TRAVAUX PRATIQUES

CLUSTERING NON SUPERVISÉ

K-Means • Hiérarchique • DBSCAN

Contexte du projet

Maven Supermarket souhaite optimiser l'organisation de ses rayons de céréales. L'objectif est d'installer des présentoirs regroupant des céréales aux caractéristiques nutritionnelles similaires pour faciliter les choix des clients.

Votre mission : Tester trois approches de clustering (K-Means, Hiérarchique, DBSCAN) pour identifier des groupes de céréales homogènes et recommander la meilleure stratégie de merchandising.

Fichier de données : cereal.csv (caractéristiques nutritionnelles des céréales : calories, protéines, lipides, sodium, fibres, glucides, sucres, vitamines, etc.)

PARTIE 1 : Exploration initiale avec K-Means

Objectif : Comprendre la structure des données et identifier le nombre optimal de clusters

Tâche 1.1 : Préparation des données

- Charger le fichier cereal.csv et explorer sa structure
- Supprimer les colonnes non numériques (Nom, Fabricant)
- Vérifier la présence de valeurs manquantes et les gérer si nécessaire

Tâche 1.2 : Premier clustering (K=2)

- Ajuster un modèle K-Means avec K=2 clusters sur les données originales
- Analyser les centres des clusters : quelles sont les caractéristiques distinctives ?

Tâche 1.3 : Méthode du coude (Elbow method)

- Créer une boucle pour tester K de 2 à 15 clusters
- Calculer et stocker l'inertie (WCSS) pour chaque valeur de K
- Créer le graphique du coude (K vs Inertie)
- Identifier visuellement le K optimal et justifier votre choix

Tâche 1.4 : Modèle K-Means optimal

- Ajuster le modèle K-Means avec le K optimal identifié
- Calculer le score de silhouette du modèle
- Créer une heatmap des centres de clusters pour interpréter les groupes

PARTIE 2 : Clustering hiérarchique

Objectif : Explorer une approche alternative basée sur la hiérarchie et comparer avec K-Means

Tâche 2.1 : Dendrogramme sur données originales

- Sélectionner 5 variables numériques pertinentes (ex: calories, protéines, lipides, glucides, fibres)
- Créer un dendrogramme avec la méthode de liaison Ward
- Identifier visuellement le nombre optimal de clusters et ajuster le seuil de couleur

Tâche 2.2 : Standardisation et nouveau dendrogramme

- Standardiser les 5 variables sélectionnées (StandardScaler)
- Créer un nouveau dendrogramme avec les données standardisées
- Comparer les deux dendrogrammes : qu'est-ce qui change ?

Tâche 2.3 : Modèle hiérarchique final

- Choisir le meilleur ensemble de données (original ou standardisé) avec justification
- Ajuster un modèle de clustering hiérarchique avec le nombre optimal de clusters
- Calculer le score de silhouette
- Créer une heatmap pour visualiser et interpréter les clusters

PARTIE 3 : Clustering basé sur la densité (DBSCAN)

Objectif : Découvrir des groupes naturels et détecter automatiquement les outliers

Tâche 3.1 : Recherche des paramètres optimaux

- Tester différentes combinaisons de paramètres :
  - eps : de 0,1 à 2,0 par incréments de 0,1
  - min_samples : de 2 à 10 par incréments de 1
- Calculer le score de silhouette pour chaque combinaison valide (ignorer les modèles à 1 cluster)
- Identifier les meilleurs paramètres (eps, min_samples)

Tâche 3.2 : Analyse des outliers

- Ajuster le modèle DBSCAN final avec les paramètres optimaux
- Identifier les céréales classées comme bruit (label = -1)
- Analyser leurs caractéristiques : pourquoi sont-elles des outliers ?
- Créer une visualisation 2D avec les outliers en couleur distincte

PARTIE 4 : Analyse comparative et recommandation

Objectif : Comparer les trois méthodes et formuler une recommandation stratégique

Tâche 4.1 : Tableau comparatif quantitatif

- Créer un tableau comparant : Méthode | Nombre de clusters | Score de silhouette | Outliers détectés | Temps de calcul

Tâche 4.2 : Analyse des affectations

- Créer un tableau avec : Nom de céréale | Cluster K-Means | Cluster Hiérarchique | Cluster DBSCAN
- Identifier les céréales qui changent de cluster selon la méthode et analyser pourquoi

Tâche 4.3 : Avantages et limites

- Pour chaque méthode, lister les avantages et inconvénients observés sur ce dataset
- Quelle méthode a le mieux géré les outliers ? Justifier

Tâche 4.4 : Recommandation stratégique pour Maven Supermarket

- Quelle méthode de clustering recommandez-vous ? Justifiez en fonction de :
  - Qualité du clustering (score de silhouette)
  - Interprétabilité business
  - Gestion des produits atypiques
- Combien de présentoirs Maven Supermarket devrait-il installer ?