Machine Learning Analysis of Framingham Heart Disease Dataset

Mostafa Badawe, Mariam Yaser, Mohamed Maher, Mohamed Azoz, Mariam Melad

December 22, 2025

1 Introduction

This project focuses on analyzing the **Framingham Heart Disease Dataset**, a widely used medical dataset for studying cardiovascular risk factors and predicting the likelihood of developing coronary heart disease (CHD). The dataset contains clinical, demographic, and behavioral attributes collected as part of the Framingham Heart Study.

The dataset consists of **4,240 patient records** and **16 features**, including age, gender, smoking habits, blood pressure measurements, cholesterol levels, glucose levels, and diabetes indicators.

The target variable, **TenYearCHD**, is a binary label indicating whether an individual is expected to develop coronary heart disease within a ten-year period:

- 0 → No coronary heart disease
- 1 → Coronary heart disease within 10 years

The primary objective of this project is to perform thorough **data preprocessing and exploratory data analysis**, address data quality issues, and prepare the dataset for machine learning classification and regression models aimed at predicting cardiovascular risk.

2 Data Quality Issues

2.1 Missing Values

Initial inspection revealed missing values in several numerical features: **glucose, education, BPMeds, total cholesterol (totChol), BMI, cigarettes per day (cigsPerDay), and heart rate**.

Handling Strategy: Missing values were imputed using the **median** to reduce the influence of extreme values.

2.2 Duplicate Records

No duplicate rows were detected, ensuring that all patient records are unique.

2.3 Outlier Detection

Outliers were identified using **box plots** and **interquartile range (IQR)** analysis:

- Total Cholesterol (totChol): up to 696
- Systolic Blood Pressure (sysBP): >290
- BMI: up to 56.8
- Glucose: >390

Handling Strategy: Outliers were clipped using the IQR method ($Q11.5\times IQR, Q3+ 1.5 \times IQR$) to reduce their influence while preserving all samples.

2.4 Target Variable Imbalance

The **TenYearCHD** target is imbalanced: $\sim$15% of individuals developed CHD.

Handling Strategy: Stratified train-test splitting and proper evaluation metrics were used during modeling.

2.5 Feature Scaling

All numerical features were standardized using **StandardScaler** to have zero mean and unit variance, essential for distance- and gradient-based algorithms.

3 Summary of Data Preprocessing

- Missing values imputed with median.
- No duplicate records detected.
- Outliers clipped using IQR.
- Features standardized.
- Stratified train-test split applied to maintain class distribution.

4 Regression Analysis and Model Comparison

Regression models were used to predict **systolic blood pressure (sysBP)** using selected features: **Age, BMI, Current Smoking Status**. Data was split 80/20 for training/testing with a fixed random seed.

4.1 Linear Regression

Dataset	MSE	RMSE	R^2
Training	333.71	18.27	0.2312
Testing	306.98	17.52	0.2201

Table 1: Linear Regression Performance

4.2 Lasso Regression (L1 Regularization)

Dataset	MSE	RMSE	R^2
Training	333.71	18.27	0.2312
Testing	306.98	17.52	0.2201

Table 2: Lasso Regression Performance

Feature Importance (Lasso Coefficients):

Feature	Coefficient
Age	0.878
BMI	1.480
Current Smoker	0.000

4.3 Regression Comparison

Model	Dataset	MSE	RMSE	R^2	Observations
Linear Regression	Training	333.71	18.27	0.2312	Captures linear trends
Linear Regression	Testing	306.98	17.52	0.2201	Moderate generalization
Lasso Regression	Training	333.71	18.27	0.2312	Regularized, improves interpretability
Lasso Regression	Testing	306.98	17.52	0.2201	Similar accuracy, feature selection applied

Table 3: Regression Models Performance Comparison

5 Classification Models Analysis

— Model — Accuracy — Precision (CHD) — Recall (CHD) — F1-Score (CHD) — Strengths — Weaknesses — ———-———————-————-————-———————— — Logistic Regression — 0.61 — 0.23 — 0.68 — 0.35 — High recall; effective for positive cases — Low precision; high false positives — — Decision Tree — 0.67 — 0.25 — 0.57 — 0.34 — Better accuracy; interpretable rules — Slightly lower recall than Logistic Regression —

Interpretation: Logistic Regression prioritizes CHD detection (high recall), while Decision Tree provides better balance and interpretability.

6 Clustering Analysis

6.1 Data Preparation

- Removed **TenYearCHD** target.
- Outliers removed (IQR method).
- Standardized features.
- PCA applied for visualization.

6.2 Optimal Clusters

Elbow method suggested $K = 3$.

6.3 Clustering Evaluation

Silhouette Score ≈ 0.16 (moderate separation).

6.4 Cluster Distribution (Expected)

Cluster	Samples	Interpretation
Cluster 0	1,250	Lower-risk, stable indicators
Cluster 1	900	Moderate-risk, elevated BMI/BP
Cluster 2	1,350	Higher-risk, multiple cardiovascular risk factors

Table 4: Cluster Distribution

7 Conclusion

This project provided a comprehensive machine learning analysis of the Framingham Heart Disease Dataset:

- **Data preprocessing:** Missing values, outliers, scaling, and class imbalance addressed.
- **Regression:** Age and BMI significant predictors; Lasso adds interpretability.
- **Classification:** Logistic Regression prioritizes recall; Decision Tree offers accuracy and interpretability.
- **Clustering:** Three distinct patient groups identified with varying cardiovascular risk.

Combining regression, classification, and clustering provides actionable insights into cardiovascular risk, supporting data-driven clinical decision-making.