

La fouille de données, ou "data mining", est une démarche qui vise à extraire des informations pertinentes et inédites à partir de vastes ensembles de données.

Objectif du TP N°3 :

Ce TP a pour but de vous familiariser avec les techniques de clustering à travers l'analyse d'un benchmark. Les étapes clés sont les suivantes :

1. Préparation des données :

- Chargement et prétraitement du benchmark.

2. Détermination du nombre optimal de clusters :

- Utilisation de la méthode du coude (courbe d'elbow) pour identifier le nombre de groupes présents dans les données.

3. Application des algorithmes de clustering hiérarchique :

- Mise en œuvre des algorithmes AGNES et DIANA, en utilisant le nombre de clusters déterminé précédemment.

4. Application de l'algorithme de clustering basé sur la densité :

- Utilisation de l'algorithme DBSCAN, en ajustant les paramètres MinPts et EPS pour obtenir le même nombre de clusters que celui trouvé avec la courbe d'elbow.

5. Évaluation des performances :

- Calcul des métriques de performance pour les clusters obtenus avec chaque algorithme (AGNES, DIANA et DBSCAN).

6. Comparaison des résultats :

- Création d'un histogramme des inerties pour les cinq méthodes étudiées (K-Means, K-Medoids, AGNES, DIANA et DBSCAN).
- Analyse comparative des performances de chaque méthode.

Rapport à remettre le 17 /11/2025