COURSE PROJECT – DELIVERABLE 2

Heart Failure Prediction Using Machine Learning: A Binary Classification Approach

SAINZOLBOO ANUJIN – 1311002
SYED ABDUL RAHMAN – 1260544
TAO XUE – 1316845

Table of Contents

1. Abstract

Cardiovascular diseases (CVDs) represent the major cause of death worldwide, which make for approximately 32% of all global deaths with an estimated 19.8 million fatalities in 2022 . Early detection and accurate prediction of heart disease are needed for timely intervention and improved patient outcomes. This project explore the application of machine learning techniques to predict heart failure using clinical data from the Kaggle Heart Failure Prediction Dataset. The dataset consists of 918 patient records with 12 clinical features, including both numerical measurements (age, blood pressure, cholesterol levels, heart rate) and categorical indicators (chest pain type, ECG results, exercise-induced symptoms). The main objective is to make a reliable binary classification model capable of identify between patients with heart disease (class 1) and normal patients (class 0).

This preliminary report establishes the foundation for comprehensive analysis by examining the problem significance, reviewing relevant literature on machine learning applications in cardiovascular disease prediction, and conducting extensive exploratory data analysis with practical statistical investigations. Current research demonstrates that ensemble methods such as Random Forest and advanced techniques like XGBoost consistently achieve prediction accuracies exceeding 85-95% on cardiovascular datasets. The practical data exploration reveals important characteristics including a class imbalance of 2.12:1 (disease to normal ratio), gender

disparity (80% male patients), and weak but significant age correlation with disease status (r=0.065, p=0.049).

Comprehensive statistical analysis consists descriptive statistics, correlation matrices, chi-square tests, t-tests, and outlier detection using the IQR method provide actionable insights for preprocessing strategies. The analysis identifies 4.68% of records with zero cholesterol values (likely missing data), 4.90% outliers in ST depression measurements, and low inter-feature correlations indicating independent predictive variables suitable for machine learning models. These findings directly inform the feature engineering, data preprocessing, and model selection strategies that will be done in subsequent project phases. The expected outcomes of this project include the implementation of multiple supervised learning algorithms, comparative performance evaluation across different classifiers, and findings of the most influential clinical features for heart disease prediction. This work contributes to the growing body of research demonstrating that machine learning-based diagnostic tools can enhance early detection capabilities, reduce healthcare costs through improved efficiency, and support clinical decision-making processes in cardiovascular medicine.

2. Introduction

2.1. Background and Motivation

Cardiovascular diseases have formed as a major global health crisis, with the burden continuing to escalate despite advances in medical technology and treatment protocols. According to the World Health Organization, CVDs claimed 19.8 million lives in 2022, representing a 60% increase from the 12.1 million deaths recorded in 1990. This huge threat trend reflects both population growth and aging demographics, as well as the persistence of preventable metabolic, environmental, and behavioral risk factors. Ischemic heart disease remains the leading cause of CVD mortality globally, with an age-standardized rate of 108.8 deaths per 100,000 population.

2.2.The Problem Statement and Objectives

Despite the demonstrated potential of machine learning in cardiovascular disease prediction, several challenges arises . These include **dataset limitations** such as small sample sizes, class imbalance between disease-positive and disease-negative cases, and missing or inconsistent data quality. **Feature selection and engineering** play critical roles in model performance, as cardiovascular datasets typically contain numerous clinical variables with varying degrees of predictive significance. Identifying the optimal subset of features that maximizes diagnostic accuracy while minimizing computational complexity remains an active area of research.

Model interpretability presents another important consideration, particularly in clinical applications where healthcare providers require transparent explanations for diagnostic predictions. While complex deep learning models may achieve marginally higher accuracy, simpler algorithms with clear decision pathways often prove more acceptable in medical practice. Moreover, **generalizability across populations** represents a significant concern, as

models trained on data from specific geographic regions or demographic groups may not perform equally well when applied to different patient populations. This project addresses these challenges through systematic investigation of machine learning techniques applied to the Kaggle Heart Failure Prediction Dataset.

3. Data Analysis

3.1. Dataset Characteristics and Structure

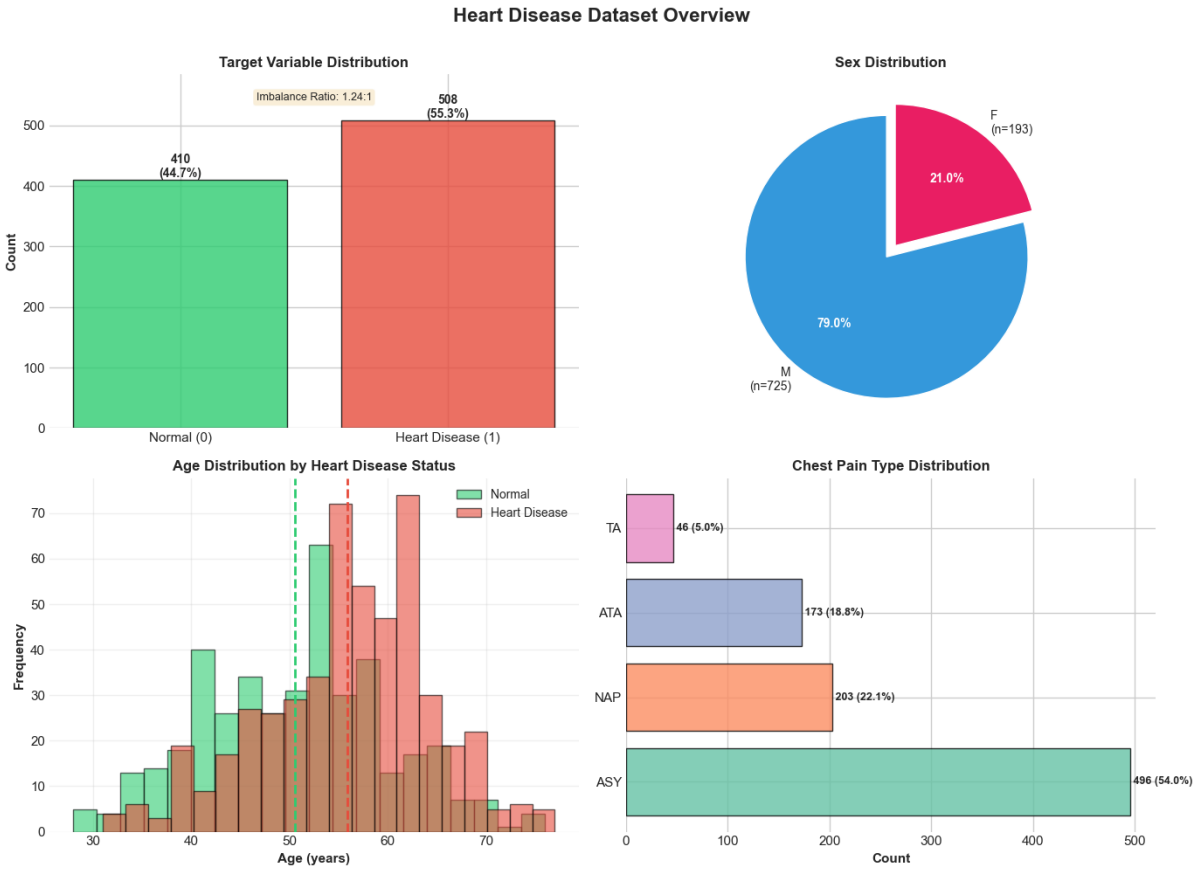
Characteristics	Value
Total Records	918 patients
Total Features	12 (11 predictors + 1 target)
Numerical	6 (Age, RestingBP, Cholesterol, MaxHR, Oldpeak, FastingBS)
Categorical Features	5 (Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope)
Target Variable	HeartDisease (Binary: 0=Normal, 1=Disease)
Explicit Missing Values	0
Implicit Missing Values	172 records (18.74%) with Colesterol = 0 1 record (0.11) with RestingBP = 0
Memory Usage	317.21 KB

Table 1: Dataset Overview

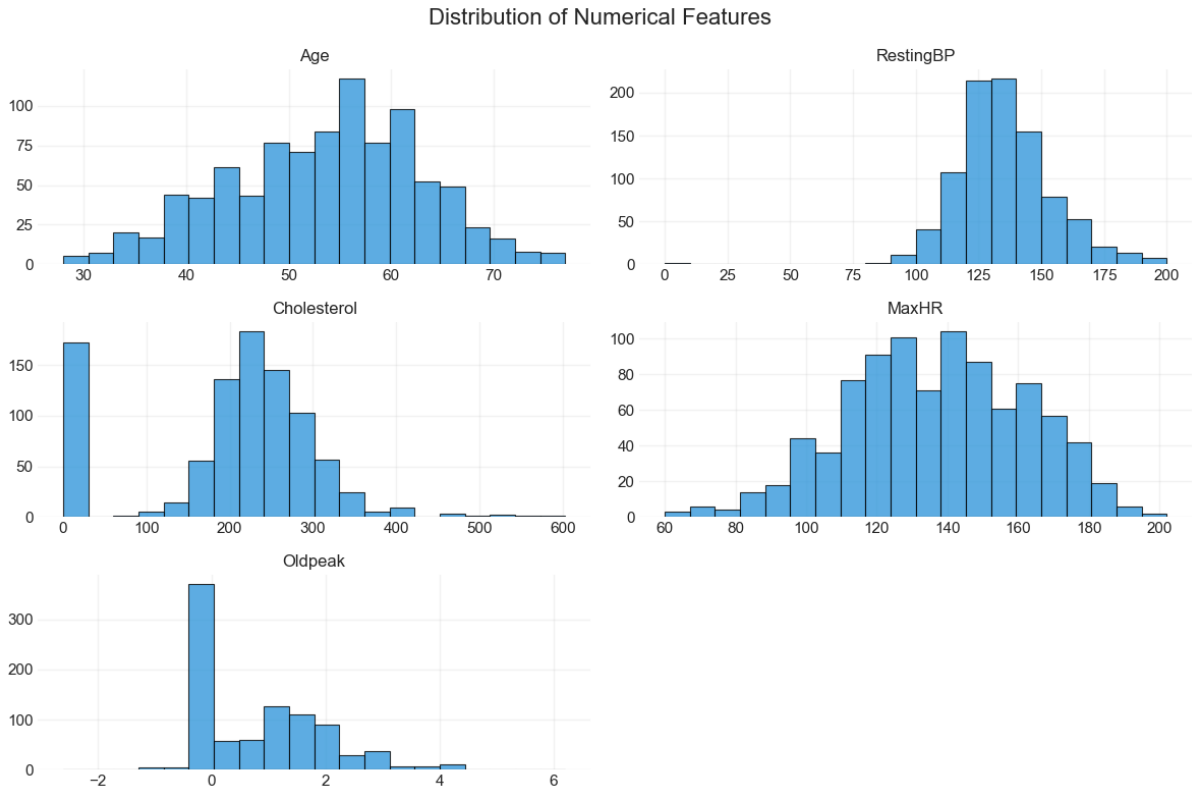
The dataset demonstrates excellent memory efficiency at 317.21 KB, enabling rapid iterative model development without computational constraints. While no explicit missing values exist, implicit missing data patterns require preprocessing attention, particularly the 172 cholesterol zero values that cannot represent true physiological measurements.

3.2. Target Variable Distribution

The class distribution reveals a relatively balanced dataset with 1.24:1 disease-to-normal ratio, markedly better than many medical datasets exhibiting 5:1 or 10:1 imbalances. This moderate imbalance minimizes the need for aggressive oversampling techniques like SMOTE, though stratified cross-validation and class-weighted loss functions remain advisable. The baseline accuracy of 55.34% (achieved by always predicting disease) establishes the minimum acceptable model performance threshold.



3.3. Numerical Features Distribution and Analysis



Age (years): Mean age of 53.51 years represents middle-aged to older adults. The age distribution is approximately bell-shaped, centered around 55 years, with most samples between 40 and 65 years old. The sample includes few young (<35) or elderly (>70) individuals. There are no major outliers or strong skewness, so age can be used directly as a continuous input variable.

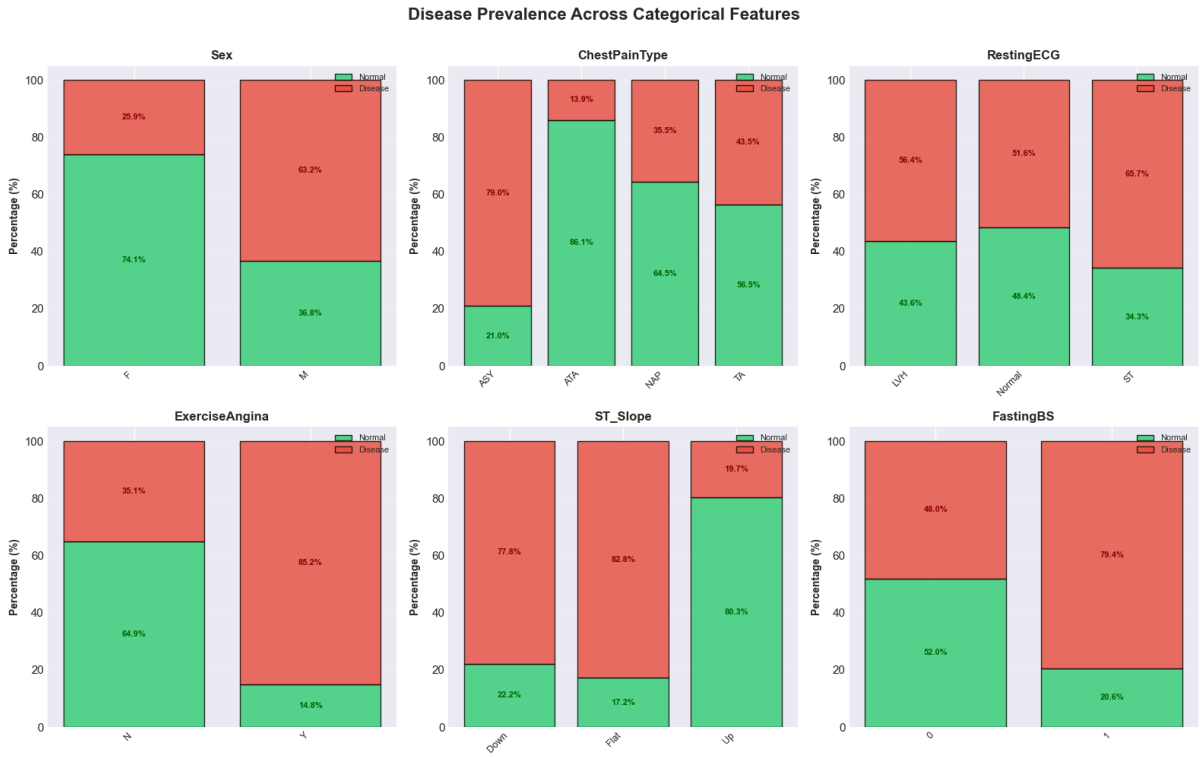
The resting blood pressure (RestingBP mm/Hg) : Average 132.40 mmHg exceeds the hypertension threshold of 130 mmHg. Distribution is obviously right-skewed, concentrated between 110–150 mmHg with a median around 125 mmHg. The extremely low value (near 0, not reasonable) should be treated as outliers. Other than that, most values fall in the normal range. One zero value requires removal or imputation.

Cholesterol (mm/dl): The value distribution centered around 200 mg/dL, typical for adult populations. However, there are a large number of zero values, which are likely data entry errors or missing-value codes.

MaxHR: The variable is approximately normally distributed, ranging from 60 to 200 and centered near 140 bpm. No significant outliers appear, and the data seem symmetric. Therefore MaxHR can be used directly as a continuous feature after standardization. In terms of medical area, lower MaxHR may reflect poorer heart performance, which might correlate with a higher likelihood of heart disease.

Oldpeak = ST [Numeric value measured in depression], 0 – normal. High value means high risk: The variable is highly right-skewed, with most values near 0 and a few extreme cases above 4. A small number of negative values likely to be measurement errors. If only consider the values in range 0–4, the distribution remains slightly skewed but realistic.

3.4. Numerical Features Distribution and Analysis



ST_Slope emerges as the strongest categorical predictor ($\chi^2 = 355.92$, $p < 0.001$) with dramatic disease prevalence differences: flat (82.8%), downsloping (77.8%), versus upsloping (19.8%). This 4.2-fold difference reflects the clinical significance of ST segmentmorphology during exercise testing

ChestPainType shows the second-highest chi-square (268.07, $p < 0.001$). Paradoxically, asymptomatic patients exhibit the highest disease prevalence (79.0%), representing the dangerous "silent ischemia" phenomenon where severe coronary disease exists without typical angina symptoms. Atypical angina shows the lowest prevalence (13.9%), suggesting these symptoms often have non-cardiac etiologies.

ExerciseAngina demonstrates exceptionally strong discriminative power ($\chi^2 = 222.26$, $p < 0.001$): patients with exercise-induced angina show 85.2% disease prevalence versus 35.1% without, representing a 2.4-fold risk increase

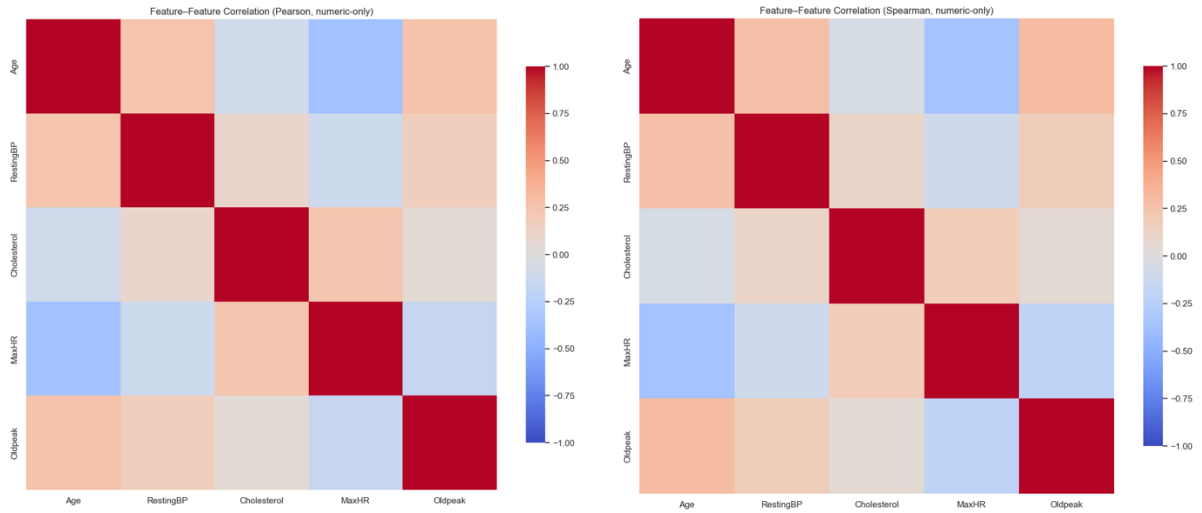
Sex exhibits highly significant association ($\chi^2 = 84.15$, $p < 0.001$) with males showing 2.4× higher disease prevalence (63.2% vs 25.9%). This substantial gender disparity reflects both biological differences in cardiovascular disease presentation and the dataset's strong male predominance (79% male), which may limit model generalizability to female populations[^53].

RestingECG shows the weakest (though still significant) association ($\chi^2 = 10.93$, $p = 0.004$), with ST abnormalities conferring 65.7% disease prevalence versus 51.6% for normal ECG.

3.5. Correlation Analysis

The Analysis will be applied on numerical features only, and the result of how strongly different features are linearly related to each other. Heatmap will be used as tool in the analysis. The abnormal in ‘Cholesterol’ and ‘RestingBP’ should be handled before the analysis to avoid mistakes.

The two Feature-Feature heatmaps (Pearson & Spearman) are as follows:



Feature A	Feature B	Pearson	|Pearson|
MaxHR	Age	-0.382045	0.382045
Age	Oldpeak	0.258612	0.258612
Age	RestingBP	0.254399	0.254399
MaxHR	Cholesterol	0.235792	0.235792
RestingBP	Oldpeak	0.164803	0.164803
MaxHR	Oldpeak	-0.160691	0.160691
MaxHR	RestingBP	-0.112135	0.112135
Cholesterol	RestingBP	0.100893	0.100893
Cholesterol	Age	-0.095282	0.095282
Cholesterol	Oldpeak	0.050148	0.050148

Both Pearson and Spearman correlation analyses showed similar patterns, so the relationships among numerical features are consistent and linear. So, we only discussed the Pearson correlation matrix for simplicity.

The heatmap showed that the variables (Age, RestingBP, Cholesterol, MaxHR, and Oldpeak) demonstrated no strong correlations ($|r| > 0.8$). The highest correlation was a moderate negative relationship between Age and MaxHR, meaning that older patients tend to achieve lower maximum heart rates. Mild positive correlation existed between Age and Oldpeak, while other feature pairs showed weak or negligible relationships. The numerical summary of Pearson correlations also confirms the analysis. These findings suggest that the numerical features are independent and suitable to use in the modeling stage without multicollinearity concerns.

3.6. Preprocessing Strategy

Based on comprehensive exploratory analysis, the following preprocessing pipeline will be implemented:

Step 1: Handle Missing Data

Handling 0 values with Median or let the **RFT** handle itself

Step 2: Encode Categorical Variables

- One-hot encoding for ChestPainType, RestingECG, ST_Slope
- Binary Encoding for Sex, ExerciseAngina

Step 3: Feature Scaling

Step 4: Train Test Split (80/20)

Step 5: Cross Validation (5-fold)

3.7.Classification Algorithm Formulation (Random Forest Tree)

Random Forest (RF) will be used in this project since it is robust, interpretable, and has a strong performance on medical datasets. It operates based on the following mechanism:

Component— Decision Tree

Decision Tree is a binary tree structure using a recursive partitioning process to finish classification task. At each node, the data is split into two subsets with a feature and a threshold. The optimal split is determined by minimizing a node impurity measure: the Gini Impurity or Entropy.

Gini Impurity:

$$G(D) = 1 - \sum_{i=1}^{i} p_i^2$$

Where p_i is the proportion of class c samples in dataset. Value of 0 means pure (all samples in one class), and higher values means more mixed classes.

Entropy:

$$H(D) = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{i}} p_{\boldsymbol{i}} \log_2 p_i$$

Entropy shows the uncertainty of a node; In a pure node entropy = 0.

The best split at each node is the one that maximizes the information gain:

$$Gain(X_j) = I(D) - \sum_{k} \frac{|D_k|}{|D|} I(D_k)$$

Where $I()$ is the impurity measure (Gini or entropy), and D_k are the subset after splitting with the feature X_j.

The splitting recursively is executed until the conditions are met (e.g., max depth)

Selection of Random samples and Random features in RF

Each tree is trained on a random subset of the data, which is drawn with replacement from the original dataset. Mathematically:

$$D^{(b)} = \{(x_i, y_i)\}_{i=1}^{N}, where\ (xi, yi) \sim D\ with\ replacement$$

This means each tree sees about 63% unique samples (the rest are duplicates).

In the subset of each tree, only a random subset feature is selected. This makes trees de-correlated, to avoid some strong features might dominate every tree.

Aggregation and Prediction in RF

After all trees are trained independently, their predictions are aggregated (only classification is discussed here):

$$\hat{y} = mode\{f1(x), f2(x), \dots, fB(x)\}$$

Where B is the number of trees in the forest.

Because each tree is slightly different, averaging their predictions reduces random fluctuations and noise. As the number of trees B increases, the ensemble's variance decreases, and the prediction stabilizes.

Bias–Variance Tradeoff in RF

A single decision tree tends to have low bias and high variance. In RF, variance is reduced by average many de-correlation trees. If ρ is the average correlation between trees, the variance of the forest is approximately:

$$Var(RF) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Where σ^2 is the variance of an individual tree. As B (number of the trees) increase, the second term will vanish, and the variance only rely on ρ.

The mechanism of RF shows that it is an ensemble learning algorithm combining multiple decision trees and is suitable for classification tasks. While individual decision trees may be overfitting in the data, a group of different trees, which are trained on different random samples and subsets of features, can achieve better performance.

In this project, RF is supposed to be effective because it can capture complex, nonlinear interactions in medical features (e.g., Age, Cholesterol, MaxHR, Oldpeak), while remaining robust to noise and outliers.

4. Conslusion

This comprehensive preliminary report establishes a robust foundation for machine learning model development through rigorous exploratory data analysis of the 918-patient heart failure prediction dataset. The analysis reveals **exceptionally strong univariate** predictors (all $p \leq 0.004$), **moderate class imbalance** (1.24:1 ratio), and **clinically coherent disease patterns** that made well for high model performance.

Statistical Strength: Unlike many medical datasets where features show weak univariate associations, this dataset demonstrates highly significant relationships for all predictors.

Data Quality: The 18.74% missing cholesterol data and 0.11% missing blood pressure data require preprocessing attention through median imputation. The 19.93% cholesterol outliers largely reflect these missing values rather than physiological extremes.

Clinical Validity: Disease prevalence patterns align with cardiovascular pathophysiology: exercise-induced angina (85.2% disease), asymptomatic presentation (79.0% disease, silent ischemia), flat ST slope (82.8% disease), reduced maximum heart rate (20.50 bpm lower in disease group), and elevated ST depression (0.87mm higher in disease group).

Gender Disparity: The 79% male predominance and 2.4× higher male disease prevalence necessitate sex-stratified model evaluation to assess generalizability to female patients.