Exercice : Implémentation de K-means Clustering

Objectif :

Dans cet exercice, vous implémenterez l'algorithme K-means sur un ensemble de données non étiquetées. Vous explorerez les concepts de clustering, de distance euclidienne, et utiliserez des visualisations pour analyser les résultats.

Instructions :

Répondez aux questions étape par étape en utilisant du code pour justifier vos réponses.

Question 1 : Chargement et Exploration des Données

- Chargez l'ensemble de données fourni.
- Affichez les premières lignes du jeu de données.
- Réalisez une analyse rapide des statistiques descriptives des variables.

Question : Quelles caractéristiques peuvent être potentiellement utiles pour le clustering ?

Question 2 : Implémentation de l'Algorithme K-means

- Utilisez KMeans de la bibliothèque sklearn pour effectuer le clustering sur les données.
- Choisissez un nombre de clusters k=3 et entraînez le modèle.

Questions :

1. Pourquoi est-il nécessaire de choisir un nombre de clusters avant d'entraîner le modèle ?

2. Comment le modèle K-means détermine-t-il les centres de clusters ?

Question 3 : Visualisation des Clusters

- Visualisez les résultats du clustering en utilisant des graphiques de dispersion (scatter plots).

Question :

Quelle est l'importance de visualiser les clusters, et quelles informations peut-on en tirer ?

Question 4 : Mesurer la Performance du Modèle

- Calculez l'inertie (ou somme des distances au carré des points à leur centre de cluster).

- Utilisez la méthode du coude (elbow method) pour déterminer le nombre optimal de clusters.

Question :

Que représente l'inertie dans l'algorithme K-means, et comment la méthode du coude aide-t-elle à choisir le nombre optimal de clusters ?

Livrables :

1. Un Jupyter Notebook avec le code, les visualisations, et les réponses aux questions.
2. Un rapport court expliquant le processus de clustering et les résultats obtenus.