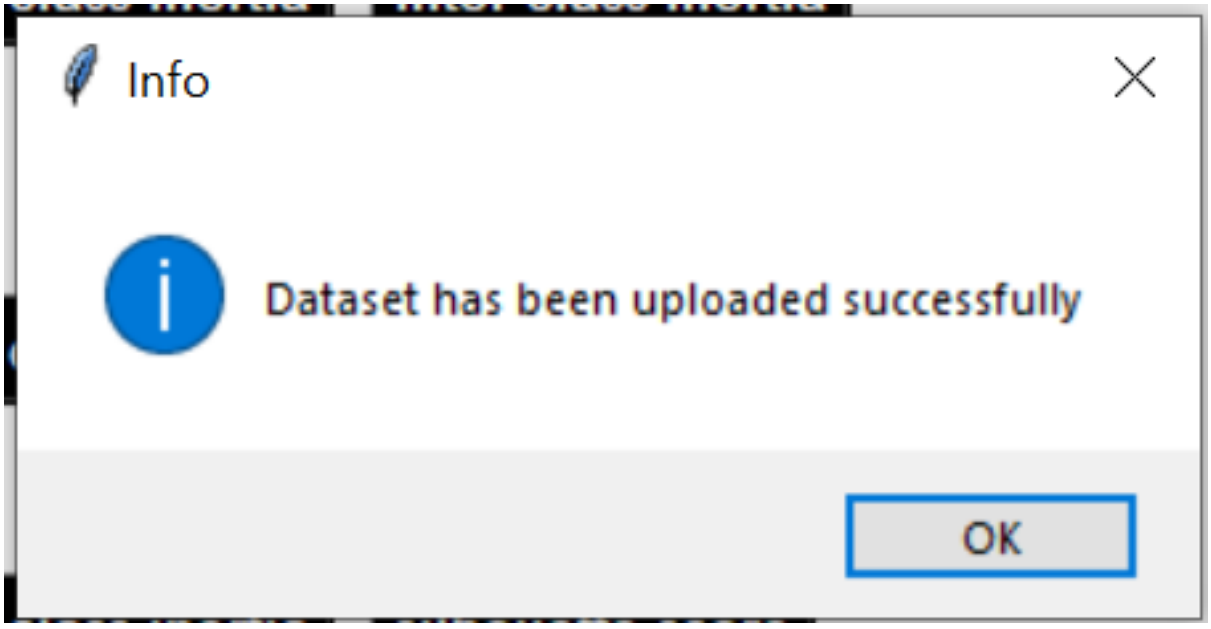

Figure 4.4: Output

4.3 Le preprocessing

Figure 4.5: Preprocessing

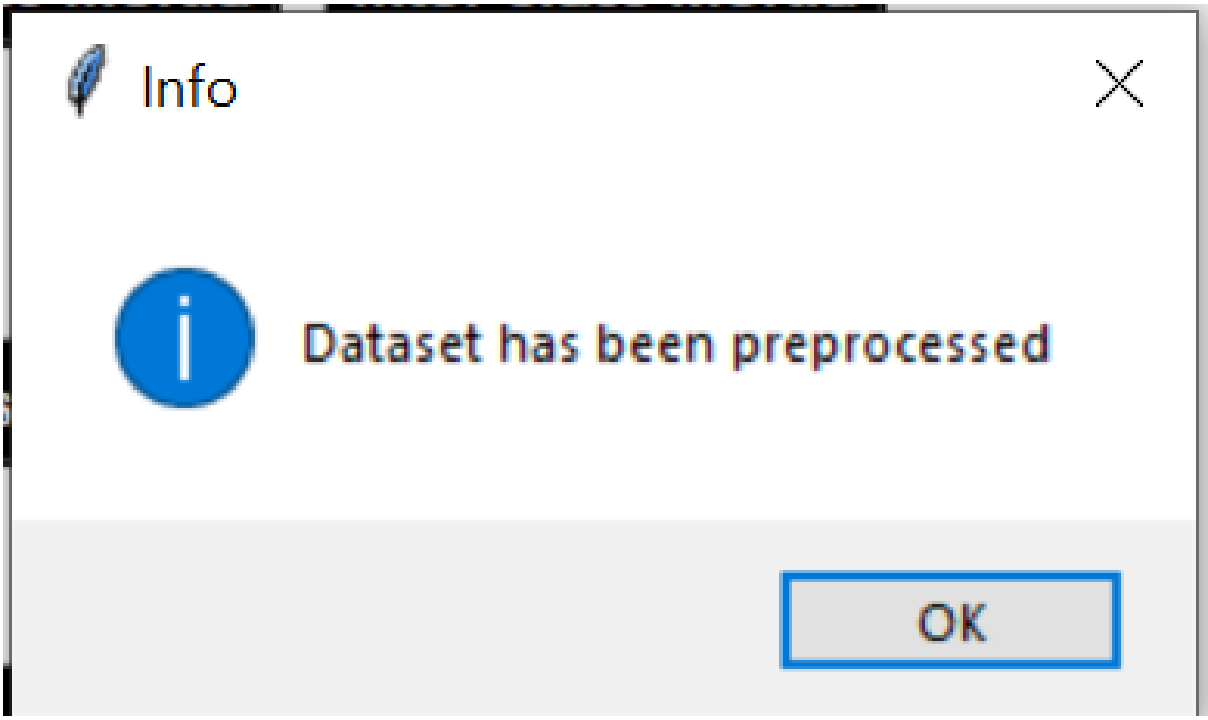

Figure 4.6: Output

4.4 La courbe d'elbow

Figure 4.7: Elbow

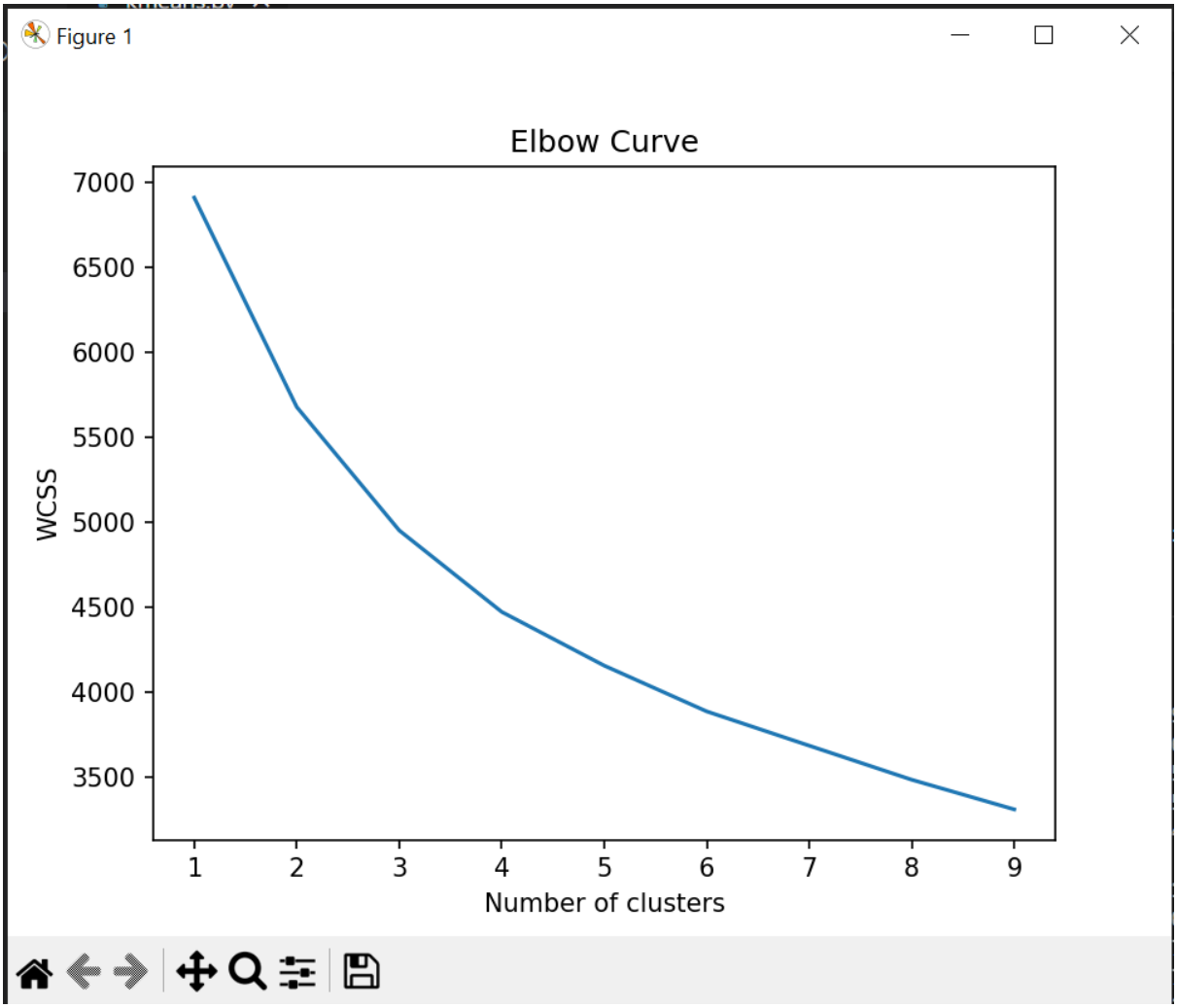

Figure 4.8: Output

Le nombre optimale des clusters est égale à : 2

4.5 Kmeans

Figure 4.9: Kmeans

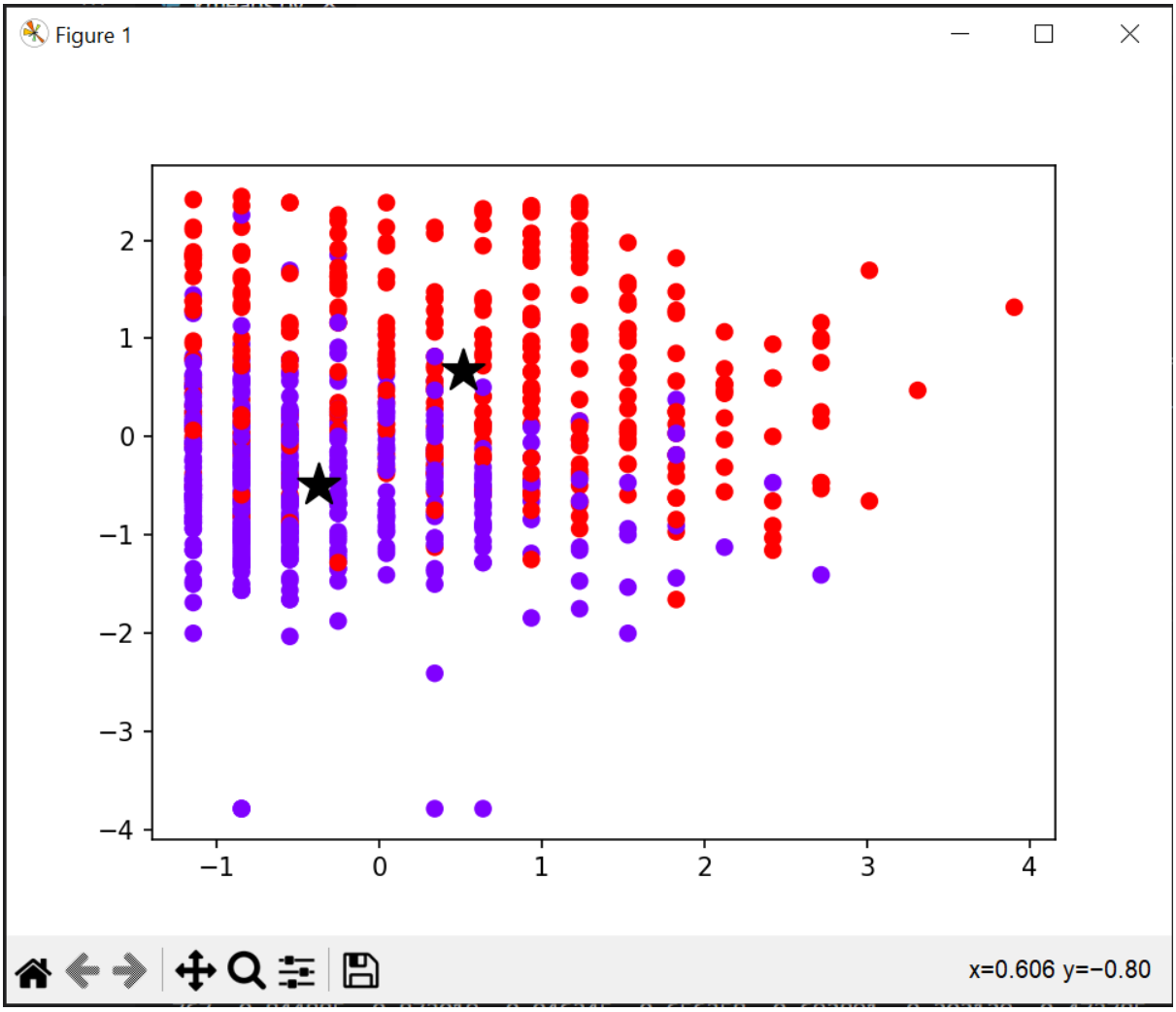

Figure 4.10: Output

4.5.1 L'inertie intra-classe

Figure 4.11: Inertie intra-classe

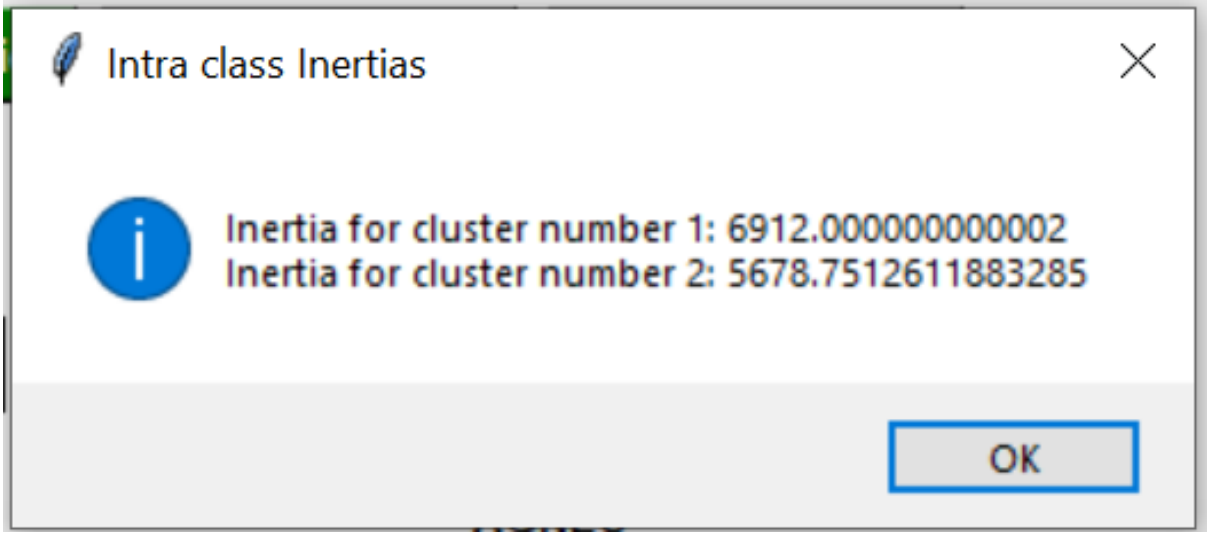

Figure 4.12: Output

4.5.2 L'inertie inter-classe

Figure 4.13: Inertie inter-classe

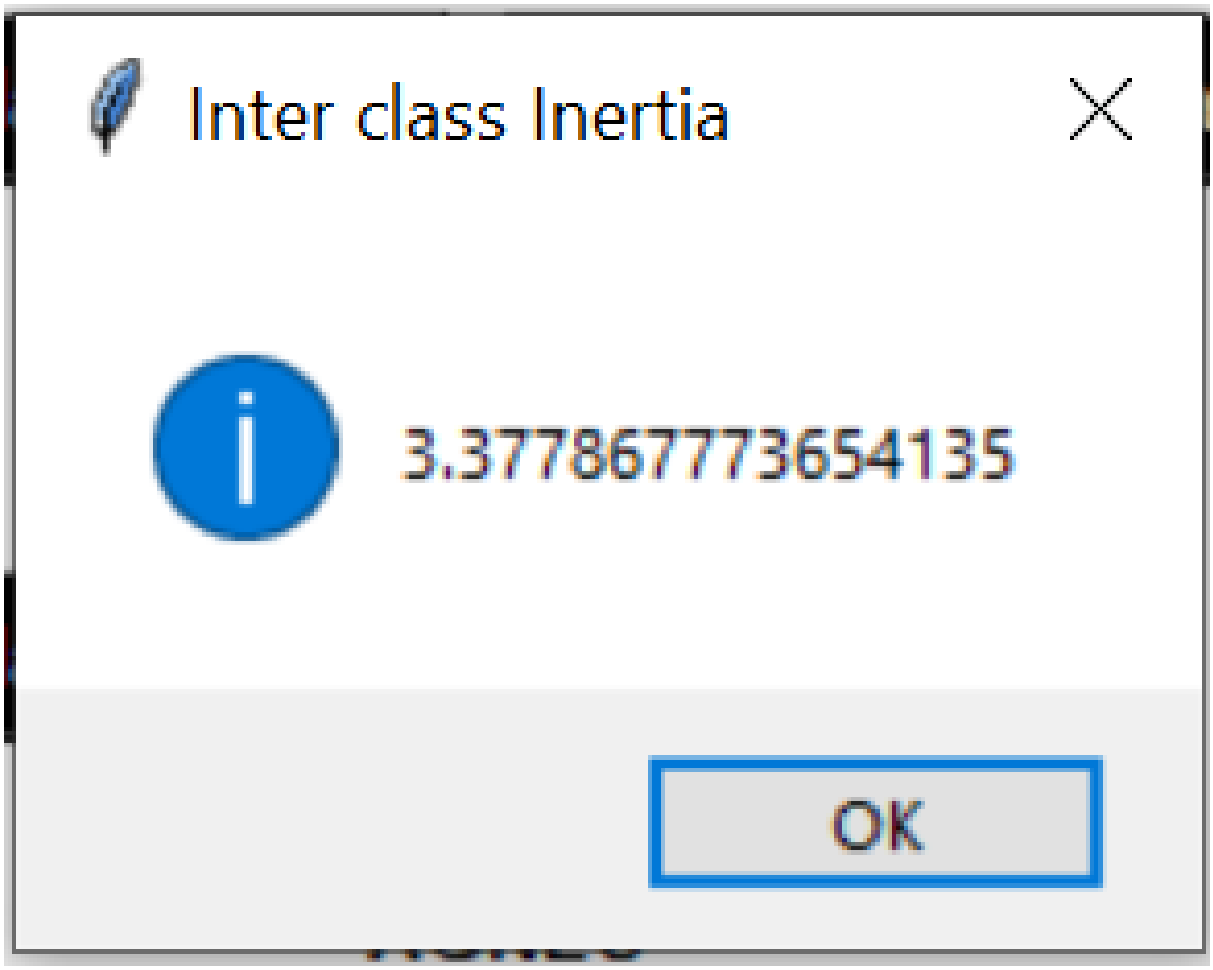

Figure 4.14: Output

4.6 Kmedoids

Figure 4.15: kmedoids

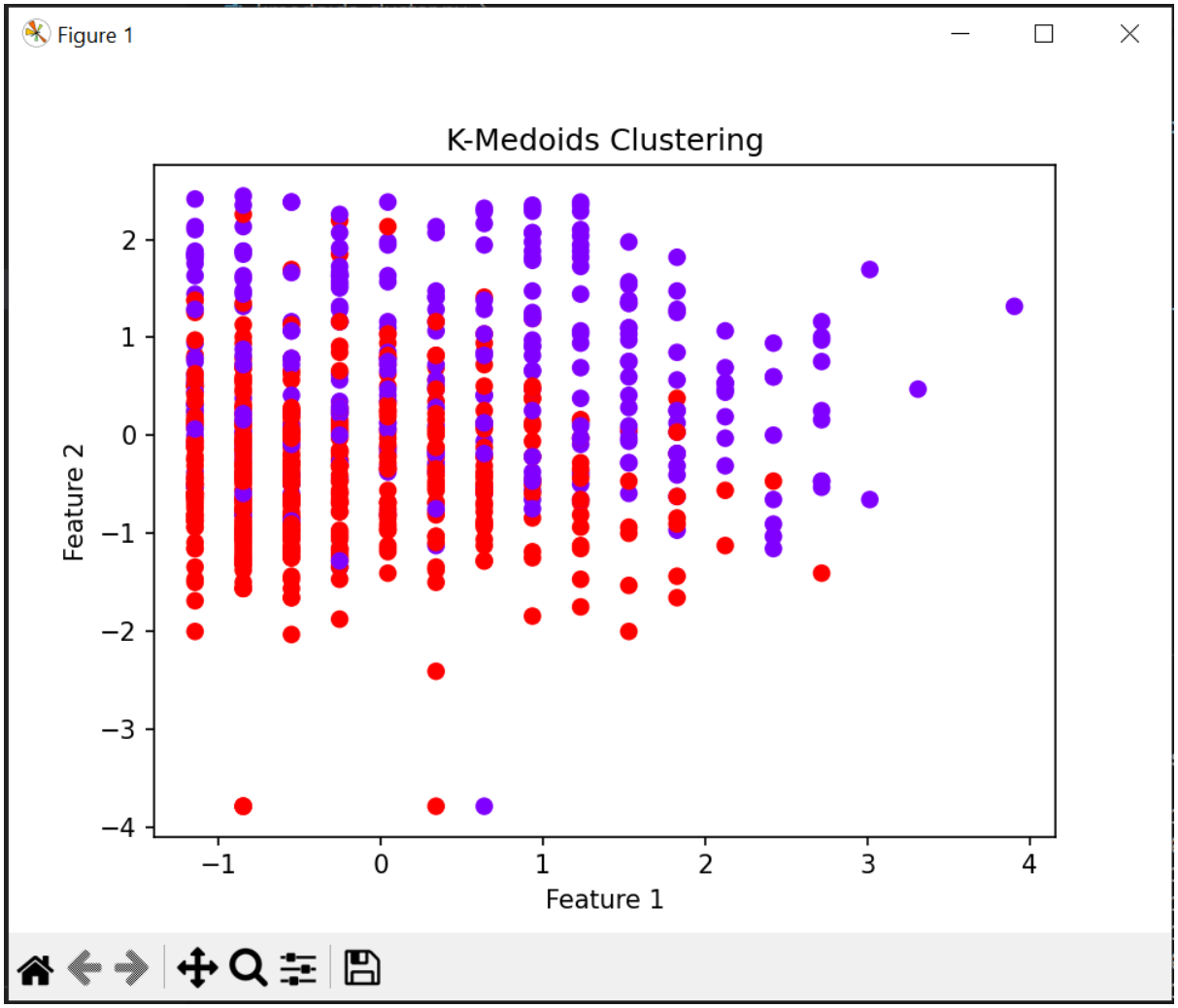

Figure 4.16: Output

4.6.1 L'inertie intra-classe

Figure 4.17: Inertie intra-classe

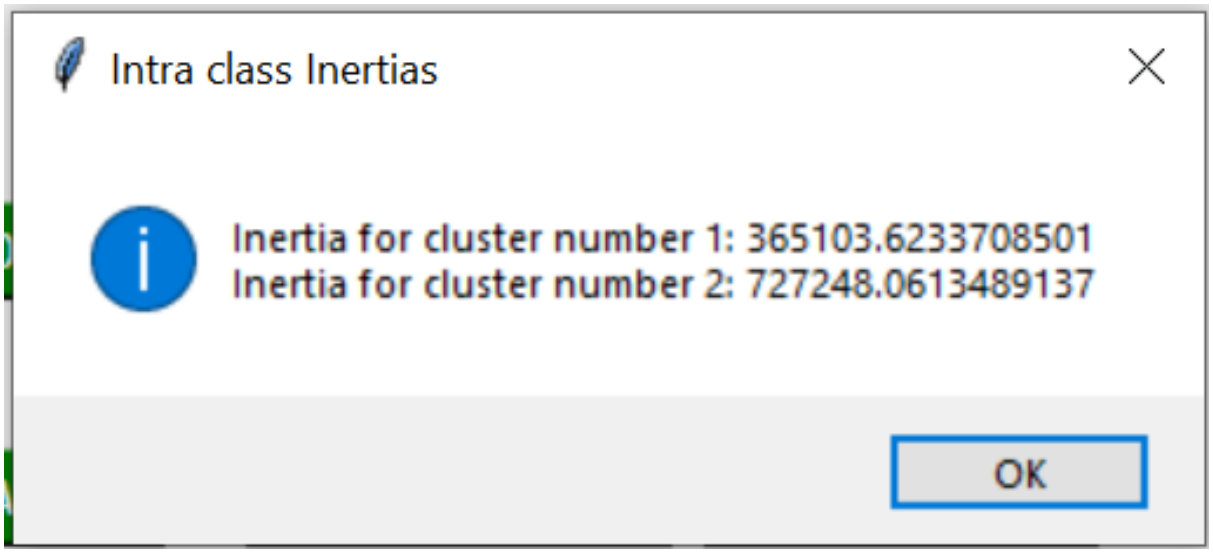

Figure 4.18: Output

4.6.2 L'inertie inter-classe

Figure 4.19: Inertie inter-classe

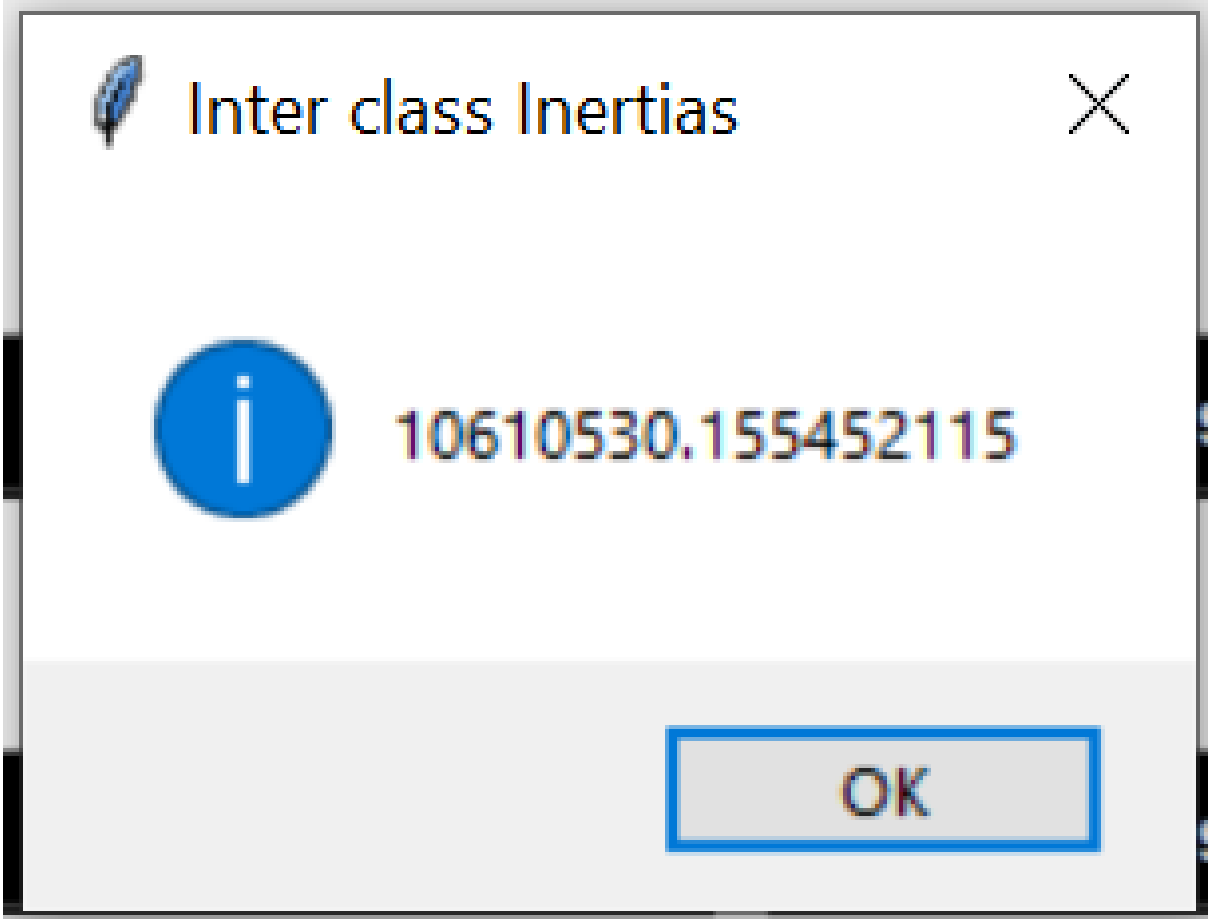

Figure 4.20: Output

4.7 AGNES

Figure 4.21: Agnes

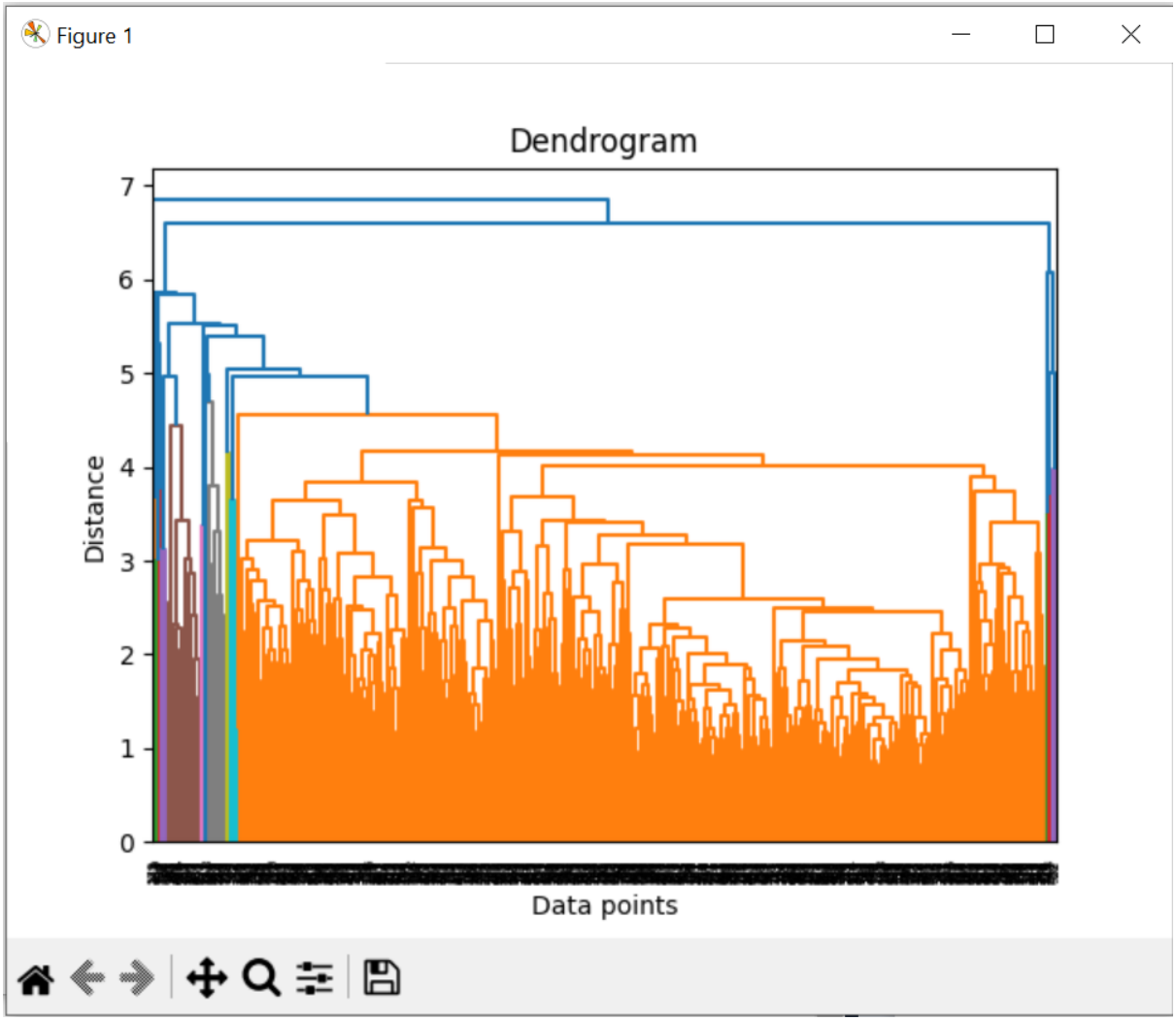

Figure 4.22: Output

4.7.1 L'inertie intra-classe

Figure 4.23: Inertie intra-classe

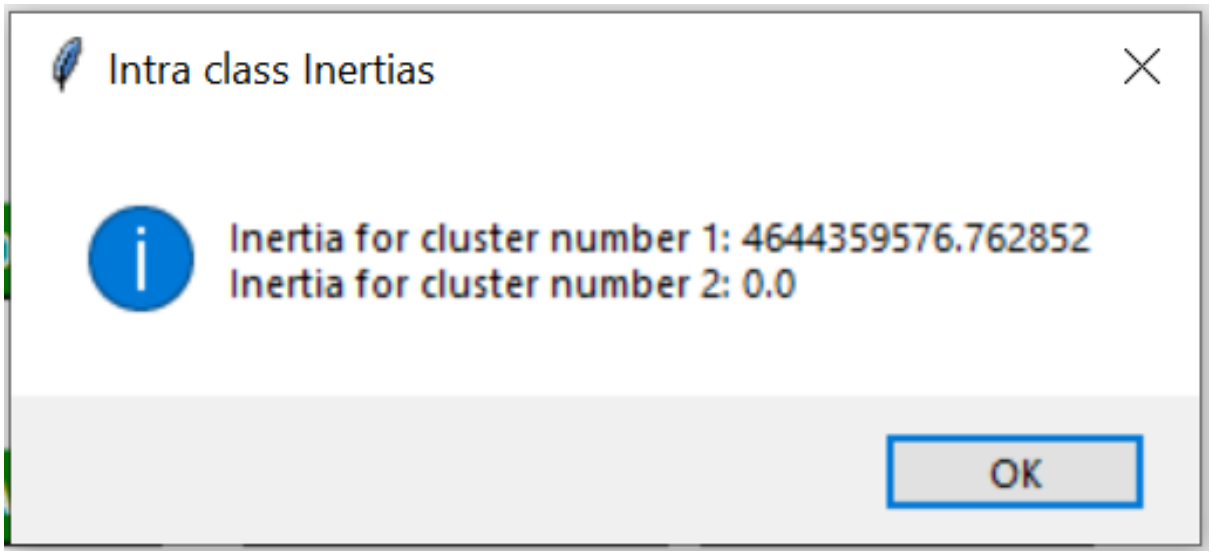

Figure 4.24: Output

4.7.2 Coefficient de silhouette

Figure 4.25: Silhouette

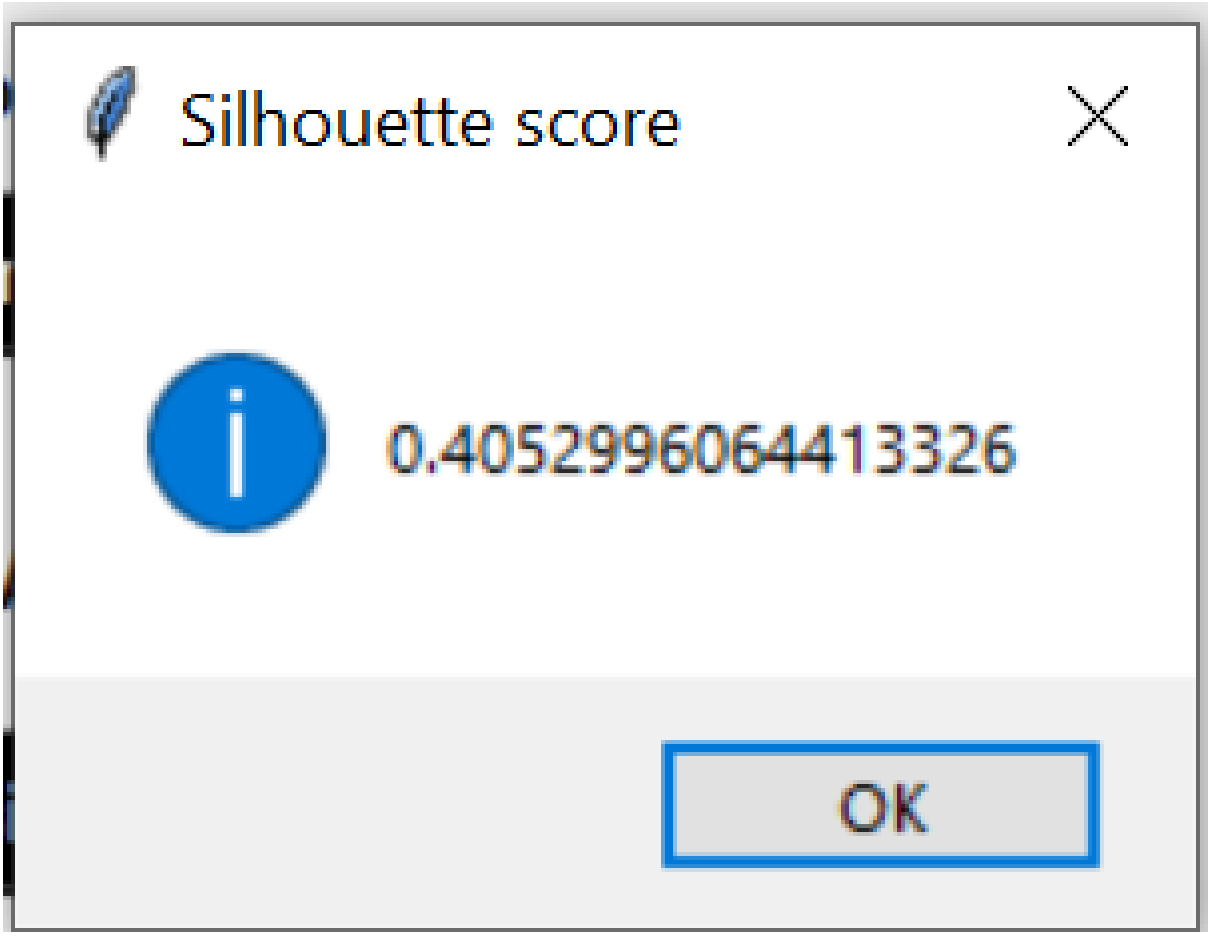

Figure 4.26: Output

4.8 DIANA

Figure 4.27: Diana

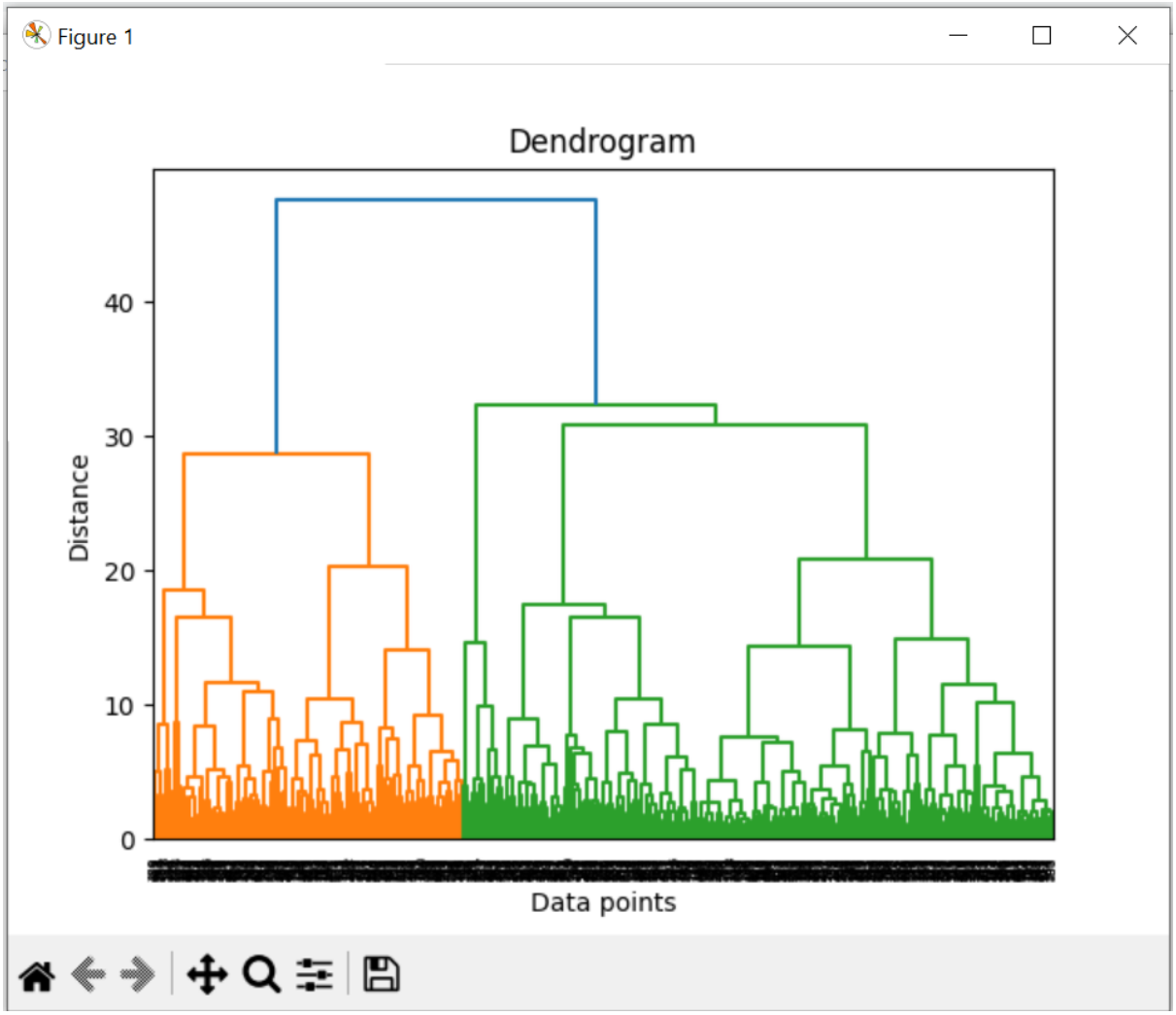

Figure 4.28: Output

4.8.1 L'inertie intra-classe

Figure 4.29: Inertie intra-classe

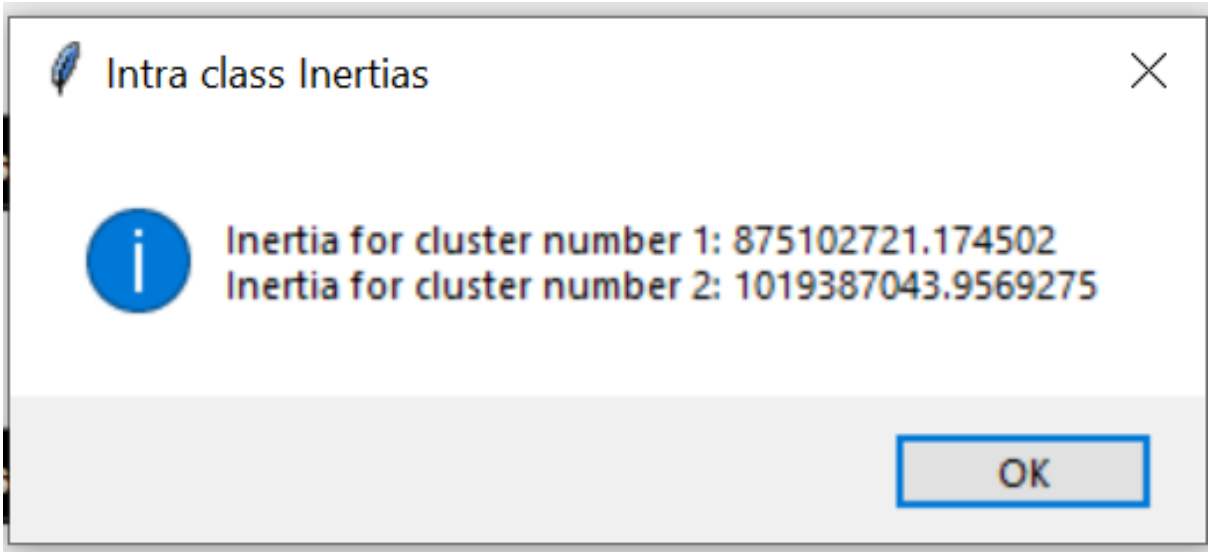

Figure 4.30: Output

4.8.2 Coefficient de silhouette

Figure 4.31: Silhouette

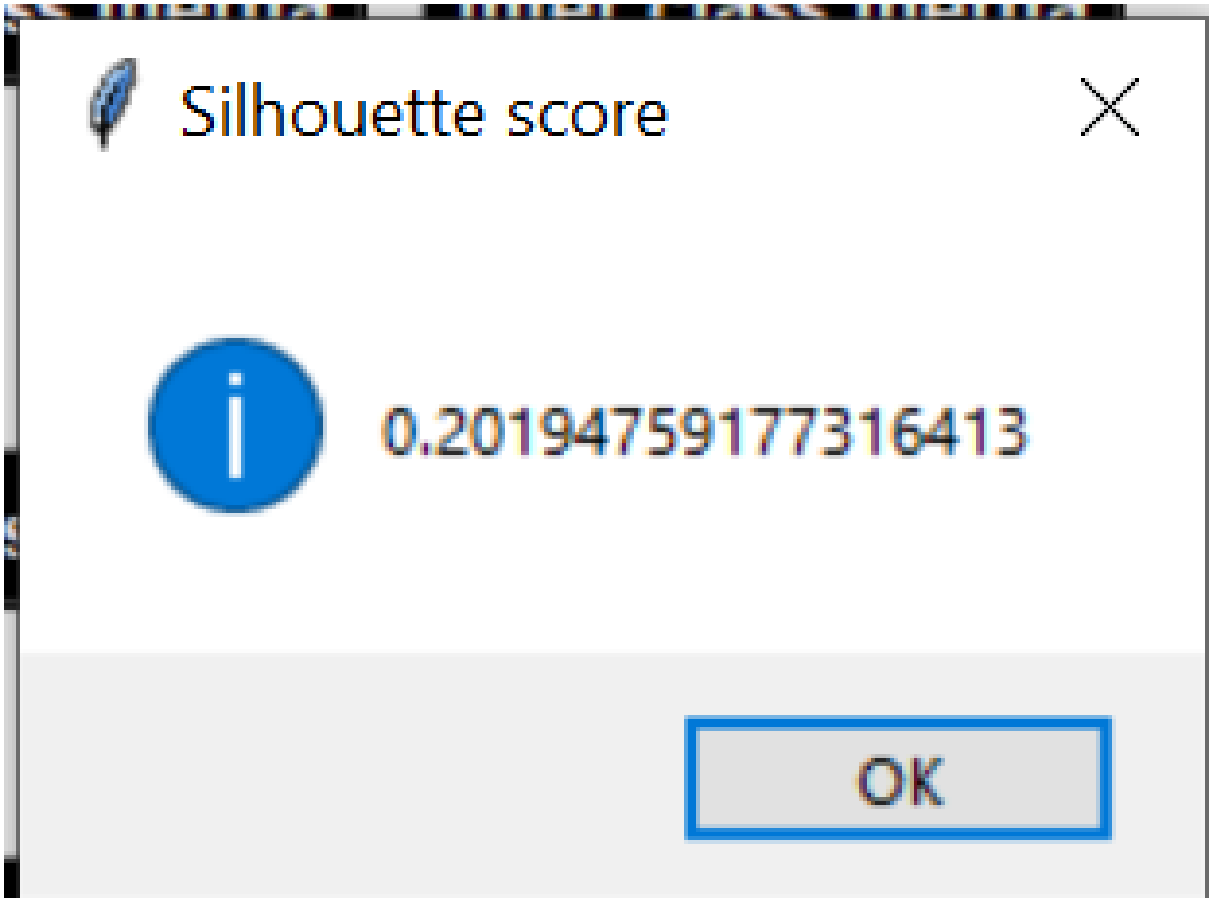

Figure 4.32: Output

4.9 DBSCAN

Figure 4.33: Dbscan

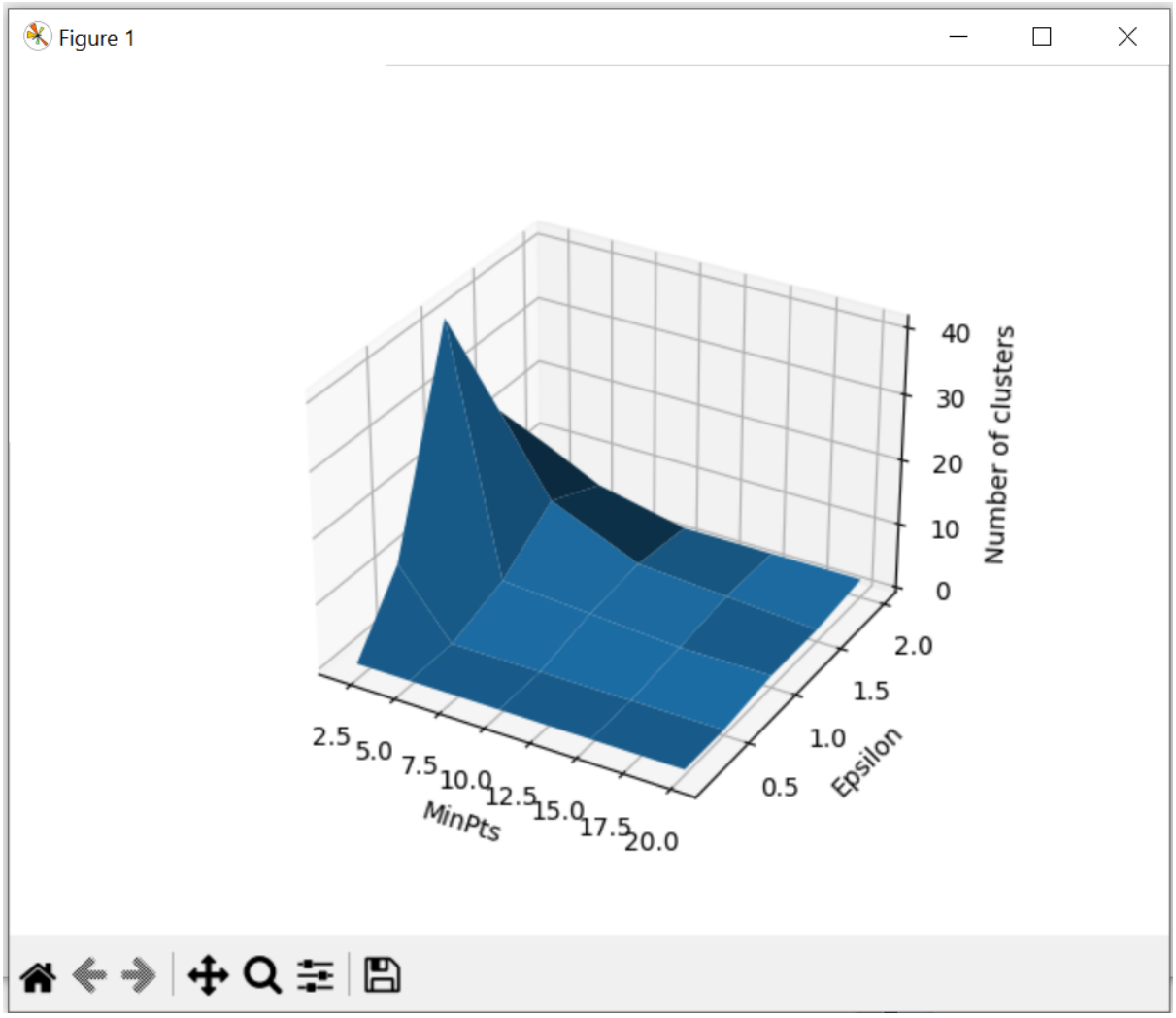

Figure 4.34: Output

4.9.1 Afficher les performanes

Figure 4.35: Performances

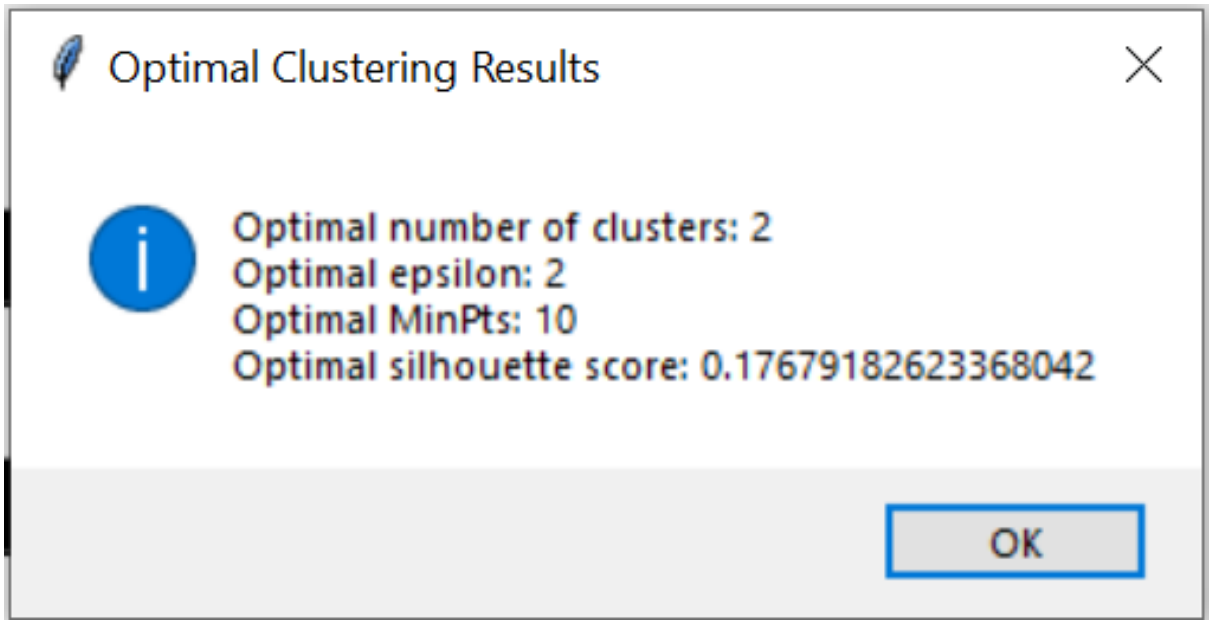

Figure 4.36: Output

Chapter 5

Conclusion générale

Le clustering est une technique d'apprentissage automatique non supervisé qui regroupe les points de données en clusters en fonction de la similitude des informations disponibles pour les points de données dans l'ensemble de données. Les points de données appartenant aux mêmes clusters sont similaires les uns aux autres à certains égards, tandis que les éléments de données appartenant à différents clusters sont différents.

K-means, PAM (Partition Around Medoids), AHC (Agglomerative Hierarchical Clustering), DBScan (Density Based Spatial Clustering of Applications with Noise) sont les algorithmes de clustering dans l'apprentissage automatique non supervisé.

Chaque algorithmes de clustering ont plusieurs différences entre eux, mais nous on va concentrer sur les différences pricipales.

5.1 Kmeans vs PAM (Kmedoids)

La principale différence entre K-means et la méthode PAM est que K-means utilise des centroïdes (généralement des points artificiels), tandis que PAM utilise des medodoïdes, qui sont toujours les points réels de l'ensemble de données.

5.2 Kmeans vs AHC (Agglomerative Hierarchical Clustering)

La principale différence entre K-means et la méthode AHC est que K-means est utilisé lorsque le nombre de classes est fixe, tandis que ce dernier est utilisé pour un nombre inconnu de classes.

5.3 Kmeans vs DBSCAN (Density Based Spatial Clustering of Applications with Noise)

la principale différence entre K-means et la méthode DBSCAN est que les clusters formées par kmeans sont de forme plus ou moins sphérique ou convexe et doivent avoir la même taille de caractéristique, par contre les clusters formés par DBSCAN sont de forme arbitraire et peuvent ne pas avoir la même taille d'entité.