

**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

[Abstract](#_Toc212230226)

[Introduction](#_Toc212230227)

[2.1 Background and Motivation](#_Toc212230228)

[2.2 The Problem Statement and Objectives](#_Toc212230229)

[Data Analysis](#_Toc212230230)

[3.1 Dataset Characteristics and Structure](#_Toc212230231)

[3.2 Target Variable Distribution](#_Toc212230232)

[3.3 Numerical Features Distribution and Analysis](#_Toc212230233)

[3.4 Categorical Features Distribution and Analysis](#_Toc212230234)

[3.5 Correlation Analysis](#_Toc212230235)

[3.6 Preprocessing Strategy](#_Toc212230236)

[Method/Algorithm](#_Toc212230237)

[Conslusion](#_Toc212230238)

# 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.

# 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.

# Data Analysis

## 3.1 Dataset Characteristics and Structure

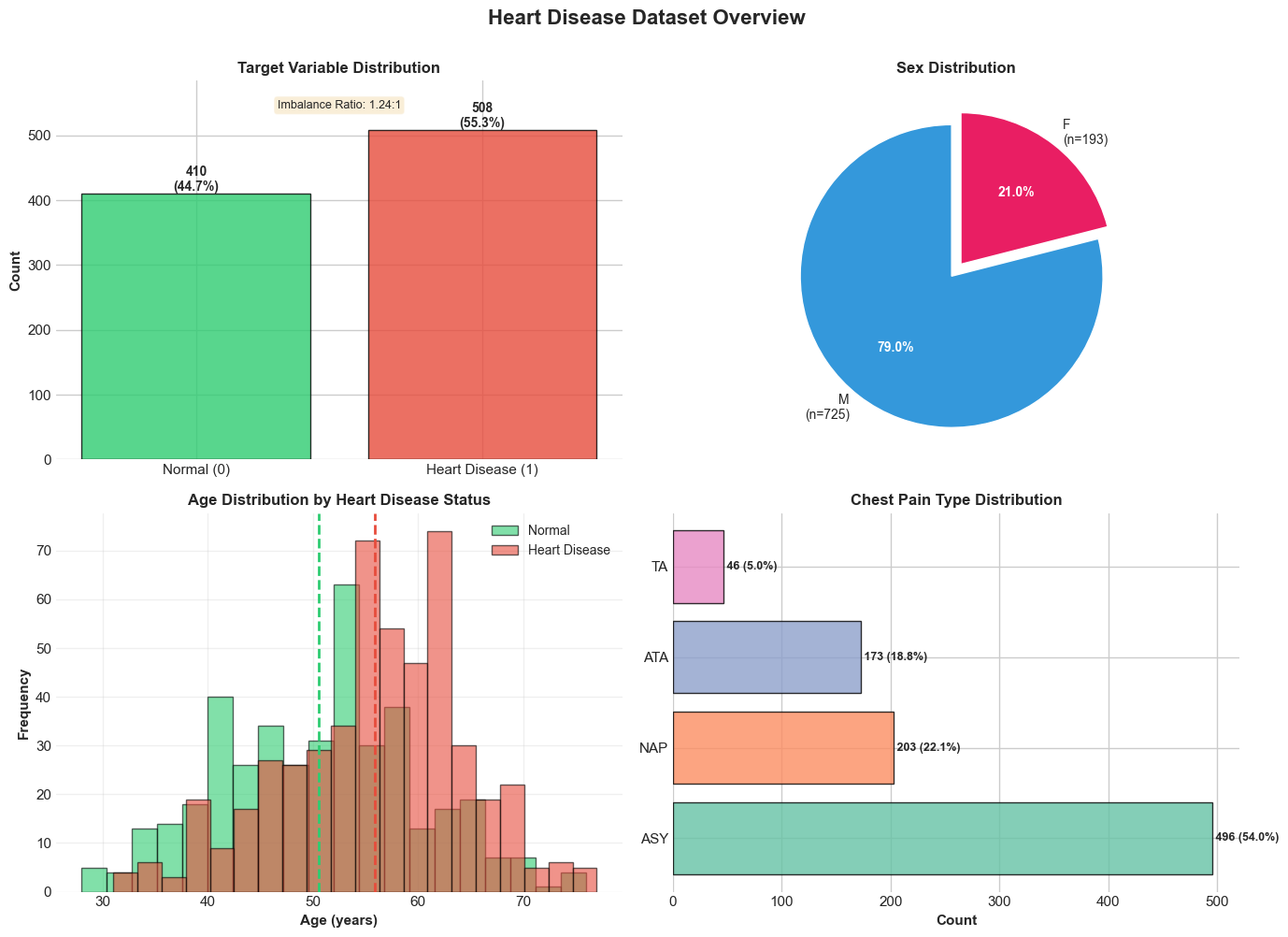
|  |  |
| --- | --- |
| Characteristics | Value |
| Total Records | 918 patients |
| Total Features | 12 (11 predictors + 1 target) |
| Numerical | 6 (Age, RestingBP, Cholesterol, MaxHR, Oldpeak, FastingBS) |
| Categorical Features | 5 (Sex, ChestPainType, RestingECG, ExerciseAngina, ST\_Slope) |
| Target 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.

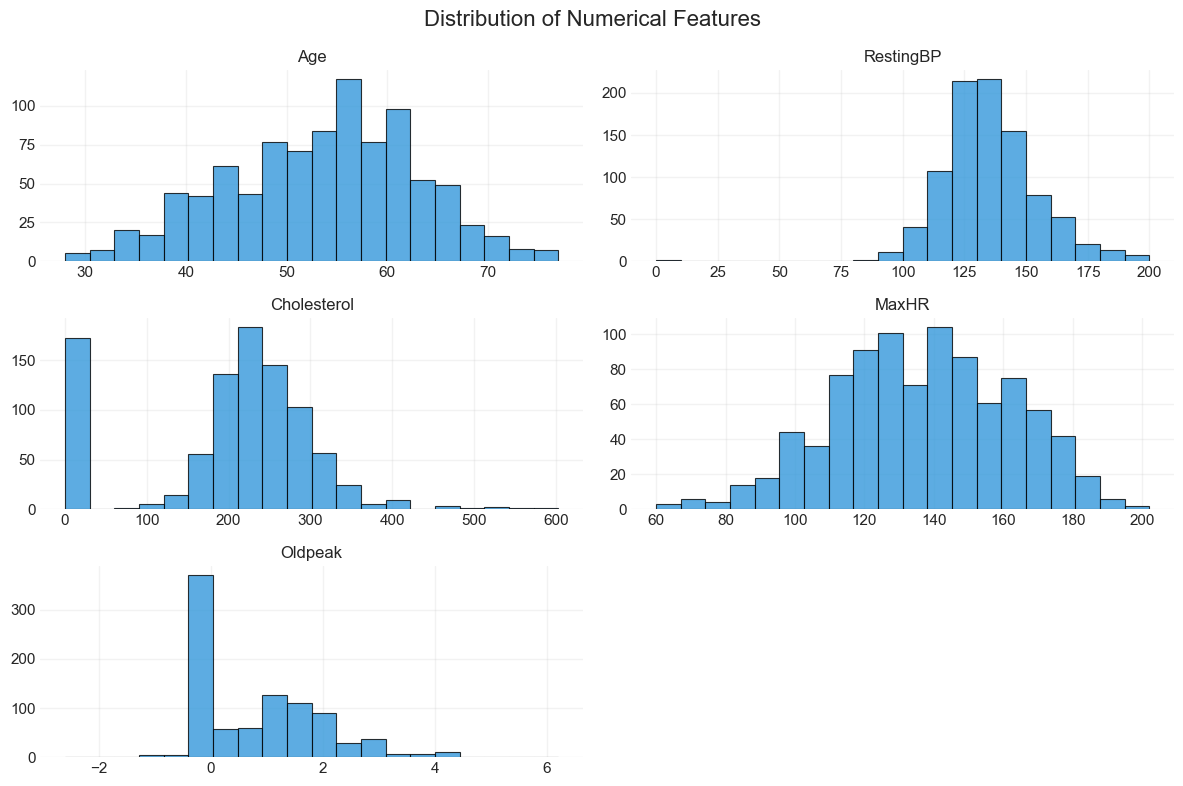
## 3.2 Target Variable Distribution

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 Univariate Analysis

### 3.3.1 Numerical Features



**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.

### 3.3.1 Categorical Features

图表, 条形图

AI 生成的内容可能不正确。

**Sex**: sex of the patient [M: Male, F: Female]. The variable has of two categories: male (M) and female (F). In the dataset, the major samples are male (725, 79%). Through our research, this gender imbalance is typical in heart-disease studies but should be noted, as it may cause a slight bias during model training if not handled carefully.

**FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise] The FastingBS variable is binary, showing if a patient’s fasting blood sugar is higher than 120 mg/dL. Most patients (about 75%) fall into the normal range (value = 0), while a smaller portion (around 25%) have higher level (value = 1). There are no outliers since the feature is categorical. The imbalance between the two categories should be further noticed in bivariate analysis to assess its relationship with the target.

**ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]The variable has four categories: ASY (496), NAP (203), ATA (173), and TA (46). ASY cases dominate the dataset (≈55%), while typical angina cases are rare (≈5%). This imbalance suggests that many patients exhibit no classical chest pain symptoms despite having heart disease. The categorical imbalance should be noted, as it may influence how models interpret this feature.

**RestingECG**: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]. The variable contains three categories: Normal (552), LVH (188), and ST (178). Most patients (≈60%) have normal resting ECG results, while the other groups show LVH or ST-T wave abnormalities. The distribution is moderately imbalanced but reasonable. This feature should be one-hot encoded for use in modeling.

**ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]. There are 547 patients (60%) reporting N and 371 (40%) reporting Y. This moderate imbalance is reasonable. The presence of exercise-induced angina is expected to be positively correlated with heart-disease risk.

**ST\_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]. The variable has three categories: Flat (460), Up (395), and Down (63). The distribution is moderately imbalanced: “Flat” is the most common and “Down” is the least. Since a flat or downsloping ST segment often indicates abnormal cardiac response, this feature is important and will be one-hot encoded. The rare “Down” category (indicate serious situation) should be treated carefully depending on model choice.

## 3.4 Bivariate Analysis (Feature vs Target)

### 3.4.1 Numerical Features

图表, 箱线图

AI 生成的内容可能不正确。**图表, 箱线图

AI 生成的内容可能不正确。图表, 箱线图

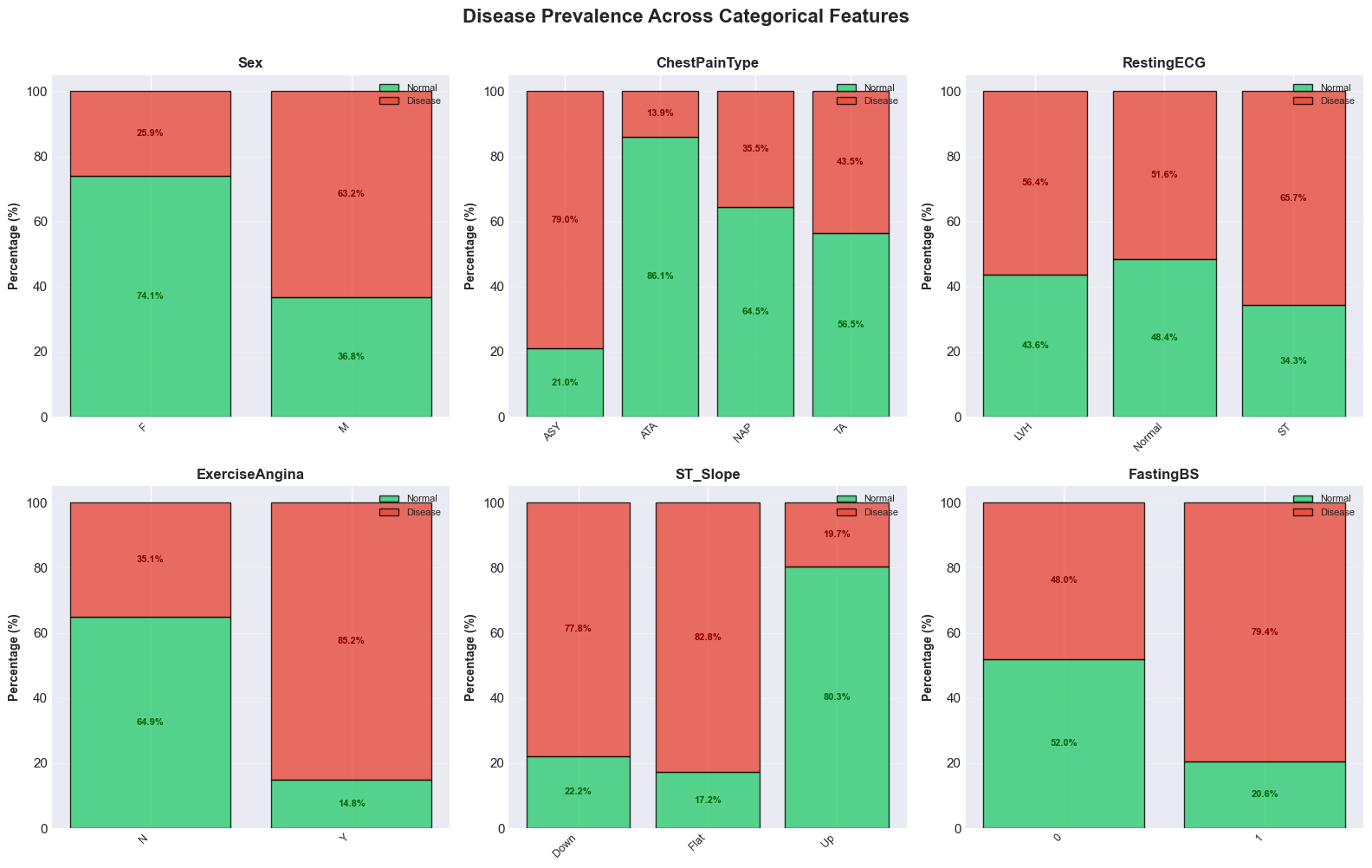
AI 生成的内容可能不正确。**

**MaxHR** distribution reveals that individuals with lower maximum heart rates are more prone to heart disease. Healthy individuals reach higher maximum heart rates during exercise, while those with cardiac issues often have restricted or abnormal heart rate responses. This variable thus captures an important functional difference between the two classes.

**Oldpeak** measures the ST depression induced by exercise relative to rest. The plot shows that higher *Oldpeak* values are strongly associated with heart disease. In medical terms, a larger ST depression suggests greater levels of myocardial ischemia. Therefore, this variable provides a **quantitative measure** of the heart’s abnormal response to stress.

**Age** distribution indicates that heart disease becomes more prevalent among older patients. While younger individuals are less likely to show disease symptoms, the risk increases steadily after age 45–50.This trend is consistent with real-world cardiovascular risk factors.

### 3.4.1 Categorical Features



**ST\_Slope** emerges as the strongest categorical predictor (χ² = 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.

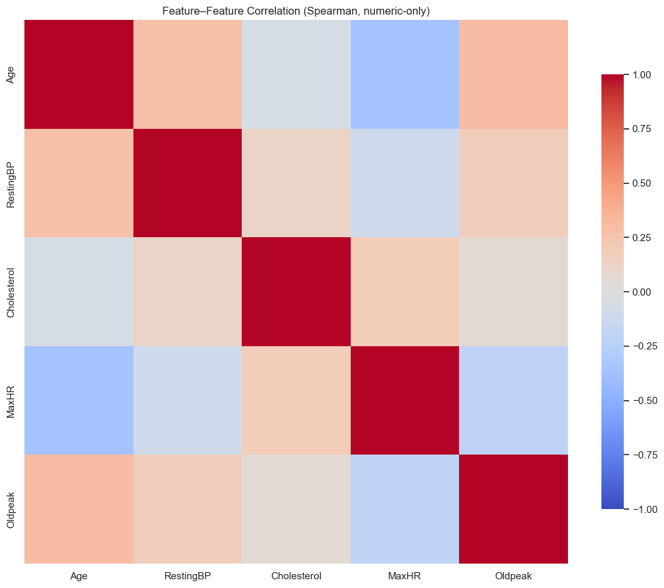
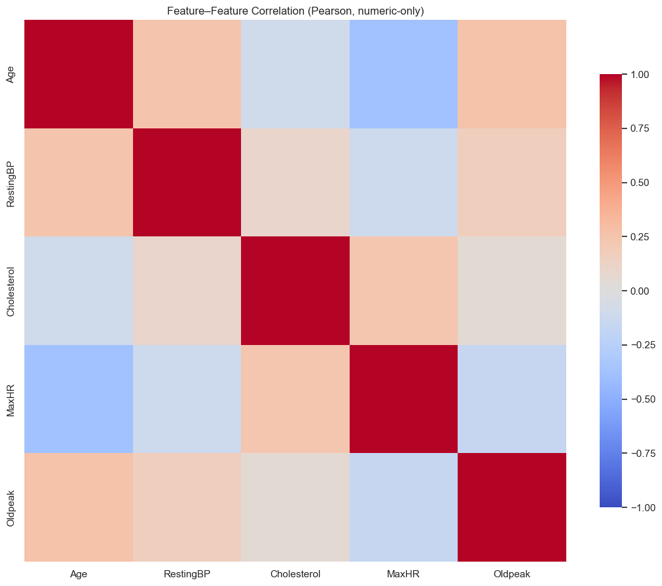
**ExerciseAngina** demonstrates exceptionally strong discriminative power (χ² = 222.26, p < 0.001): patients with exercise-induced angina show 85.2% disease prevalence versus 35.1% without, representing a 2.4-fold risk increase

**Sex** exhibits highly significant association (χ² = 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 (χ² = 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)

# Method/Algorithm

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:

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

Entropy:

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:

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

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 withreplacement from the original dataset. Mathematically:

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):

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 is the average correlation between trees, the variance of the forest is approximately:

Where 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 .

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.

# 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 ≤ 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.