

# Harvard Data Science Capstone: Heart Disease Prediction Model

Andrea De Nardi

2025-12-01

## Introduction

Cardiovascular diseases (CVDs) are a major public health concern, and one of the leading causes of death globally. World Health Organization estimates that 19.8 million people died from CVDs in 2022, representing about a third of all deaths globally (WHO CVDs, 2022). Given the burden of CVDs on mortality and healthcare systems, early detection and targeted care are critical challenges. In this project, we will be using the UCI “Cleveland Heart” dataset (Dua & Graff, 2019) to train machine learning models that predict the presence of heart disease from a set of physiological measures. We will start our analysis by producing a set of visualization that will help us understand what clinical measures can successfully predict the risk of heart disease. Our modeling approach will include simple linear models, classic and penalized logistic regression, decision trees, and final model selection via cross-validation and threshold tuning.

## Variables Definition

Before jumping into our analysis, it is important to define the variables of the Cleveland Heart Dataset (Kaggle, Heart Disease Dataset content description):

- “Age: Patients Age in years (Numeric)
- Sex: Gender (Male : 1; Female : 0) (Nominal)
- cp: Type of chest pain experienced by patient. This term categorized into 4 category.  
0 typical angina, 1 atypical angina, 2 non- anginal pain, 3 asymptomatic (Nominal)
- trestbps: patient’s level of blood pressure at resting mode in mm/HG (Numerical)
- chol: Serum cholesterol in mg/dl (Numeric)
- fbs: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
- restecg: Result of electrocardiogram while at rest are represented in 3 distinct values  
0 : Normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)  
2: showing probable or definite left ventricular hypertrophyby Estes’ criteria (Nominal)
- thalach: Maximum heart rate achieved (Numeric)
- exang: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
- oldpeak: Exercise induced ST-depression in relative with the state of rest (Numeric)
- slope: ST segment measured in terms of slope during peak exercise  
0: up sloping; 1: flat; 2: down sloping(Nominal)
- ca: The number of major vessels (0–3)(nominal)

- thal: A blood disorder called thalassemia  
0: NULL 1: normal blood flow 2: fixed defect (no blood flow in some part of the heart) 3: reversible defect (a blood flow is observed but it is not normal(nominal))
- target: It is the target variable which we have to predict 1 means patient is suffering from heart disease and 0 means patient is normal.”

## Data Wrangling

Prior to modeling, several preprocessing steps were applied to prepare the dataset for analysis. The original **target** variable in the UCI Cleveland Heart dataset is coded as 0 or 1, indicating the absence or presence of heart disease. For interpretability and to support classification workflows in **caret**, this variable was converted into a labeled factor with levels **"no\_disease"** and **"disease"**. All categorical predictors (e.g. chest pain type, resting ECG results, exercise-induced angina, slope, thal, and the number of major vessels) were also recoded as factors to ensure they were correctly handled by algorithms that distinguish between nominal and numeric inputs. In addition to the categorical target, we created a secondary numeric outcome variable, **target\_numeric**, which retains the original 0/1 encoding. This version of the target was used for exploratory data analysis tasks such as computing correlations, where numeric encoding is required. These preprocessing steps ensured that the dataset was properly structured for both statistical modeling and exploratory analyses.

## Analysis

### Exploratory Data Analysis

Let us take a look at the summary statistics of the variables from our dataset:

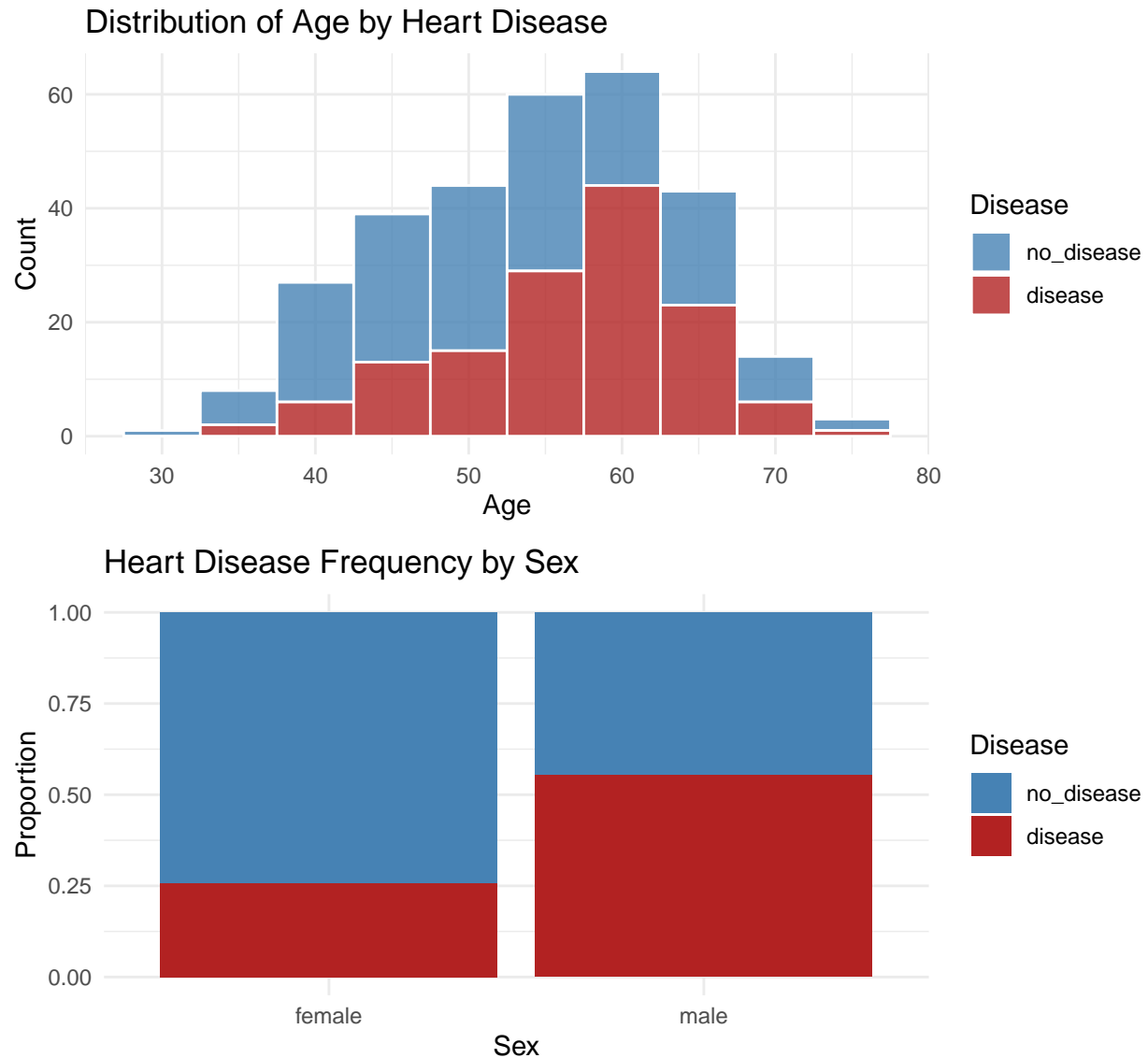
```
summary(heart)
```

```
##      age      sex      cp      trestbps      chol      fbs
## Min.   :29.00  female: 97  1: 23  Min.    : 94.0  Min.    :126.0  0:258
## 1st Qu.:48.00  male   :206  2: 50  1st Qu.:120.0  1st Qu.:211.0  1: 45
## Median :56.00                                Median :130.0  Median :241.0
## Mean   :54.44                                Mean    :131.7  Mean    :246.7
## 3rd Qu.:61.00                                3rd Qu.:140.0  3rd Qu.:275.0
## Max.   :77.00                                Max.    :200.0  Max.    :564.0
## restecg  thalach  exang  oldpeak  slope  ca      thal
## 0:151    Min.    : 71.0  0:204  Min.    :0.00  1:142  0    :176  3    :166
## 1: 4     1st Qu.:133.5  1: 99  1st Qu.:0.00  2:140  1    : 65  6    : 18
## 2:148    Median :153.0              Median :0.80  3: 21  2    : 38  7    :117
##          Mean   :149.6              Mean    :1.04  3    : 20  NA's: 2
##          3rd Qu.:166.0              3rd Qu.:1.60  NA's: 4
##          Max.   :202.0              Max.    :6.20
##      target  target_numeric
## no_disease:164  Min.    :0.0000
## disease      :139  1st Qu.:0.0000
##               Median :0.0000
##               Mean   :0.4587
##               3rd Qu.:1.0000
##               Max.   :1.0000
```

The final dataset contains 303 patients, with a mean age of approximately 54 years (range 29–77). The sample includes 206 males and 97 females. Heart disease is present in 139 patients (46%) and absent in 164

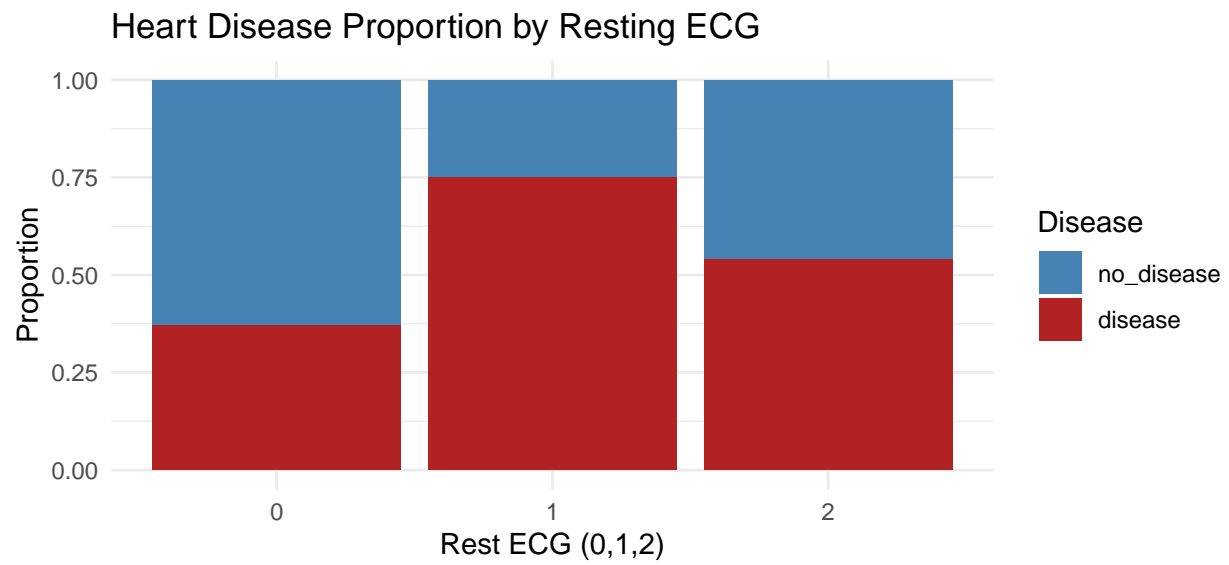
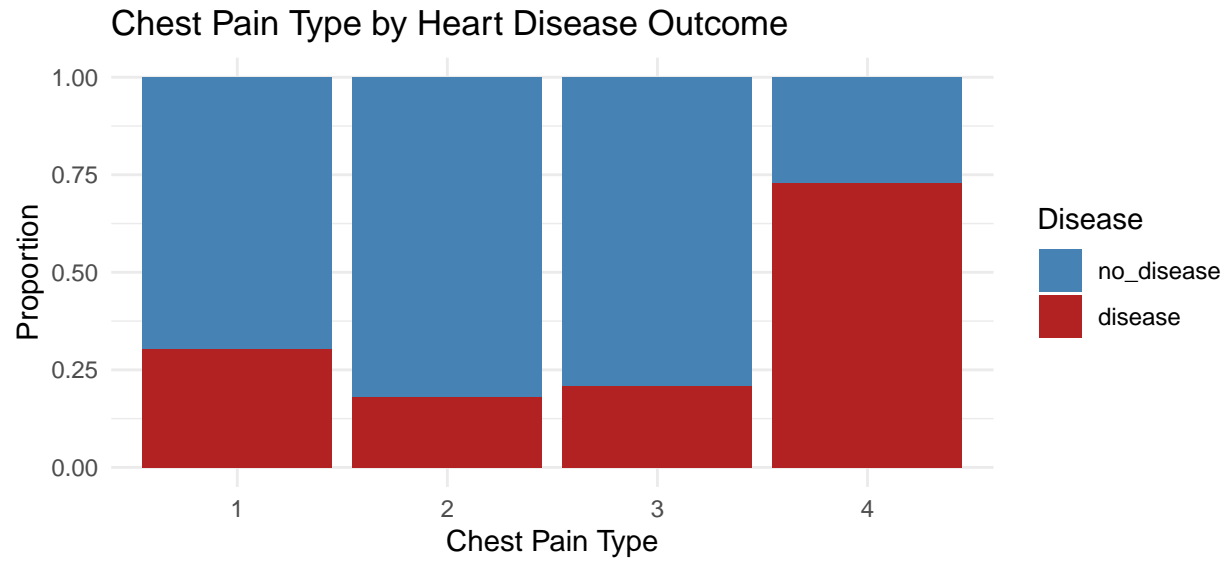
patients (54%), reflecting a reasonably balanced outcome distribution. Several clinical measurements show substantial variability, including resting blood pressure (94–200 mmHg), cholesterol levels (126–564 mg/dl), and maximum heart rate achieved (71–202 bpm). Categorical predictors such as chest pain type, resting ECG results, exercise-induced angina, the number of major vessels observed via fluoroscopy, and thallium stress test results also exhibit diverse distributions across levels. A small number of missing values is present in the thal and ca variables, which will need to be handled before modelling. Overall, the dataset provides a heterogeneous set of demographic, clinical, and physiological features suitable for predicting heart disease. Let us now visualize the relationships between the predictors and the dependent variable.

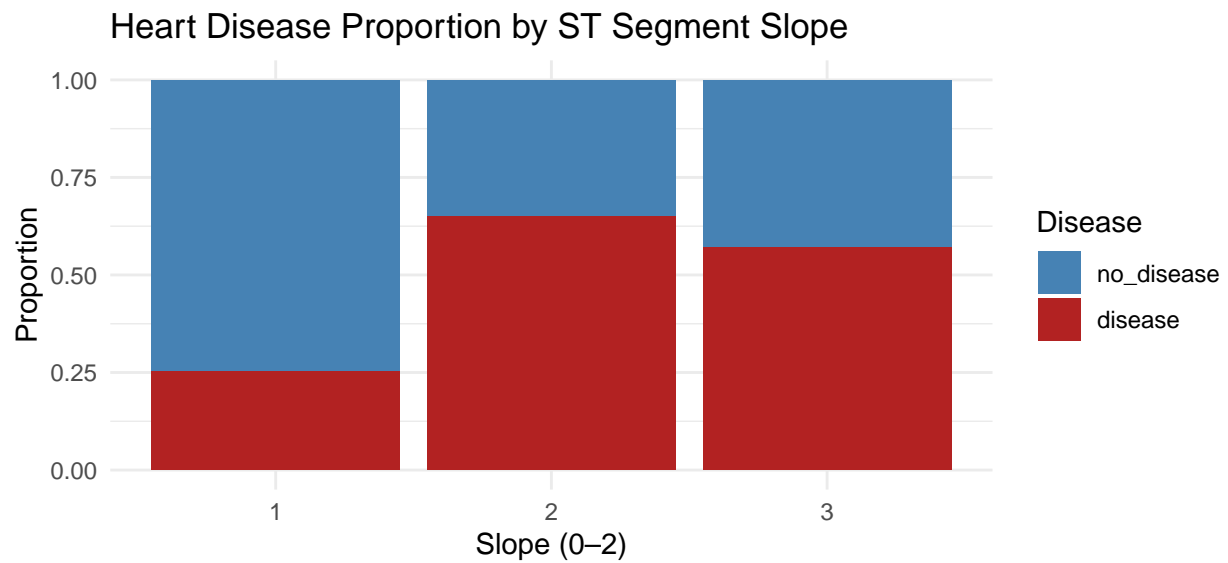
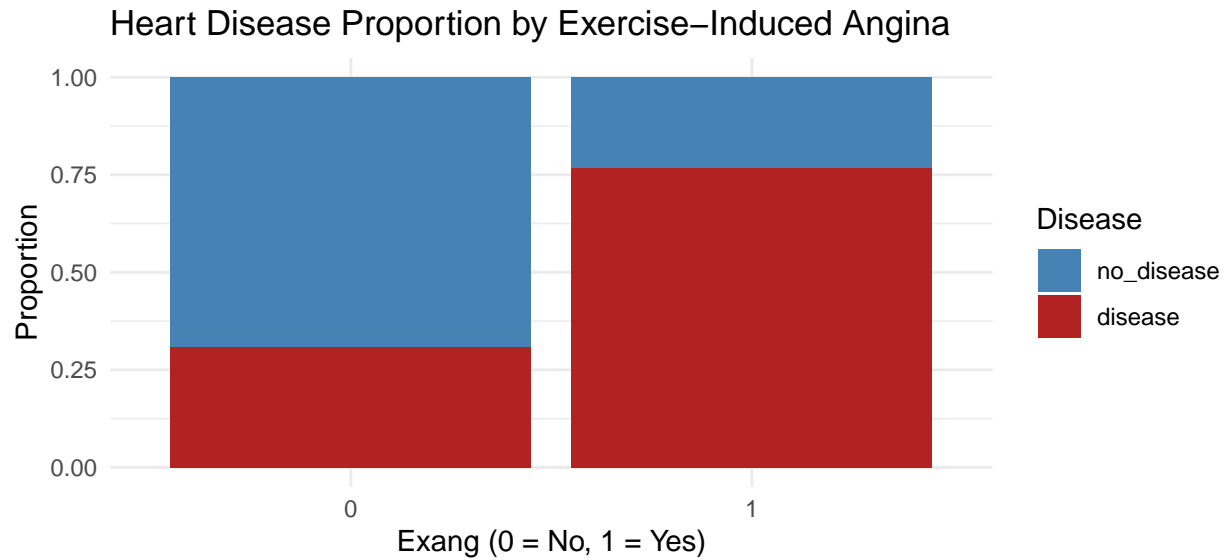
## Demographic Factors



The above histograms shows the distribution of age and sex in our dataset. The bins have been colored to reflect the diagnosis within each group (disease = red, no disease = blue). A pattern seems to be emerge, and to indicate that there is a positive relationship between age and heart disease diagnosis, although the strength of the link appears to weaken after 70 years (perhaps due to survivorship bias). In addition, male seem to be more prone to heart disease than women.

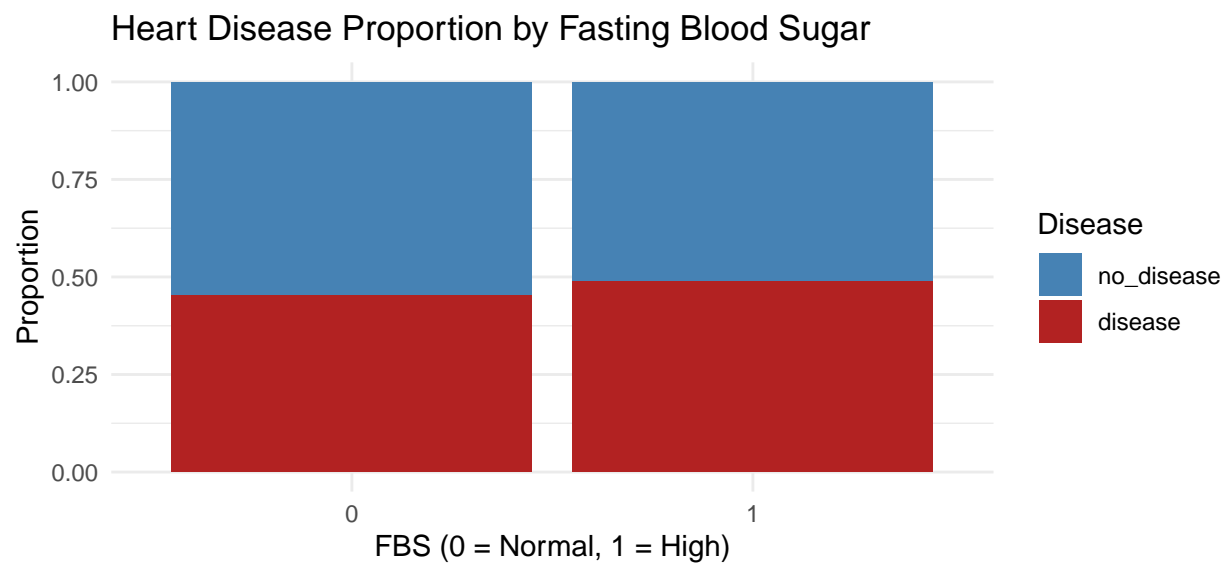
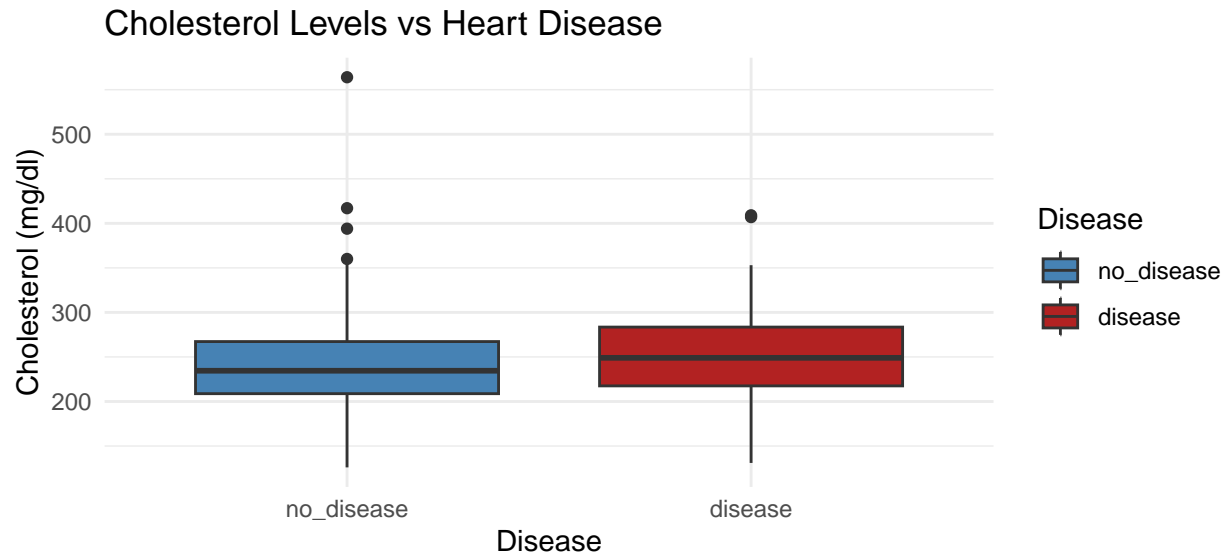
## Chest Pain and ECG-Related Predictors

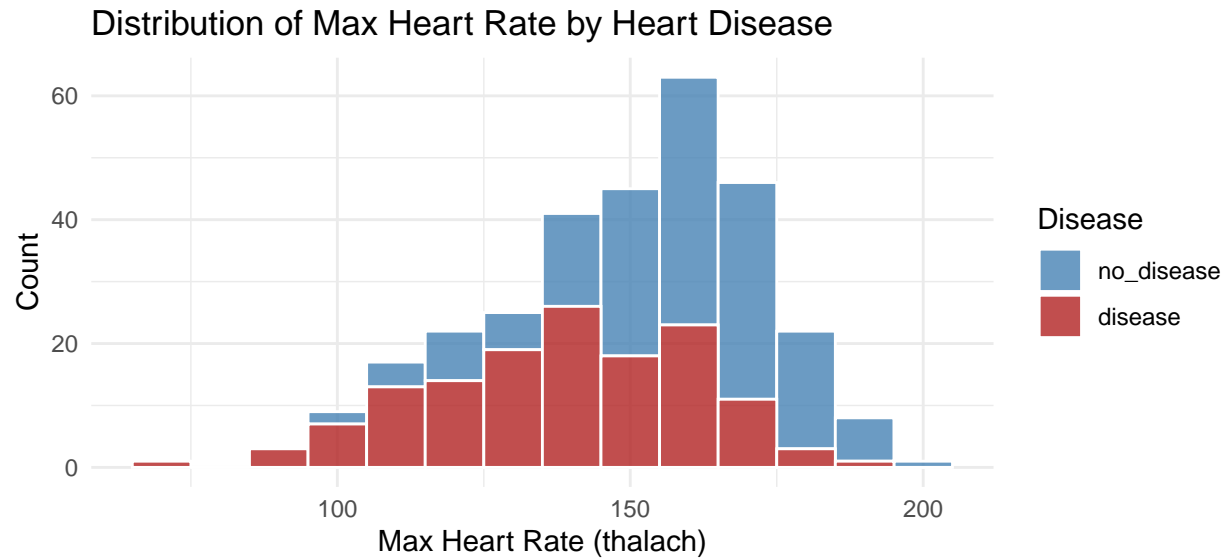




The graphs show that chest pain type #4 and resting ECG's of type 1 (having ST-T wave abnormality) are associated with higher risks of heart disease. In addition, the presence of chest pain during exercise (Exang = 1) and flat or negativs ST Segment Slope are associated with higher risks of CVDs.

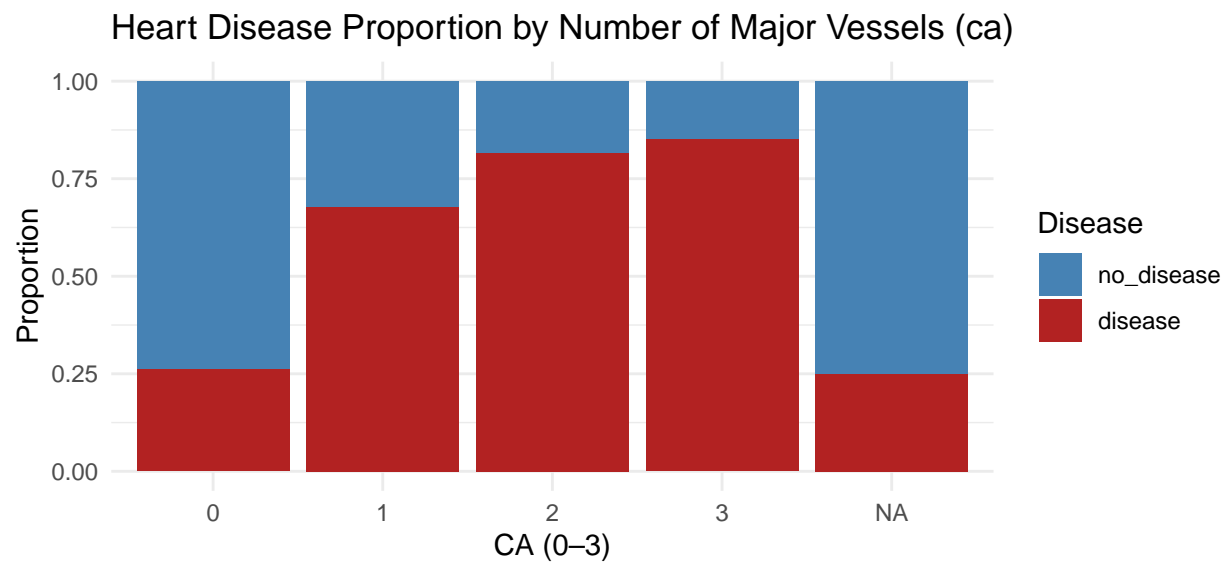
#### Clinical Risk Factors

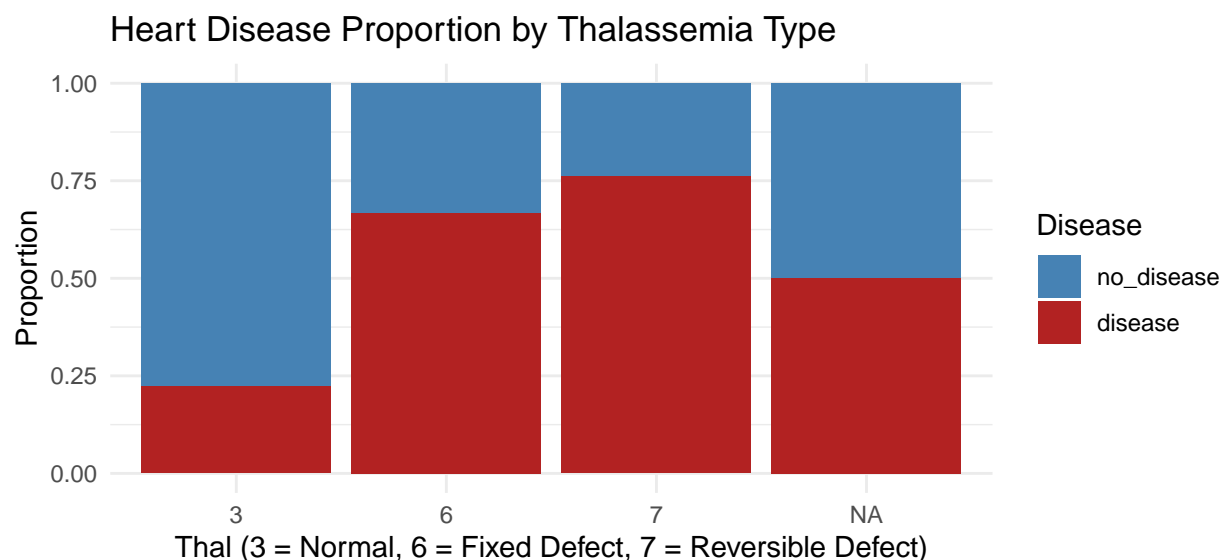




The visualizations seem to indicate a moderate positive relationship between cholesterol levels and disease prevalence (higher summary statistics for cholesterol levels within the disease group). In addition, there is a modest difference in heart disease prevalence between normal and high fasting blood sugar groups (slightly higher for high FBS group). Maximum heart rate appears to be a solid predictor of heart disease outcome, with a significantly higher prevalence of heart disease for lower levels of thalach.

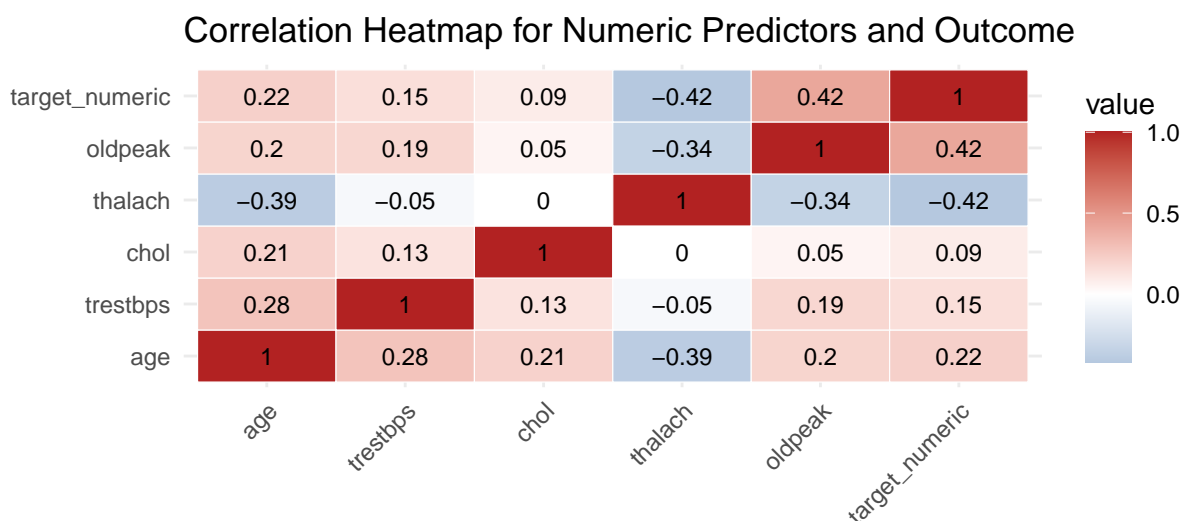
#### Imaging Based Predictors





The plots show a positive relationship between heart disease proportion and number of major vessels, as well as a higher heart disease prevalence among the groups with fixed defect and reversible defect thalassemia types.

#### Correlations Between Numeric Predictors and Outcome



The correlation matrix shows that thalach (maximum heart rate) is moderately negatively correlated with heart disease (-0.42), indicating that patients achieving lower exercise heart rates are more likely to have the condition. In contrast, variables such as oldpeak, age, and resting blood pressure show weak positive correlations with disease (around 0.15-0.22), suggesting only mild linear associations. Overall, no single numeric predictor is strongly correlated with the outcome, supporting the need for multivariate modeling rather than relying on individual variables.

#### Modelling

To build our predictive models for heart disease, we first created an 80/20 train-test split using only complete cases to ensure clean and reproducible evaluation. We began with two baseline methods: a simple



linear regression model and a logistic regression model trained on the numeric predictors, converting predicted probabilities into binary disease classifications using a standard 0.5 threshold. We then expanded the analysis to capture nonlinear patterns by training a decision tree model using the rpart algorithm, tuning its complexity parameter through 10-fold cross-validation. To further improve predictive performance, we implemented a penalized logistic regression model using glmnet, which performs regularization and variable selection. This model was trained using caret’s ROC-optimized cross-validation strategy. Finally, we evaluated a grid of probability cutoff values on the training set and selected the value that maximized balanced accuracy before applying it once to the held-out test set, thereby avoiding data leakage. Altogether, this stepwise progression allowed us to systematically develop increasingly robust predictors of heart disease.

## Results

### Model Selection

Table 1: Comparison of Model Performance Metrics

Model	Accuracy	Sensitivity	Specificity	Balanced_Accuracy
Linear Regression	0.767	0.743	0.800	0.771
Logistic Regression (GLM)	0.750	0.714	0.800	0.757
Decision Tree (rpart)	0.746	0.741	0.750	0.745
Penalized Logistic Regression (glmnet)	0.847	0.938	0.741	0.839

Across the different models we evaluated, we observed meaningful variation in predictive performance. Logistic regression (GLM) performed the weakest overall, with an accuracy of 0.75 and a balanced accuracy of 0.757, indicating that this baseline classifier struggled to reliably distinguish between patients with and without heart disease. Linear regression performed slightly better, achieving an accuracy of 0.767 and a balanced accuracy of 0.771, although it remains limited by the fact that it is not inherently designed for binary classification. The decision tree achieved comparable performance, with an accuracy of 0.746 and a balanced accuracy of 0.745, benefiting from its ability to capture simple nonlinear effects and interactions. The best results were obtained using penalized logistic regression (glmnet), which reached an accuracy of 0.847 and the highest balanced accuracy of 0.839. These results demonstrate that regularization and embedded feature selection substantially improve generalization, leading to more robust predictions in this clinical context.

### Final Model Cutoff Tuning

In the previous models, we relied on the conventional probability cutoff of 0.5 to convert predicted probabilities into binary disease classifications. However, this threshold is arbitrary and may not yield optimal clinical performance, especially when sensitivity and specificity are imbalanced. In this final modeling step, we treated the cutoff as a tunable hyperparameter and evaluated the penalized logistic regression model over a grid of probability thresholds ranging from 0.01 to 0.99 in increments of 0.01. We selected the value that maximized balanced accuracy on the training set and then assessed the model once on the held-out test set using this optimized threshold. The resulting performance metrics are summarized below.

Table 2: Performance of Penalized Logistic Regression with Tuned Probability Cutoff (Cutoff = 0.49)

Model	Cutoff	Accuracy	Sensitivity	Specificity	Balanced_Accuracy
Glmnet with Tuned Cutoff	0.49	0.864	0.938	0.778	0.858

Using the optimized probability cutoff substantially improved the performance of the penalized logistic regression model. The selected threshold of 0.49 achieved an accuracy of 0.864, with sensitivity increasing to 0.938 and specificity reaching 0.778. This balance between correctly identifying patients with heart disease and avoiding false positives is reflected in the high balanced accuracy of 0.858. Overall, tuning the probability cutoff allowed the model to better align with the classification priorities of the task, resulting in a more clinically meaningful performance profile.

## Conclusion

In this project, we analyzed the Cleveland Heart Disease dataset to examine how clinical, demographic, and exercise-related predictors relate to the presence of heart disease. Following an exploratory analysis, we evaluated a range of predictive modeling approaches of increasing complexity, including linear regression, logistic regression, decision trees, and penalized logistic regression. Our findings showed that models capable of capturing nonlinear relationships or leveraging regularization achieved noticeably stronger performance than simpler baselines. In particular, the penalized logistic regression model—with its tuned probability cutoff—provided the highest predictive accuracy and most balanced classification, underscoring the value of choosing an appropriate decision threshold and using regularization to prevent overfitting. Although the dataset is relatively small and lacks external validation, the results demonstrate that thoughtfully designed statistical models can offer meaningful support for early heart disease detection. Future work could benefit from larger and more diverse datasets, integration of richer clinical features such as imaging or longitudinal biomarker data, and exploration of advanced algorithms such as ensemble or deep learning methods.

## References

- Kaggle (Feature Description) Ritwikb3. Heart Disease – Cleveland Dataset. Kaggle. Available at: <https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland>
- UCI Machine Learning Repository (Dataset Source) Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease Dataset (Cleveland subset). UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/heart+disease>
- Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine. <https://archive.ics.uci.edu/ml/index.php>
- World Health Organization (CVD Statistics) World Health Organization (2022). Cardiovascular diseases (CVDs) – Fact sheet. [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))