USTHB - Faculty of Computer Science

Mini Project

Fondements de l'IA - 3e année Software Engineering

Unsupervised Learning with K-Means Clustering

Importante note :

1. This work should be done in **pairs**
2. It should be done in the **Python language**
3. The code must **run correctly**; otherwise, the points will not be granted.
4. It must be submitted **before January 25, 2026**.

Problem context and objective:

The goal of this project is to apply K-Means clustering to the Heart Disease dataset to identify hidden patterns and patient groups without relying on labeled outcomes.

The Heart Disease dataset is a real-world medical dataset originally collected to investigate factors associated with heart disease in patients. The dataset contains clinical and demographic information about patients, such as age, sex, blood pressure, cholesterol levels, and results of medical tests.
It includes both numerical and categorical features.

Typical features in the dataset include:

1. Age: Age of the patient
2. Sex: Gender of the patient
3. Chest pain type (cp): Type of chest pain experienced
4. Resting blood pressure (trestbps)
5. Cholesterol level (chol)
6. Fasting blood sugar (fbs)
7. Resting ECG results (restecg)
8. Maximum heart rate achieved (thalach)
9. Exercise-induced angina (exang)
10. Oldpeak: ST depression induced by exercise
11. Slope: Slope of the peak exercise ST segment
12. Number of major vessels (ca)
13. Thal: Thalassemia type

Using K-Means, you will attempt to group patients into clusters with similar clinical profiles.

Instructions :

Part 1 – Data Exploration (3.5 points)

1. Download the Heart Disease dataset and load it into your code using Pandas. **(0.25pt)**

2. Display: **(0.5pt)**

- The first rows of the dataset
- Dataset shape (number of rows and columns)
- Column names and data types

3. Remove the target column **(0.25pt)**

4. Compute basic statistics (mean, standard deviation, minimum, maximum) **(0.5pt)**

5. Analyze feature distributions and visualize them using: **(0.5pt)**

- Histograms
- Box plots
- Scatter plots (for selected features)

6. Identify missing values and remove them if found **(0.25pt)**

7. Remove irrelevant columns (e.g., patient identifiers, if present) **(0.25pt)**

8. Check for duplicates and remove them **(0.25pt)**

9. Encode categorical variables (Use the One Hot Encoding) **(0.25pt)**

10. Select the features to be used for clustering **(0.25pt)**

11. Normalize or standardize numerical features **(0.25pt)**

Part 2 – Model Training (3 points)

1. Split the dataset into a training set (80%) and a test set(20%), and explain why splitting can still be useful in unsupervised learning. **(0.5pt)**

2. Apply the Elbow Method and choose an appropriate number of clusters **k**, explain how. **(0.75pt)**

3. Import the K-Means algorithm from sklearn and train it using the training set. **(0.75pt)**

4. Obtain cluster labels for both training and test data. **(0.5pt)**

5. Display the cluster centers. **(0.5pt)**

Part 3 – Evaluation (3.5 points)

1. Evaluate the clustering using appropriate metrics: **(0.75pt)**

    - Inertia
    - Silhouette Score
    - Davies–Bouldin Index

2. Compare metric values for different choices of k. **(0.75pt)**

3. Interpret what these metrics say about cluster quality. **(1pt)**

4. Visualize the clustering results using matplotlib plots, color data points according to their assigned cluster, and visualize the cluster centroids. **(1pt)**

__Deliverables :__

You must submit **one ZIP folder** named in the form *Nom1_Prénom1_Nom2_prénom2_Groupe_x* containing:

1. Report (PDF) that includes the dataset description, an explanation of the implementation of the 3 parts, and the interpretation of the clustering results.

2. Python files (.py or .ipynb), well-commented and executable with clear separation between steps.