Unsupervised Learning with K-Means Clustering

Heart Disease Dataset Analysis

SURNAME: Louni	**SURNAME:** Rouane
NAME: Mohammed Said	**NAME:** Mohamed Walid
NUMBER: 232331499716	**NUMBER:** 232331665310
GROUP: 1	**GROUP:** 1

29th January 2026

Contents

1 Introduction

This report presents an implementation of K-Means clustering applied to the Heart Disease dataset. The objective is to identify hidden patterns and patient groups based on clinical and demographic features without using labelled outcomes. The analysis follows a systematic approach: data exploration and preprocessing, model training using K-Means, and evaluation of clustering quality using standard metrics.

2 Dataset Description

2.1 Overview

The Heart Disease dataset is a real-world medical dataset originally collected to investigate factors associated with heart disease in patients. It contains clinical and demographic information for 1,025 patients with 14 features.

2.2 Features

The dataset includes the following attributes:

Numerical Features:

- `age`: Age of the patient (29–77 years)
- `trestbps`: Resting blood pressure (94–200 mm Hg)
- `chol`: Cholesterol level (126–564 mg/dl)
- `thalach`: Maximum heart rate achieved (71–202 bpm)
- `oldpeak`: ST depression induced by exercise (0.0–6.2)

Categorical Features:

- `sex`: Gender (0 = female, 1 = male)
- `cp`: Chest pain type (0–3)
- `fbs`: Fasting blood sugar $>$ 120 mg/dl (0 = no, 1 = yes)
- `restecg`: Resting ECG results (0–2)
- `exang`: Exercise-induced angina (0 = no, 1 = yes)
- `slope`: Slope of the peak exercise ST segment (0–2)
- `ca`: Number of major vessels (0–4)
- `thal`: Thalassaemia type (0–3)
- `target`: Heart disease presence (0 = no, 1 = yes) − removed for clustering

Table 1: Basic statistics of numerical features

Feature	Mean	Std Dev	Min	Max
age	54.43	9.07	29	77
trestbps	131.61	17.52	94	200
chol	246.00	51.59	126	564
thalach	149.11	23.01	71	202
oldpeak	1.07	1.18	0.0	6.2

2.3 Dataset Statistics

Key statistics computed from the dataset:

The dataset is complete with no missing values and no patient identifiers requiring removal.

3 Part 1: Data Exploration and Preprocessing

3.1 Data Loading and Initial Inspection

The dataset was loaded using Pandas with cross-platform compatibility using `pathlib.Path`:

```python
from pathlib import Path
DATA_DIR = Path(__file__).parent
HEART_CSV = DATA_DIR / 'heart.csv'
dataset = pd.read_csv(HEART_CSV)
```

Initial inspection revealed 1,025 rows and 14 columns with mixed data types (integers and floats). The target column was removed as required for unsupervised learning.

3.2 Data Cleaning

Missing Values: The dataset contained zero missing values across all features, requiring no imputation.

Duplicates: Duplicate removal was performed using `pandas.drop_duplicates()`, though the dataset contained no exact duplicates.

3.3 Feature Engineering

3.3.1 Categorical Encoding

Five categorical features (`cp`, `restecg`, `slope`, `ca`, `thal`) were encoded using one-hot encoding, expanding the feature space from 13 to 27 features:

```python
categorical_cols = ['cp', 'restecg', 'slope', 'ca', 'thal']
dt = pd.get_dummies(dt, columns=categorical_cols, dtype=int)
```

3.3.2 Feature Selection

Feature importance analysis was performed using the `featclus` library with time-series shifts to identify the most relevant features for clustering. Based on importance scores, 11 features were selected:

- Original features: `age`, `sex`, `trestbps`, `chol`, `thalach`, `exang`, `oldpeak`
- One-hot encoded: `cp_0`, `cp_1`, `cp_2`, `cp_3`

The top features by importance were chest pain types (`cp_0`: 0.176) and resting ECG results, indicating their significance for patient grouping.

3.3.3 Feature Scaling

Numerical features were standardised using `StandardScaler` to ensure equal weighting in distance calculations:

```
numerical_cols = [’age’, ’trestbps’, ’chol’, ’thalach’, ’oldpeak’]
scaler = StandardScaler()
dt_scaled[numerical_cols] = scaler.fit_transform(dt_scaled[numerical_cols])
```

This transformation centres each feature at mean 0 with standard deviation 1, preventing features with larger ranges (e.g., cholesterol) from dominating the clustering.

3.4 Data Visualisation

Three visualisation types were generated:

- **Histograms**: Revealed distributions of all features, showing age follows a roughly normal distribution centred at 54 years, while cholesterol shows a right-skewed distribution.
- **Box plots**: Identified potential outliers, particularly in cholesterol levels (maximum 564 mg/dl).
- **Scatter plots**: Age vs. cholesterol and age vs. maximum heart rate showed no strong linear relationships, suggesting clusters may not separate along simple axes.

4 Part 2: Model Training

4.1 Train-Test Split

The dataset was split into training (80%, 241 samples) and test sets (20%, 61 samples):

```
X_train, X_test = train_test_split(dt_scaled, test_size=0.2, random_state=42)
```

Rationale for splitting in unsupervised learning:
While clustering is unsupervised, train-test splitting serves three purposes:

1. **Cluster stability validation**: Tests whether clusters are consistent and reproducible on unseen data.
2. **Generalisation assessment**: Ensures the model does not overfit to training data patterns.
3. **Metric comparison**: Allows evaluation of clustering quality on both sets to detect instability.

4.2 Elbow Method for Optimal k

The Elbow Method was applied to determine the optimal number of clusters by plotting within-cluster sum of squares (inertia) against k:

```
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, init="random", n_init=10, random_state=1)
    kmeans.fit(X_train)
    sse.append(kmeans.inertia_)
```

Selection of k=3:

The elbow curve showed a clear inflection point at $k = 3$, where the rate of decrease in inertia significantly diminishes. This indicates that:

- Adding more clusters beyond 3 provides diminishing returns in variance reduction
- $k = 3$ balances model complexity with clustering quality
- Three clusters offer medically interpretable risk groups (low, medium, high risk)

4.3 K-Means Training

The K-Means algorithm was trained using scikit-learn with the following parameters:

```
kmeans = KMeans(init="random", n_clusters=3, n_init=10, random_state=1)
kmeans.fit(X_train)
```

Parameters:

- `n_clusters=3`: Based on elbow method
- `init="random"`: Random initialisation of centroids
- `n_init=10`: Algorithm runs 10 times with different seeds
- `random_state=1`: Ensures reproducibility

4.4 Cluster Assignment

Cluster labels were obtained for both training and test sets:

Training set distribution:

- Cluster 0: 104 samples (43.2%)
- Cluster 1: 73 samples (30.3%)
- Cluster 2: 64 samples (26.6%)

Test set distribution:

- Cluster 0: 24 samples (39.3%)
- Cluster 1: 22 samples (36.1%)
- Cluster 2: 15 samples (24.6%)

The similar distributions between training and test sets indicate stable clustering.

4.5 Cluster Centres

The three cluster centroids in standardised feature space are:

Table 2: Cluster centres (key features shown)

Feature	Cluster 0	Cluster 1	Cluster 2
age	-0.823	0.475	0.737
sex	0.712	0.808	0.484
trestbps	0.260	0.836	0.328
chol	0.240	0.014	0.203
thalach	0.442	0.123	0.328
oldpeak	-0.488	0.850	-0.300

Cluster interpretation:

- **Cluster 0**: Younger patients with higher heart rates, lower ST depression
- **Cluster 1**: Middle-aged patients with elevated blood pressure and high ST depression
- **Cluster 2**: Older patients with moderate clinical indicators

5 Part 3: Evaluation and Results

5.1 Clustering Quality Metrics

Three metrics were used to evaluate clustering quality:

5.1.1 Inertia

Value: 1023.37

Inertia measures within-cluster sum of squared distances. A value of 1023.37 indicates moderate spread around cluster centres. While lower values are preferred, this must be balanced against the number of clusters.

5.1.2 Silhouette Score

Value: 0.179

The silhouette score ranges from -1 to 1, measuring how similar points are to their own cluster compared to other clusters. A score of 0.179 indicates:

- Clusters have some overlap
- Points are closer to their own cluster than others, but not strongly separated
- The clustering structure is weak but meaningful

5.1.3 Davies-Bouldin Index

Value: 1.845

The Davies-Bouldin index measures the average similarity ratio of each cluster with its most similar cluster. Lower values indicate better separation. A value of 1.845 suggests:

- Clusters are not well-separated
- Some cluster pairs are relatively close
- This is typical for medical data with continuous features

5.2 Comparison Across Different k Values

Table 3: Clustering metrics for different k values

k	Silhouette Score	Davies-Bouldin Index
2	0.199	1.835
3	0.179	1.845
4	0.147	1.831
5	0.119	1.955

Analysis:

- $k = 2$ achieves the highest silhouette score (0.199) and lowest Davies-Bouldin index (1.835), indicating the most stable clusters by metrics
- $k = 3$ scores slightly lower but provides more granular patient groupings
- $k = 4$ and $k = 5$ show deteriorating metrics, indicating over-segmentation

Decision: Despite $k = 2$ having superior metrics, $k = 3$ was chosen because:

1. The elbow method clearly indicated $k = 3$
2. Three clusters provide clinically meaningful risk stratification (low/medium/high)
3. The metric difference from $k = 2$ is small (0.02 in silhouette score)
4. Medical interpretability is prioritised over marginal metric improvements

5.3 Interpretation of Cluster Quality

Overall Assessment:

The clustering reveals moderate but meaningful patient groupings. The low silhouette score (0.179) and elevated Davies-Bouldin index (1.845) indicate:

1. **Overlapping clusters**: Heart disease risk exists on a continuum rather than in discrete categories. Patients near cluster boundaries share characteristics with multiple groups.

2. **Complex relationships**: Medical data rarely exhibits clear geometric separation. The 11-dimensional feature space contains subtle patterns that K-Means, which uses Euclidean distance, may not fully capture.

3. **Feature interdependencies**: Clinical features interact in complex ways. For example, age influences multiple other features (blood pressure, heart rate), creating correlation that reduces cluster separation.

Practical Significance:
Despite modest metrics, the clustering provides value:

- Identifies distinct patient profiles for targeted treatment strategies
- Reveals that a single risk score may be insufficient; multidimensional grouping is necessary
- Suggests that more sophisticated clustering methods (e.g., DBSCAN, hierarchical clustering) might improve separation

5.4 Visualisation

Figure 1 shows the clustering results in a 2D projection using age and cholesterol:

Figure 1: K-Means clustering visualisation: patients coloured by cluster assignment, with centroids marked as black X symbols. The 2D projection shows moderate cluster separation consistent with the silhouette score of 0.179.

The visualisation confirms metric findings: clusters overlap but maintain distinct centres, particularly in the age dimension where Cluster 0 (younger patients) separates from Clusters 1 and 2 (older patients).

6 Conclusion

This project successfully applied K-Means clustering to the Heart Disease dataset, identifying three patient groups with distinct clinical profiles. The implementation followed a rigorous methodology: comprehensive data preprocessing including one-hot encoding and standardisation, systematic cluster number selection via the elbow method, and multi-metric evaluation.

Key Findings:

1. The optimal configuration uses $k = 3$ clusters, balancing statistical metrics with medical interpretability.

2. Clustering metrics (silhouette: 0.179, Davies-Bouldin: 1.845) indicate moderate cluster quality, typical for continuous medical data.

3. Three distinct patient profiles emerged: younger with higher heart rates, middle-aged with elevated blood pressure, and older with moderate indicators.

4. Feature selection identified chest pain type and resting ECG as most important for patient grouping.

Limitations:

The relatively low silhouette score suggests K-Means may not be the optimal algorithm for this dataset. Future work could explore:

- Density-based clustering (DBSCAN) to handle overlapping regions
- Hierarchical clustering to reveal nested patient subgroups
- Feature engineering to extract non-linear combinations
- Dimensionality reduction (PCA, t-SNE) for improved visualisation

Despite these limitations, the clustering provides actionable insights for patient stratification and demonstrates the application of unsupervised learning to real-world medical data.