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

**Dataset : Iris (scikit-learn)**

## Contexte & objectifs

L’objectif de ce TP est de découvrir le clustering non supervisé à travers deux méthodes classiques : K-Means et DBSCAN. Vous travaillerez sur le dataset Iris (150 échantillons, 4 variables continues : longueur/largeur sépale et pétale). Le but est de :

• Charger et explorer le dataset Iris

• Projeter les données en 2D avec PCA pour visualiser les structures

• Appliquer K-Means et interpréter les clusters obtenus

• Appliquer DBSCAN et comparer les résultats à K-Means

• Discuter des avantages/limites de chaque approche

## Consignes pratiques

Durée : ~1h

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

Livrables : votre notebook (.ipynb) complété et commenté

## Étapes à suivre

1) Chargement des données

- Importez le dataset Iris depuis scikit-learn (sklearn.datasets.load\_iris).  
- Placez les variables dans un DataFrame pandas et affichez les 5 premières lignes.  
- Affichez la taille des données (shape) et des statistiques descriptives (describe).

2) Visualisation avec PCA (2 composantes)

- Appliquez une ACP (PCA) pour projeter les données en 2D.  
- Tracez un nuage de points (matplotlib) des deux composantes principales.  
- Commentez visuellement la structure apparente.

3) Clustering avec K-Means

- Appliquez K-Means avec k=3 (random\_state=42 pour la reproductibilité).  
- Affichez les centres (dans l’espace PCA) et visualisez les clusters.  
- Calculez le score silhouette sur les données d’origine et commentez brièvement.

4) Clustering avec DBSCAN

- Appliquez DBSCAN (essayez eps=0.6, min\_samples=5 puis ajustez si nécessaire).  
- Visualisez le résultat dans l’espace PCA (bruit = étiquette -1).  
- Comparez avec K-Means : nombre de clusters, points marqués comme bruit, cohérence visuelle.

5) Discussion

- Dans quels cas K-Means est-il pertinent ? Dans quels cas DBSCAN l’est-il davantage ?  
- Que change la mise à l’échelle des variables (StandardScaler) ?  
- Limites : choix de k (K-Means), sensibilité à eps/min\_samples (DBSCAN).

## Questions à remettre (réponses courtes)

Q1. Donnez la forme (shape) de X et les noms des variables.

Q2. Après PCA (2D), observez-vous des groupes visuels distincts ? Justifiez en une phrase.

Q3. Avec K-Means (k=3), indiquez le score silhouette obtenu (≈) et commentez.

Q4. Avec DBSCAN, indiquez le nombre de clusters (hors bruit) et la proportion de bruit.

Q5. En une ou deux phrases, comparez K-Means et DBSCAN sur ce dataset.

## Astuces

- Pensez à standardiser les données avant DBSCAN (souvent utile) et parfois avant K-Means.  
- Pour évaluer le nombre de clusters en K-Means : méthode du coude (optionnel) ou silhouette.  
- Fixez random\_state pour la reproductibilité.

Fichier généré le 2025-09-22 06:45:11