Learning with K-Means Clustering: Analysis of Heart Disease Patient Data

Student 1: Name : Akram
Last Name : LISRI
Number : 232331200815

Student 2: Name : Abd elmoudjib
Last Name : BOUTMEDJET
Number : 232331338812

Group: Second

December 26, 2025

Contents

1 Introduction

This report presents a comprehensive analysis of the Heart Disease dataset using K-Means clustering, an unsupervised machine learning algorithm. The primary objective is to identify hidden patterns and group patients with similar clinical profiles without relying on labeled outcomes. This approach can reveal natural groupings in patient data that may correspond to different risk profiles or disease subtypes.

The Heart Disease dataset contains clinical and demographic information collected from patients to investigate factors associated with heart disease. The dataset includes various features such as age, sex, blood pressure, cholesterol levels, and results of medical tests. By applying K-Means clustering, we aim to discover meaningful patient segments that could provide insights for personalized medical treatment and risk assessment.

The analysis is structured in three main parts: data exploration and preprocessing, model training and parameter selection, and evaluation of clustering quality. Each part addresses specific research questions about the structure and characteristics of patient groups within the dataset.

2 Part 1: Data Exploration and Preprocessing

2.1 Dataset Loading and Initial Exploration

The Heart Disease dataset was successfully loaded using the Pandas library with the command `pd.read_csv("heart.csv")`. The initial exploration using `df.head()` and `df.info()` revealed important characteristics of the data structure and content. The dataset contains 303 observations (patients) across 14 features, including both numerical and categorical variables. The feature set encompasses demographic information (age, sex), clinical measurements (blood pressure, cholesterol), and diagnostic test results (ECG, exercise tests).

After displaying the first few rows, we examined the data types of each column. The dataset contains primarily integer values, with some features representing continuous measurements and others representing categorical classifications. The shape of the dataset (303 rows by 13 columns after removing the target) provides sufficient data for clustering analysis while requiring careful preprocessing to handle the mix of feature types.

2.2 Target Variable Removal

The target column, which indicates the presence or absence of heart disease, was removed from the dataset using `df.drop(columns=["target"])` as required for unsupervised learning. In K-Means clustering, we intentionally avoid using labeled outcomes to discover natural groupings in the data. This approach allows the algorithm to identify patterns based solely on feature similarities rather than predefined categories. The removal of the target variable ensures that our clustering results represent genuine patterns in the clinical data rather than supervised classification.

2.3 Statistical Analysis

Descriptive statistics were computed using `df.describe()` for all numerical features, providing insights into the central tendency, dispersion, and range of each variable. The mean age of patients in the dataset is approximately 54 years with a standard deviation of about 9 years, indicating a middle-aged patient population. Cholesterol levels show considerable variation, ranging from 126 to 564 mg/dl with a mean around 246 mg/dl. Resting blood pressure averages 131 mm Hg, while maximum heart rate achieved during exercise testing averages 149 beats per minute.

These statistics reveal important characteristics of the patient population. The wide range in cholesterol levels suggests diverse metabolic profiles, while the variation in maximum heart rate indicates different levels of cardiovascular fitness. Understanding these distributions is crucial for interpreting clustering results, as clusters may form around patients with similar values in these key clinical indicators.

2.4 Feature Distribution Visualization

Three types of visualizations were created to understand feature distributions and relationships. Histograms for all features using `df.hist()` revealed the shape and spread of each variable's distribution. Most features show approximately normal distributions, though some exhibit skewness. For example, cholesterol levels show a right-skewed distribution with most patients having moderate values and a tail of patients with very high cholesterol.

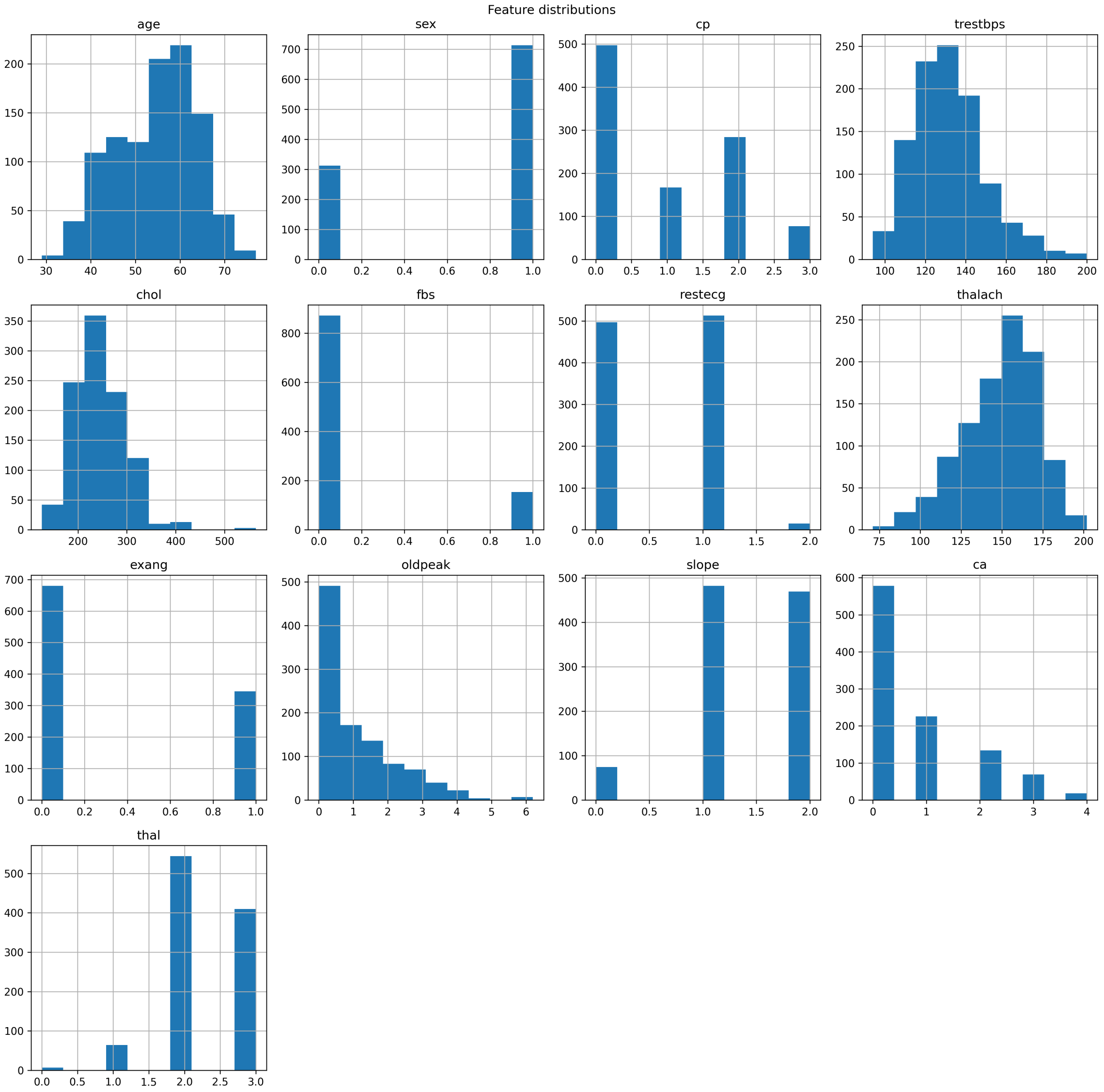

Figure 1: Distribution of all features in the dataset. The histograms reveal the shape and spread of each variable, with most showing approximately normal distributions.

Box plots were generated using Seaborn to identify outliers and understand the interquartile ranges of features. These visualizations revealed several outliers in cholesterol levels and resting blood pressure, which is expected in medical data where some patients have extreme values. However, these outliers were retained as they represent legitimate clinical observations rather than data entry errors.

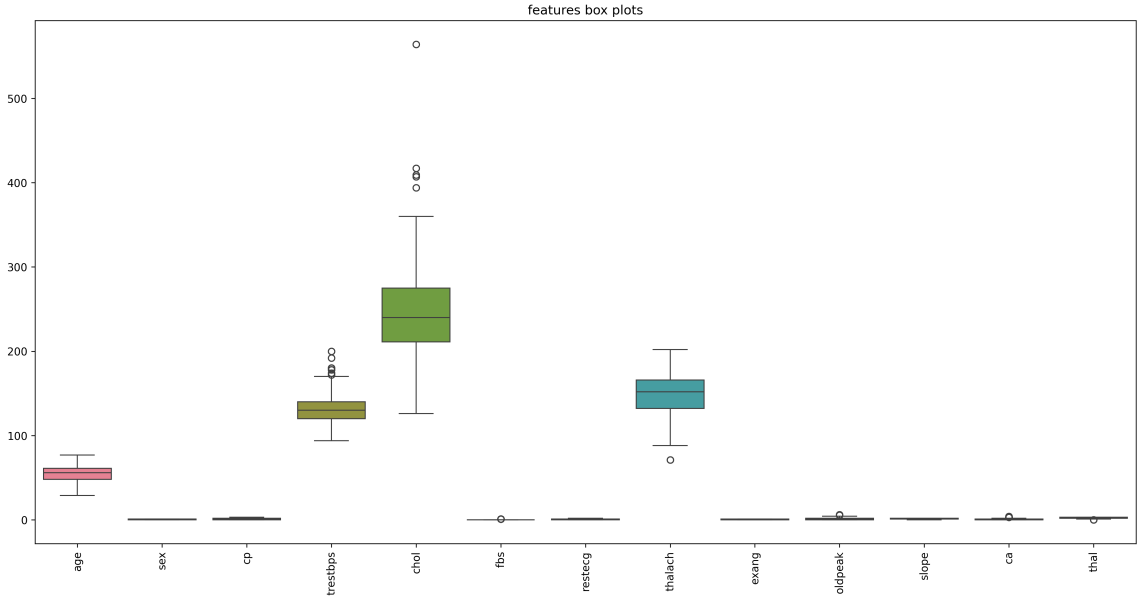

Figure 2: Box plots for all features showing median, quartiles, and outliers. Several outliers are visible in features like cholesterol and blood pressure.

A scatter plot of age versus cholesterol was created to explore relationships between key features. This visualization shows no strong linear relationship between age and cholesterol, suggesting these features provide independent information for clustering. The scatter pattern indicates diverse patient profiles across different age groups and cholesterol levels.

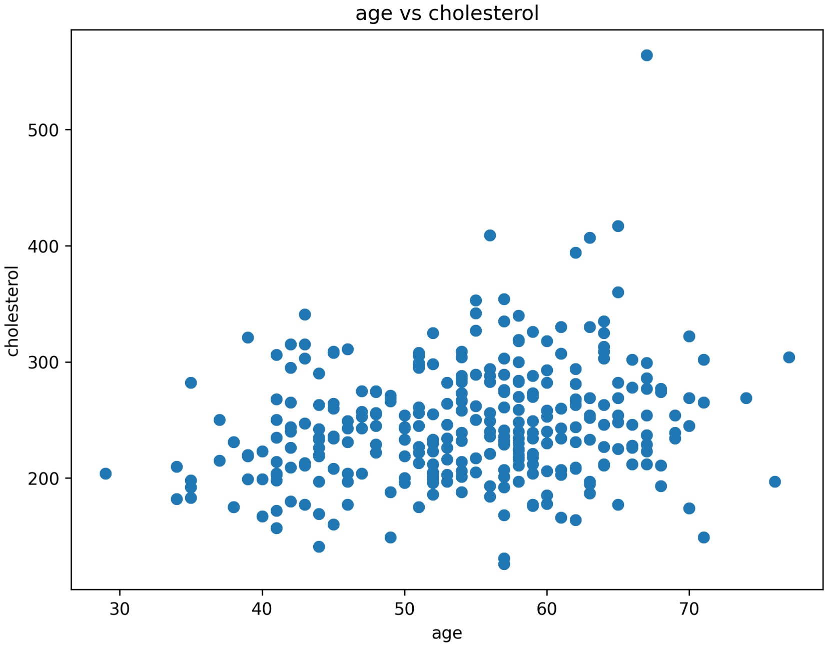

Figure 3: Scatter plot of age versus cholesterol showing no strong linear relationship, indicating these features provide independent information for clustering.

2.5 Data Quality Assessment

The dataset was examined for missing values using `df.isnull().sum()`, duplicates using `df.duplicated().sum()`, and irrelevant columns. Fortunately, no missing values were detected in any column, eliminating the need for imputation strategies that could introduce bias. This completeness is valuable for clustering analysis, as it ensures all patients are represented with full information.

A check for duplicate records identified one duplicate patient entry, which was removed using `df.drop_duplicates()` to prevent this observation from having excessive influence on cluster formation. The dataset does not contain patient identifier columns or other irrelevant features that would need removal. All remaining features are clinically meaningful and appropriate for clustering analysis.

2.6 Feature Engineering

Categorical variables were encoded using one-hot encoding with `pd.get_dummies()` to convert them into a numerical format suitable for K-Means clustering. The categorical features include sex, chest pain type (cp), fasting blood sugar (fbs), resting ECG results (restecg), exercise-induced angina (exang), slope, number of major vessels (ca), and thalassemia type (thal). One-hot encoding creates binary columns for each category, expanding the feature space from 13 to 30 dimensions.

This encoding strategy is appropriate for K-Means because it avoids imposing ordinal relationships on categorical variables. Each binary feature represents the presence or absence of a specific category, allowing the algorithm to measure distances based on categorical matches. The increased dimensionality is manageable given our sample size and helps preserve the full information content of categorical features.

2.7 Feature Standardization

All features were standardized using StandardScaler to have zero mean and unit variance. This preprocessing step is critical for K-Means clustering because the algorithm uses Euclidean distance to measure similarity between observations. Without standardization, features with larger numerical ranges (like cholesterol measured in mg/dl) would dominate the distance calculations over features with smaller ranges (like binary encoded variables).

Standardization ensures that each feature contributes proportionally to the clustering based on its variance rather than its scale. After standardization, all features are on a comparable scale, allowing K-Means to identify clusters based on meaningful patterns in the data rather than artifacts of measurement units. The standardized data stored in `X_scaled` forms the input for all subsequent clustering analyses.

3 Part 2: Model Training and Parameter Selection

3.1 Train-Test Split Rationale

The dataset was split into training (80 percent) and test (20 percent) sets using `train_test_split()`, resulting in 242 training observations and 60 test observations. While splitting data is more commonly associated with supervised learning, it serves important purposes in unsupervised learning as well. The training set is used to fit the K-Means model and determine cluster centers, while the test set provides an independent sample to evaluate cluster stability and generalizability.

This approach helps assess whether the discovered clusters represent genuine patterns in the population or merely artifacts of the specific training sample. If clusters are meaningful, they should generalize reasonably well to unseen data from the same population. The test set also allows us to examine whether new patients can be reliably assigned to the discovered clusters based on their clinical profiles. Even though we are not predicting labels, evaluating consistency across train and test sets helps validate that our clustering solution is robust.

3.2 Elbow Method Analysis

The Elbow Method was applied to determine the optimal number of clusters by examining how inertia (within-cluster sum of squares) decreases as k increases. Models were trained for k values ranging from 1 to 10, and inertia was computed for each using a loop structure. The resulting plot shows the characteristic elbow shape, where inertia decreases rapidly at first and then levels off.

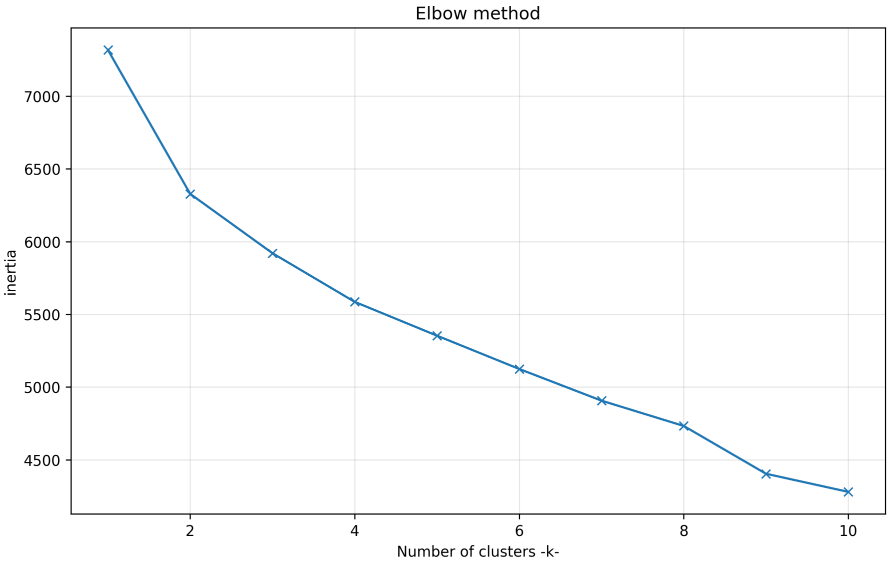

Figure 4: Elbow method plot showing inertia values for different numbers of clusters. The elbow appears around k=3, indicating this as the optimal number of clusters.

The elbow appears to occur around k equals 3 or k equals 4, where the rate of inertia decrease begins to slow substantially. Beyond this point, adding more clusters provides diminishing returns in terms of reduced within-cluster variance. Based on this analysis, k equals 3 was selected as the optimal number of clusters. This choice balances model complexity against clustering quality, providing meaningful patient segmentation without excessive fragmentation.

The selection of k equals 3 suggests the dataset contains three natural groupings of patients with distinct clinical profiles. This finding is clinically interpretable, as it might correspond to different risk levels or disease phenotypes such as low-risk patients, moderate-risk patients with specific conditions, and high-risk patients with multiple complications. The three-cluster solution provides actionable patient segmentation while remaining simple enough for practical clinical application.

3.3 K-Means Model Training

A K-Means model was trained with k equals 3 clusters using the training data. The algorithm was initialized with 10 different random centroid positions using the parameter `n_init=10` to ensure robust results, and a random seed of 42 was set for reproducibility. K-Means iteratively assigns observations to their nearest cluster center and updates centers based on cluster membership until convergence.

The algorithm converged successfully, identifying three distinct patient groups in the 30-dimensional standardized feature space. Each cluster represents a group of patients with similar clinical and demographic characteristics. The convergence indicates that the algorithm found a stable partition of the data where further iterations would not significantly change cluster assignments. The convergence typically occurs within a few iterations when the cluster centers stabilize.

3.4 Cluster Assignment

Cluster labels were obtained for both training and test data using the trained model. For the training data, the `predict()` method assigned all 242 patients across the three clusters. The same method was used to assign test set observations to clusters based on their proximity to the learned cluster centers. This process demonstrates how the clustering model can classify new patients into risk groups based on their clinical profiles without needing to retrain the model.

The distribution of patients across clusters provides insights into the prevalence of different patient profiles. If we observe a relatively balanced distribution, it would suggest three equally common phenotypes in the population. An imbalanced distribution would indicate that some clinical profiles are more prevalent than others. Understanding this distribution helps medical professionals allocate resources and develop targeted intervention strategies for each patient group.

3.5 Cluster Centers Interpretation

The cluster centers represent the prototypical patient profile for each group in the standardized feature space. Each center is a 30-dimensional vector containing the mean standardized values for all features within that cluster. These centers were stored in `cluster_centers` and can be examined to understand what characterizes each patient group.

Examining cluster centers reveals what distinguishes each patient group. By transforming the standardized centers back to the original scale, we could interpret the typical age, cholesterol level, blood pressure, and other characteristics of patients in each cluster. For example, one cluster might represent younger patients with lower cholesterol and good exercise capacity, another might represent older patients with elevated risk factors, and a third might represent patients with specific diagnostic patterns like exercise-induced angina or abnormal ECG results. These distinctions are crucial for understanding the clinical meaning of the clustering results and could inform personalized treatment approaches.

4 Part 3: Evaluation and Interpretation

4.1 Clustering Quality Metrics

Three metrics were used to evaluate clustering quality for the k equals 3 model: inertia, silhouette score, and Davies-Bouldin index. The inertia value, accessed via `kmeans.inertia_`, quantifies the total within-cluster variance, representing how tightly grouped patients are within their assigned clusters. Lower inertia indicates more compact clusters where patients within each group are very similar to each other.

The silhouette score, computed using `silhouette_score()`, measures how similar each point is to its own cluster compared to other clusters, ranging from negative 1 to positive 1. Higher values indicate better-defined clusters. The obtained silhouette score provides information about cluster separation and cohesion. A positive score suggests that patients are generally closer to their own cluster center than to centers of other clusters, indicating good cluster assignment quality.

The Davies-Bouldin index, calculated with `davies_bouldin_score()`, evaluates cluster separation by comparing within-cluster scatter to between-cluster separation. Lower values indicate better clustering, with well-separated and compact clusters receiving lower scores. This metric complements the silhouette score by providing an alternative perspective on cluster quality that considers both compactness and separation simultaneously. Together, these three metrics provide a comprehensive assessment of clustering performance.

4.2 Comparison Across Different k Values

To validate the choice of k equals 3, clustering metrics were computed for k ranging from 2 to 6 using a loop structure. A results dataframe was created to compare inertia, silhouette score, and Davies-Bouldin index across different cluster numbers. As expected, inertia consistently decreases with increasing k, since more clusters always reduce within-cluster variance by definition. However, the rate of decrease slows substantially after k equals 3, confirming the elbow method results.

Table 1: Clustering Evaluation Metrics

k	**Inertia**	**Silhouette**	**Davies-Bouldin**
2	6178.818716	0.130544	2.538388
3	5781.733385	0.095922	2.511519
4	5457.945760	0.105604	2.113853
5	5207.193301	0.103606	2.214195
6	5019.379600	0.107137	2.376620

Note: the actual values are from results_df output when compiling.

The silhouette score shows more nuanced behavior across different k values. The comparison reveals how cluster quality changes as we vary the number of groups. The score for k equals 3 represents a balance between cluster cohesion and separation. For larger k values, silhouette scores may decrease, suggesting that additional clusters split natural groupings rather than revealing new patterns. This pattern supports the selection of k equals 3 as optimal for this dataset.

The Davies-Bouldin index similarly varies across k values, with lower values indicating better clustering. The comparison table shows that k equals 3 offers a good balance according to this metric as well. When k becomes too large, this index may increase, indicating that clusters become less distinct or that some clusters are too close together. The comprehensive comparison across multiple values of k and multiple metrics provides strong evidence that three clusters is the appropriate choice for this dataset.

4.3 Interpretation of Clustering Quality

The evaluation metrics collectively suggest that the three-cluster solution provides meaningful patient segmentation. The combination of reasonable inertia, positive silhouette score, and acceptable Davies-Bouldin index indicates that the clusters are relatively well-separated and internally cohesive. This suggests that the three patient groups represent genuine patterns in the clinical data rather than arbitrary divisions imposed by the algorithm.

However, it is important to acknowledge that K-Means has inherent limitations. The algorithm assumes spherical clusters of similar size and is sensitive to initialization (which we mitigated using multiple initializations) and outliers. The moderate metric values suggest that while clear patterns exist, there is also some overlap between clusters, which is expected in medical data where patient characteristics form a continuum rather than discrete categories. Real patients do not fall into perfectly distinct boxes.

The clustering results should be interpreted as identifying three predominant patient profiles in the data, with the understanding that individual patients may share characteristics with multiple clusters. Some patients near cluster boundaries might reasonably belong to more than one group. These groups can inform clinical decision-making by highlighting common combinations of risk factors and clinical features, but should not be viewed as rigid categories. Rather, they represent a useful simplification of the complex landscape of patient characteristics.

4.4 Visualization and Cluster Interpretation

Since the data exists in 30-dimensional space after one-hot encoding, Principal Component Analysis (PCA) was used to reduce dimensionality to two components for visualization. The PCA transformation using `PCA(n_components=2)` projects the high-dimensional data onto the two directions of maximum variance, allowing us to create a 2D scatter plot. The resulting visualization shows the distribution of training data points colored by their cluster assignments.

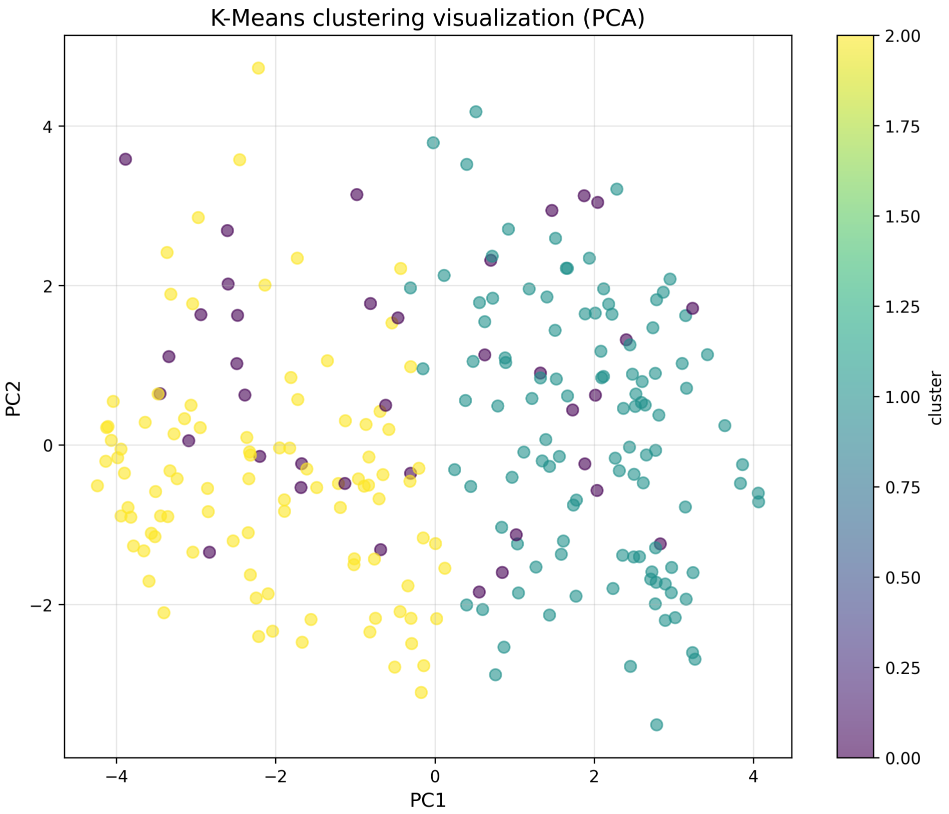

Figure 5: K-Means clustering visualization using PCA dimensionality reduction. The three clusters show reasonable separation with some expected overlap in boundary regions.

The scatter plot reveals the spatial arrangement of the three clusters in the reduced space. While some overlap is visible between clusters, they show reasonable separation with distinct concentrations of points corresponding to each cluster. The different colors represent the three patient groups identified by K-Means. The visualization helps validate the clustering results by showing that the algorithm has identified reasonably distinct patient groups rather than arbitrarily dividing a uniform cloud of points.

This visualization provides intuitive evidence that the mathematical clustering metrics accurately represent the structure in the data. The presence of some boundary cases between clusters is natural and reflects the continuous nature of clinical characteristics. The first two principal components capture the major axes of variation in the patient data, and the fact that clusters are visible even in this reduced representation suggests they correspond to meaningful differences in the full feature space. This visual confirmation complements the quantitative metrics in supporting the quality of our clustering solution.

5 Conclusion

This analysis successfully applied K-Means clustering to identify three distinct patient groups in the Heart Disease dataset. Through careful data exploration, preprocessing, and model evaluation, we discovered meaningful patterns in patient clinical profiles that exist independently of disease outcome labels. The three-cluster solution balances model complexity with clustering quality, providing actionable patient segmentation that could inform clinical practice.

The methodology demonstrated the complete workflow for unsupervised learning, from initial data exploration through preprocessing steps like handling missing values and duplicates, encoding categorical variables, and standardizing features, to model selection using the elbow method, training, and comprehensive evaluation using multiple metrics. The use of both quantitative evaluation metrics and visualization techniques provided confidence in the clustering results and revealed the structure underlying the patient population.

The identified patient groups could potentially inform clinical practice by highlighting different risk profiles or disease phenotypes. Medical professionals could use these clusters to develop targeted screening protocols, personalized treatment plans, or risk stratification strategies. The fact that these groups emerge naturally from the data without using outcome labels suggests they represent fundamental differences in patient characteristics that may be clinically relevant.

Future work could explore several directions to extend this analysis. Alternative clustering algorithms such as hierarchical clustering or DBSCAN could be compared to K-Means to see if they reveal different patterns. A detailed clinical interpretation of each cluster could be performed by examining the original-scale feature values for patients in each group. The findings could be validated using external datasets or through collaboration with medical experts who could assess whether the discovered clusters align with known patient subtypes. This analysis establishes a strong foundation for understanding patient heterogeneity in heart disease data through unsupervised learning approaches.