# HW1 Results — Heart Disease (KNN & K-means)

## 1) Dataset & Preprocessing

Rows × Columns: 918 × 5

Features: Age, RestingBP, Cholesterol, MaxHR

Label: HeartDisease (0 = No disease, 1 = Disease)

Cleaning: Replaced RestingBP=0 and Cholesterol=0 with the median of non-zero values computed from the TRAIN split (to avoid leakage).

Train medians used → RestingBP: 130.0, Cholesterol: 237.0

## 2) Splits

Train: 660 rows | Validation: 74 rows | Test: 184 rows

Stratified by label, random\_state=42

## 3) KNN (with StandardScaler)

We trained KNN with K = 3, 9, 21 on the training set and evaluated validation accuracy:

K=3 → Validation Accuracy = 0.6351

K=9 → Validation Accuracy = 0.6486

K=21 → Validation Accuracy = 0.6486

Both K=9 and K=21 achieved the same validation accuracy (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.

Chosen K: 21 (acc=0.6486)

## 4) Final Test Accuracy

Retrained KNN with K=21 on Train+Val and evaluated on Test.

Test Accuracy: 0.7283

Confusion Matrix (rows=actual 0/1, cols=predicted 0/1):

[[52 30]  
 [20 82]]

## 5) K-means (K=2) on All Features

We scaled features then ran K-means with 2 clusters. The plot shows clusters on Age vs MaxHR, with black stars marking cluster means.

