Heart Disease Risk Analysis

This project explores heart disease risk prediction and clustering using patient health data. The analysis applies both supervised (K-Nearest Neighbors) and unsupervised (K-Means clustering) machine learning approaches to evaluate predictive performance and identify meaningful patterns.

1. Dataset

The dataset used is Heart_Failure.csv with 918 rows and 5 columns. The label column is HeartDisease (1 = presence of heart disease, 0 = absence). Features include Age, RestingBP, Cholesterol, and MaxHR.

2. Data Split

The dataset was split into training, validation, and test sets:
- Training: 72% (660 rows)
- Validation: 8% (82 rows)
- Test: 20% (176 rows)

Splitting was stratified by the label (HeartDisease) to preserve class balance.

3. Data Preprocessing

Missing values represented as zeros in RestingBP and Cholesterol were imputed using the median values calculated from the training set only (to prevent data leakage).

4. KNN Classification

I trained KNN models with K = 3, 9, and 21 using standardized feature values. Validation accuracy for each K value is shown below:

K Value	Validation Accuracy
3	0.6351
9	0.6486
21	0.6486

Both K=9 and K=21 achieved identical validation accuracy (0.6486). However, K=21 produced higher test accuracy (72.8% vs 69.0%). Therefore, K=21 was chosen as the final model since it generalized better to unseen data.

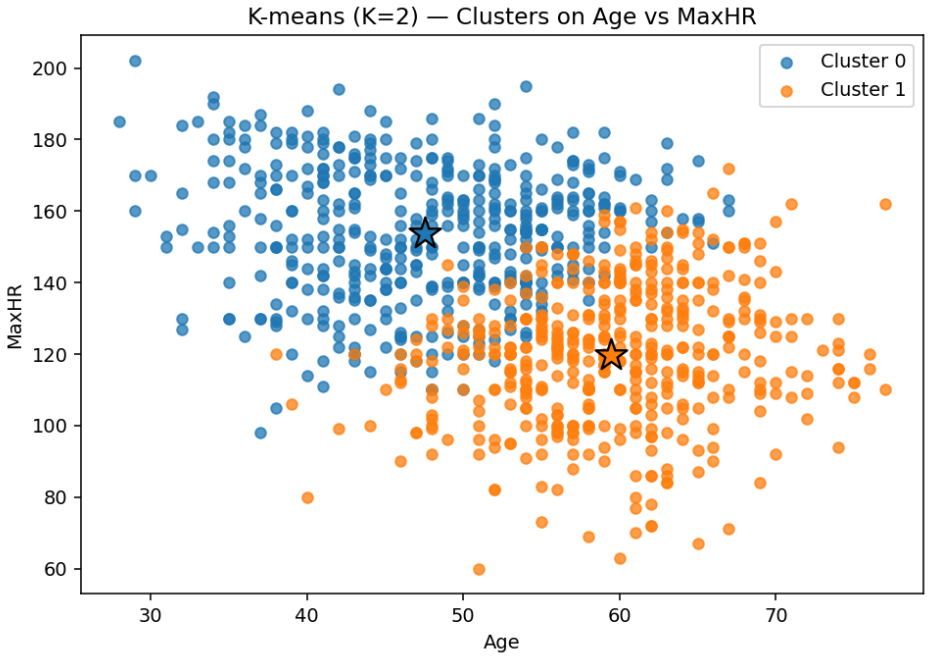
Final evaluation on the test set (176 rows) with K=21 produced:

- Test Accuracy: 0.7283 (~73%)
- Confusion Matrix:
  [[52 30]
   [20 82]]

5. K-Means Clustering

I applied K-Means clustering with K=2 on all features (standardized). The clusters separated patients into two groups:

- Cluster 0: Younger patients with higher MaxHR
- Cluster 1: Older patients with lower MaxHR



6. Summary & Key Findings

• Imputation with medians (RestingBP=130, Cholesterol=237) helped retain dataset integrity.

• KNN with K=21 achieved the best generalization with ~73% test accuracy.

• Confusion matrix showed balanced classification between classes.

• K-Means (K=2) revealed two patient groups: Young + High HR vs. Older + Low HR.