Université Joseph-Ky ZERBO/IFOAD
UFR : Sciences Informatiques Appliquées
Filière : **Sciences des Données**
Niveau : M1-S2
Enseignant : Dr Serge SONFACK SOUNCHIO

Burkina Faso
La Patrie ou la Mort, nous Vaincrons
Année académique 2024-2025
Ouagadougou le 13/11/2025

PROJET K-MEANS

Équipe de développement :

NOM & PRENOM	INE
OUDRAOGO Lassina	N00069220051
OUDRAOGO Rasmané	N00017820091
POUBERE Abdourazakou	N00145620141

Note	**Observation**

Lien du dépôt git : _https://github.com/POUBERE/TP_K-means.git

NB : Le fichier de rapport se trouve dans le dossier docs

TP 2 : Machine Learning Non Supervisé

Clustering avec K-means

Dataset : Estimation des Niveaux d'Obésité

Rapport Préliminaire

1. Introduction

Ce rapport présente le cadre de travail pour le TP 2 de Machine Learning non supervisé, portant sur l'algorithme de clustering K-means. L'objectif est de préparer une analyse complète avant l'implémentation pratique dans un notebook Jupyter.

Le clustering par K-means est une technique d'apprentissage non supervisé qui permet de regrouper des observations similaires en clusters homogènes, sans avoir besoin d'étiquettes prédéfinies. Cette approche est particulièrement adaptée pour découvrir des patterns naturels dans les données.

Ce document structuré justifie le choix du dataset, définit la problématique et les objectifs, décrit les données en détail, et identifie les outils nécessaires pour mener à bien ce projet.

2. Justification du Choix du Dataset

2.1. Présentation du Dataset

Le dataset choisi est le **"Obesity Levels Dataset"**, qui contient des informations sur les habitudes alimentaires, les caractéristiques physiques et le niveau d'obésité de 2111 individus. Ce dataset synthétique a été créé à partir de données réelles collectées auprès d'individus du Mexique, du Pérou et de la Colombie.

2.2. Raisons du Choix

Ce dataset est particulièrement adapté pour le clustering K-means pour plusieurs raisons :

- **Richesse des variables :** 17 variables incluant des données numériques (âge, taille, poids) et catégorielles (habitudes alimentaires, activité physique).
- **Pertinence thématique :** L'obésité est un enjeu de santé publique majeur, rendant l'analyse particulièrement significative.
- **Volume approprié :** 2111 observations offrent une taille suffisante pour un clustering robuste sans être trop volumineux.
- **Qualité des données :** Aucune valeur manquante, facilitant le preprocessing.
- **Potentiel de découverte :** La présence de 7 catégories d'obésité permet de valider la qualité du clustering obtenu.
- **Diversité des profils :** Distribution équilibrée entre les différents niveaux d'obésité.

3. Problématique

Comment identifier automatiquement des groupes homogènes d'individus partageant des caractéristiques similaires en termes de profil physique et de comportements liés à l'obésité, sans utiliser les étiquettes prédéfinies ?

Plus précisément, nous cherchons à répondre aux questions suivantes :

1. Peut-on découvrir des patterns naturels dans les données qui correspondent aux différents niveaux d'obésité ?
2. Quelles sont les caractéristiques principales qui différencient les différents groupes d'individus ?
3. Combien de clusters distincts émergent naturellement des données ?

4. Les clusters découverts par K-means correspondent-ils aux catégories d'obésité définies médicalement ?

4. Objectifs du Projet

4.1. Objectif Principal

Implémenter et entraîner un modèle de clustering K-means sur le dataset d'obésité pour identifier des groupes naturels d'individus partageant des profils similaires.

4.2. Objectifs Spécifiques

- **Préparation des données :** Nettoyer, encoder les variables catégorielles, normaliser les données numériques.
- **Analyse exploratoire :** Visualiser les distributions, identifier les corrélations, comprendre la structure des données.
- **Détermination du nombre optimal de clusters :** Utiliser la méthode du coude (Elbow Method) et le score de silhouette.
- **Entraînement du modèle :** Appliquer l'algorithme K-means avec différentes valeurs de K.
- **Évaluation et validation :** Mesurer la qualité du clustering avec des métriques appropriées (inertie, silhouette, indice de Davies-Bouldin).
- **Interprétation des résultats :** Analyser les caractéristiques de chaque cluster découvert et les comparer aux catégories originales.
- **Visualisation :** Créer des graphiques pour illustrer les clusters dans un espace réduit (PCA/t-SNE).

5. Description Détaillée des Données

5.1. Vue d'Ensemble

Nombre d'observations : 2111 individus

Nombre de variables : 17 variables (16 features + 1 variable cible)

Valeurs manquantes : Aucune (dataset complet)

5.2. Description des Variables

Variable	Type	Description
Gender	Catégorielle	Sexe de l'individu (Male/Female)
Age	Numérique	Âge en années (14-61)
Height	Numérique	Taille en mètres (1.45-1.98)
Weight	Numérique	Poids en kilogrammes
family_history_with_overweight	Catégorielle	Antécédents familiaux de surpoids (yes/no)
FAVC	Catégorielle	Consommation fréquente d'aliments hypercaloriques (yes/no)
FCVC	Numérique	Fréquence de consommation de légumes (0-3)
NCP	Numérique	Nombre de repas principaux par jour (1-4)
CAEC	Catégorielle	Consommation de nourriture entre les repas (no/Sometimes/Frequently/Always)
SMOKE	Catégorielle	Fume (yes/no)

Variable	Type	Description
CH2O	Numérique	Consommation d'eau quotidienne en litres (1-3)
SCC	Catégorielle	Surveillance de la consommation de calories (yes/no)
FAF	Numérique	Fréquence d'activité physique (0-3)
TUE	Numérique	Temps d'utilisation d'appareils technologiques en heures (0-2)
CALC	Catégorielle	Consommation d'alcool (no/Sometimes/Frequently/Always)
MTRANS	Catégorielle	Moyen de transport (Walking/Bike/Motorbike/Public_Transportation/Automobile)

5.3. Variable Cible

NObeyesdad : Niveau d'obésité (7 catégories)

- Insufficient_Weight : 272 individus (12.9%)
- Normal_Weight : 287 individus (13.6%)
- Overweight_Level_I : 290 individus (13.7%)
- Overweight_Level_II : 290 individus (13.7%)
- Obesity_Type_I : 351 individus (16.6%)
- Obesity_Type_II : 297 individus (14.1%)
- Obesity_Type_III : 324 individus (15.3%)

5.4. Statistiques Descriptives

Variables Numériques :

- **Âge :** moyenne = 24.3 ans, écart-type = 6.3 ans
- **Taille :** moyenne = 1.70 m, écart-type = 0.09 m
- **Activité physique :** moyenne = 1.0, écart-type = 0.85
- **Consommation d'eau :** moyenne = 2.0 litres/jour

6. Outils et Technologies Nécessaires

6.1. Environnement de Développement

- **Python 3.8+ :** Langage de programmation principal
- **Jupyter Notebook :** Pour le développement interactif et la documentation
- **Google Colab (optionnel) :** Alternative cloud pour l'exécution

6.2. Bibliothèques Python Essentielles

6.2.1. Manipulation et Analyse de Données

- **pandas :** Chargement et manipulation du dataset
- **numpy :** Opérations mathématiques et manipulation de arrays

6.2.2. Machine Learning

- **scikit-learn :** Implémentation de K-means et métriques d'évaluation

- KMeans : Algorithme de clustering

- StandardScaler : Normalisation des données

- LabelEncoder : Encodage des variables catégorielles

- PCA : Réduction de dimensionnalité pour la visualisation

- silhouette_score : Métrique d'évaluation

- davies_bouldin_score : Indice de séparation

6.2.3. Visualisation

- **matplotlib :** Création de graphiques de base
- **seaborn :** Visualisations statistiques avancées
- **plotly (optionnel) :** Visualisations interactives

6.3. Commandes d'Installation

Installation de toutes les bibliothèques nécessaires :

pip install pandas numpy scikit-learn matplotlib seaborn

7. Méthodologie Proposée

7.1. Phase 1 : Préparation des Données

5. **Chargement du dataset**
6. **Analyse exploratoire initiale** : statistiques descriptives, distributions, corrélations
7. **Traitement des variables catégorielles** : encodage (Label Encoding ou One-Hot Encoding)
8. **Normalisation des features numériques** : StandardScaler pour mettre toutes les variables sur la même échelle
9. **Sélection des features** : exclure la variable cible NObeyesdad pour l'apprentissage non supervisé

7.2. Phase 2 : Détermination du Nombre Optimal de Clusters

10. **Méthode du Coude (Elbow Method)**

- Tester K de 2 à 15 clusters

- Calculer l'inertie (somme des distances au carré) pour chaque K

- Tracer la courbe et identifier le coude

11. **Score de Silhouette**

- Calculer le score pour différentes valeurs de K

- Sélectionner le K avec le meilleur score

12. **Indice de Davies-Bouldin** : plus le score est bas, meilleure est la séparation

7.3. Phase 3 : Entraînement du Modèle K-means

13. Initialisation avec le nombre optimal de clusters identifié
14. Configuration des paramètres : n_init=10, max_iter=300, random_state=42
15. Entraînement du modèle
16. Attribution des labels de cluster à chaque observation

7.4. Phase 4 : Évaluation et Validation

17. **Métriques quantitatives**

- Inertie totale

- Score de silhouette moyen et par cluster

- Indice de Davies-Bouldin

18. **Comparaison avec les labels originaux**

- Matrice de confusion

- Adjusted Rand Index (ARI)

- Normalized Mutual Information (NMI)

7.5. Phase 5 : Analyse et Interprétation

19. Profiler chaque cluster : caractéristiques moyennes de chaque groupe
20. Identifier les features les plus discriminantes
21. Interpréter les clusters en termes de profils d'obésité
22. Comparer avec les catégories médicales d'obésité

7.6. Phase 6 : Visualisation

23. **Réduction de dimensionnalité** : PCA ou t-SNE pour visualiser en 2D/3D
24. **Scatter plots** : représentation des clusters dans l'espace réduit
25. **Heatmaps** : corrélations et caractéristiques moyennes par cluster
26. **Box plots** : distribution des variables numériques par cluster
27. **Graphiques de silhouette** : qualité de chaque cluster individuellement

8. Résultats Attendus

À l'issue de ce projet, nous devrions obtenir :

- **Un modèle K-means entraîné et optimisé** capable de segmenter automatiquement les individus
- **Des clusters cohérents** regroupant des individus avec des profils similaires
- **Une compréhension approfondie** des facteurs associés à différents niveaux d'obésité
- **Des visualisations claires** illustrant la structure des données
- **Une validation** de la pertinence du clustering par rapport aux catégories médicales
- **Des insights actionnables** pour identifier les comportements à risque

9. Limitations et Défis Potentiels

- **Choix du K :** La détermination du nombre optimal de clusters peut être subjective
- **Sensibilité à l'initialisation :** K-means peut converger vers des optima locaux
- **Forme des clusters :** K-means assume des clusters sphériques, ce qui peut ne pas correspondre à la réalité
- **Variables catégorielles :** L'encodage peut influencer les résultats du clustering
- **Dataset synthétique :** Les résultats pourraient différer sur des données réelles

10. Conclusion

Ce rapport a présenté le cadre complet pour le TP 2 de Machine Learning non supervisé avec l'algorithme K-means appliqué au dataset d'obésité. Nous avons justifié le choix de ce dataset riche et pertinent, défini une problématique claire, établi des objectifs spécifiques et mesurables, et décrit en détail les données disponibles.

La méthodologie proposée suit une approche rigoureuse en six phases, de la préparation des données à la visualisation finale, en passant par l'optimisation du nombre de clusters et l'évaluation quantitative. Les outils identifiés (Python, scikit-learn, pandas, matplotlib) constituent un écosystème complet pour mener à bien ce projet.

L'application de K-means sur ce dataset permettra non seulement de comprendre les mécanismes de l'apprentissage non supervisé, mais aussi d'obtenir des insights concrets sur les patterns de comportements liés à l'obésité. Les résultats de ce clustering pourront être comparés aux catégories médicales établies, validant ainsi l'efficacité de l'approche.

La prochaine étape consistera à implémenter cette méthodologie dans un notebook Jupyter, en suivant le plan détaillé et en documentant soigneusement chaque étape du processus. Ce travail préparatoire solide garantit une exécution efficace et des résultats de qualité.

Annexe : Structure du Notebook à Implémenter

Le notebook Jupyter devra suivre la structure suivante :

28. **Introduction et Objectifs**
29. **Importation des Bibliothèques**
30. **Chargement et Exploration des Données**
31. **Prétraitement des Données**

- Encodage des variables catégorielles

- Normalisation

32. **Détermination du Nombre Optimal de Clusters**

- Méthode du coude

- Score de silhouette

33. **Entraînement du Modèle K-means**
34. **Évaluation du Modèle**
35. **Analyse et Interprétation des Clusters**
36. **Visualisations**

Conclusions et Perspec