**BUT Informatique — TP IA : Clustering (1h) — Mall Customers**

**Dataset : Mall Customers (CSV : Mall\_Customers.csv)**

## Contexte & objectifs

Vous allez pratiquer le clustering non supervisé sur le dataset \*\*Mall Customers\*\* (clients d’un centre commercial) contenant des informations démographiques et de consommation. L’objectif est d’identifier des segments de clientèle par similarité de comportements.

Colonnes typiques : CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100). Nous concentrerons l'analyse sur \*\*Age\*\*, \*\*Annual Income (k$)\*\*, \*\*Spending Score (1-100)\*\* et encoderaons \*\*Gender\*\* si nécessaire.

## Consignes pratiques

Durée : ~1h

Pré-requis : Python, scikit-learn, numpy, pandas, matplotlib installés

Fichier requis : placez \*\*Mall\_Customers.csv\*\* dans le même dossier que votre notebook.

Livrables : votre notebook (.ipynb) complété et commenté

## Étapes à suivre

1) Chargement & préparation des données

- Chargez `Mall\_Customers.csv` dans un DataFrame pandas.  
- Affichez `head()`, `shape`, `info()` et `describe()`.  
- Sélectionnez les variables pertinentes (ex : Age, Annual Income (k$), Spending Score (1-100)).  
- Gérez \*\*Gender\*\* : encodez-le si vous décidez de l'inclure (LabelEncoder).

2) Mise à l’échelle & PCA (2D)

- Standardisez les variables numériques (StandardScaler).  
- Appliquez une ACP (PCA) 2 composantes.  
- Visualisez le nuage de points (matplotlib) et commentez la structure apparente.

3) Clustering avec K-Means

- Lancez K-Means pour différents k (ex : 2 à 8) et calculez le \*\*silhouette score\*\*.  
- Choisissez un k (par exemple via le meilleur silhouette) et entraînez le modèle final.  
- Visualisez les clusters en 2D (espace PCA) et affichez les centres.

4) Clustering avec DBSCAN

- Essayez DBSCAN (par ex. eps=0.6 puis ajustez) et `min\_samples=5`.  
- Visualisez le résultat (points bruits = étiquette -1) dans l’espace PCA.  
- Comparez avec K-Means : nombre de clusters, cohérence et bruit.

5) Discussion rapide

- Quels segments de clientèle pourriez-vous décrire ?  
- Avantages/limites de K-Means vs DBSCAN sur ces données.  
- Impact de la mise à l’échelle, de l’ajout/retrait de variables (ex : Gender).

## Questions à remettre (réponses courtes)

Q1. Donnez les dimensions du dataset retenu (après sélection des variables).

Q2. Quel k avez-vous retenu pour K-Means et pourquoi (silhouette) ? Donnez la valeur (≈).

Q3. Avec DBSCAN, combien de clusters (hors bruit) et quelle proportion de bruit ?

Q4. Donnez un court profil pour 2 clusters obtenus (ex : revenu/spending/âge).

Q5. En une ou deux phrases, comparez K-Means et DBSCAN pour ce cas.

## Astuces

- Standardiser avant le clustering améliore souvent les résultats.  
- Pour K-Means, utilisez le `silhouette\_score` pour comparer les k.  
- Pour DBSCAN, ajustez `eps` selon l’échelle des données (après standardisation).

Fichier généré le 2025-09-22 06:47:59