Voici les **étapes principales d'un projet de machine learning en apprentissage non supervisé** en utilisant l'algorithme **k-means** :

---

1. Définition du problème

- Quel est l'objectif ? (e.g. segmentation de clients, regroupement de produits similaires, détection de schémas)
- Quelles données sont disponibles ?
- Pourquoi utiliser **l'apprentissage non supervisé** et particulièrement **k-means** ?

---

2. Collecte des données

- Obtenir des données structurées : CSV, base de données, API, etc.
- Exemple : caractéristiques clients (âge, revenus, fréquence d'achat), caractéristiques produits (poids, prix, ventes…).

---

3. Préparation et nettoyage des données

- **Nettoyage** : suppression des doublons, gestion des valeurs manquantes, correction d'incohérences.
- **Encodage** des variables catégorielles si nécessaire.
- **Normalisation/standardisation** : K-means est sensible à l'échelle des variables, donc il est **indispensable de standardiser** les données (e.g. avec `StandardScaler` de `sklearn`).

---

4. Exploration des données (EDA)

- Statistiques descriptives.
- Visualisation des distributions, corrélations, nuages de points, etc.
- Vérification de la structure potentielle des clusters (avec PCA ou t-SNE par exemple).

---

5. Détermination du nombre optimal de clusters (k)

Utiliser des méthodes comme :

- **Méthode du coude (Elbow method)** : tracer l'inertie intra-cluster (inertia) en fonction de `k`.
- **Silhouette score** : qualité de la séparation entre les clusters.

- **Gap statistic** (moins courant mais utile).

---

6. Application de l'algorithme k-means

- Initialisation de l'algorithme avec `k`.
- Entraînement : `kmeans.fit(X)`
- Récupération des **labels des clusters** : `kmeans.labels_`
- Récupération des **centroïdes** : `kmeans.cluster_centers_`

---

7. Évaluation du clustering

- Analyse de la **cohérence** des clusters : taille, distance, similarité intra-cluster.
- Visualisation :
  - 2D avec réduction de dimension (PCA, t-SNE).
  - Couleurs selon les clusters attribués.
- Calcul de **silhouette score** ou d'autres métriques non supervisées.

---

8. Interprétation des clusters

- Analyser les caractéristiques moyennes de chaque cluster.
- Donner un **profil** ou une **interprétation métier** aux groupes trouvés.
- Valider la pertinence avec des experts ou des connaissances terrain.

---

9. Utilisation des clusters

- Intégration dans un système ou une base de données.
- Exploitation pour du ciblage marketing, de la recommandation, de la gestion de stock, etc.

---

10. Itérations et amélioration

- Réajuster `k` si nécessaire.
- Ajouter ou supprimer des variables.
- Essayer d'autres algorithmes de clustering (DBSCAN, Agglomerative, etc.) pour comparaison.
