Définition études de marché

Est-ce que les deux méthodes de clustering ont donné les mêmes résultats ?

Nous avons comparé les résultats de la Classification Ascendante Hiérarchique (CAH) et de l'algorithme K-means.

Le score Adjusted Rand Index (ARI) est de 1.0, ce qui signifie que les deux méthodes ont donné exactement les mêmes clusters.

Cela montre que la structure des données est bien marquée et détectée de manière cohérente par les deux approches.

Score ARI = 1 , ce qui nous dévoile que les résultats des 2 clusters sont très proches.

Est-ce qu'il y a un pays qui ressort plus que les autres ?

Oui la suède ressort plus que les autres par le revenu par habitant le plus élevé.

En quoi est-ce intéressant d'utiliser une ACP ?

But : Réduire le nombre de variables (features) tout en conservant un maximum d'information (variance).

L'ACP est intéressante en Python data car elle permet :

de simplifier les données, de mieux visualiser et comprendre leur structure,
de préparer efficacement les données pour les modèles d'analyse prédictive.

L'ACP peut révéler des regroupements naturels dans les données.
Peut être utilisée en amont d'un algorithme de clustering comme K-Means pour améliorer les performances.

Quand on a beaucoup de variables (plus que 3), on ne peut pas visualiser facilement les données.

Avec une ACP, on peut projeter les données sur 2 dimensions principales et faire un scatter plot.

But : Réduire le nombre de variables (features) tout en conservant un maximum d'information (variance)

Pourquoi c'est utile : Diminue le temps de calcul, Réduit le risque de surapprentissage (*overfitting*), Facilite la visualisation (ex. : en projetant sur 2 ou 3 composantes principales).

Classification ascendante hierarchique

C’est une méthode de regroupement (clustering) utilisée pour former des groupes d'individus similaires dans un jeu de données sans avoir besoin de spécifier le nombre de clusters à l'avance.

Identifier des groupes naturels dans les données.

Visualiser la structure hiérarchique entre les observations.
Choisir un nombre optimal de clusters en coupant le dendrogramme à un certain niveau.

Coefficient de Silhouette

Il permet de savoir si les éléments sont bien regroupés dans leur cluster et séparés des autres clusters.

Évaluer la cohérence des clusters formés.

Choisir le meilleur nombre de clusters.
Identifier si certains points sont mal classés.

Classification kmeans

C’est une méthode qui permet de regrouper automatiquement des données similaires en k groupes (clusters)

Rapide et simple

Fonctionne bien sur de grands volumes de données

Algorithme : K-means

Type : Clustering non supervisé

But : Regrouper les données en k groupes similaires

Paramètre clé : Nombre de clusters kk

Résultat : Chaque donnée reçoit un label de cluster

Méthode du coude

C’est une technique utilisée pour choisir le bon nombre de clusters kkk dans un algorithme de clustering, notamment K-means

Trouver le nombre optimal de clusters qui permet de bien regrouper les données sans en faire trop.

Nom : Méthode du coude

Utilité : Trouver le nombre optimal de clusters kk

Basée sur : L’évolution de l’inertie intra-cluster

Interprétation : Chercher l’endroit où la courbe forme un "coude"

Centroïdes

Le centroïde est la moyenne de toutes les coordonnées des points d’un cluster.

En d’autres mots, c’est le "centre de gravité" du groupe.

Plan factoriel

Un plan factoriel est une représentation graphique des données projetées sur les axes principaux obtenus après une Analyse en Composantes Principales (ACP).

Après une ACP :

On obtient des axes principaux (appelés composantes principales), comme F1, F2, etc.

Ces axes sont des combinaisons linéaires des variables d'origine.

Le plan factoriel est le plan formé par les deux premières composantes principales (F1 et F2), qui contiennent généralement le plus d'information.