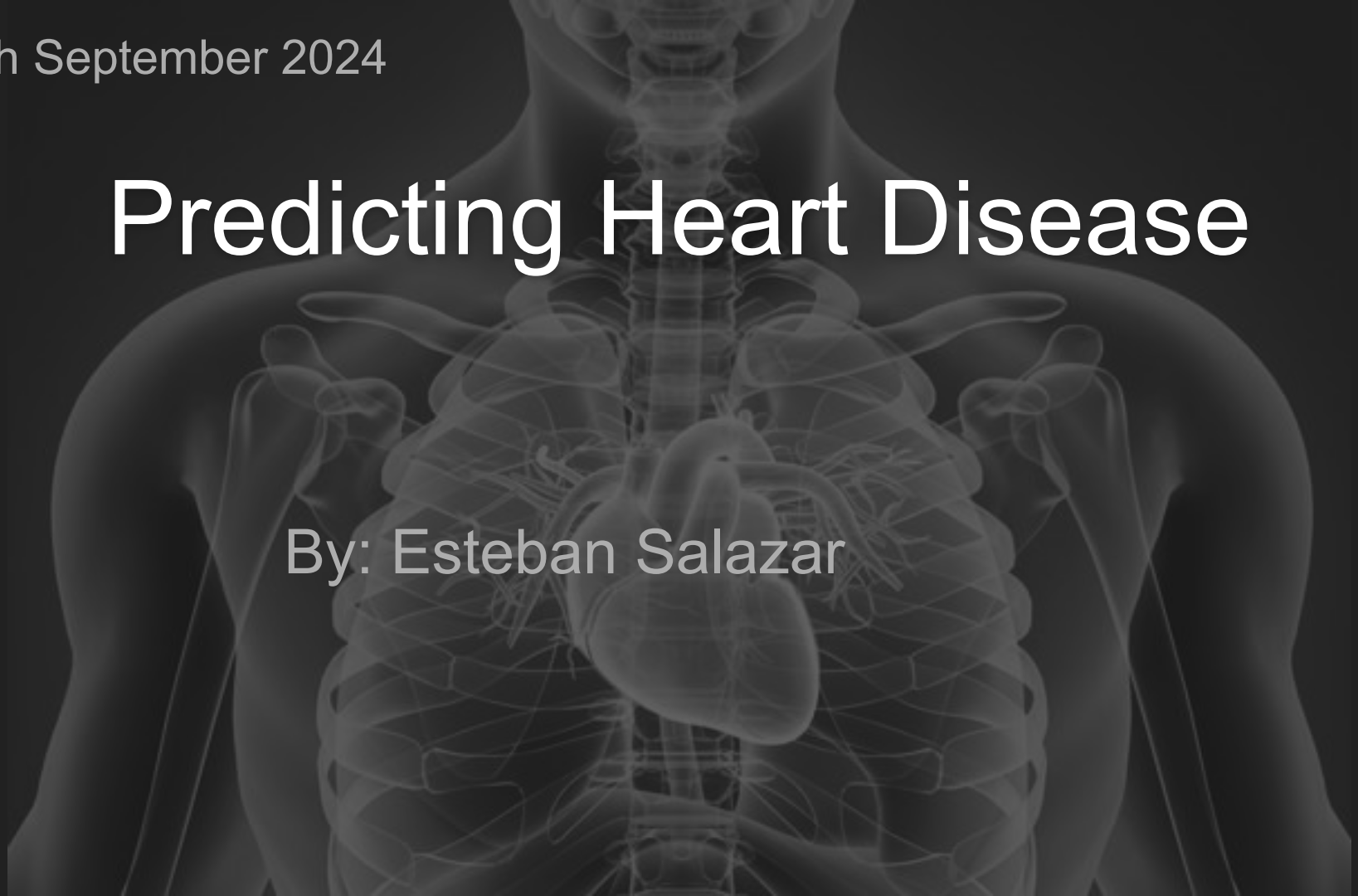
10th September 2024

# Predicting Heart Disease



By: Esteban Salazar

# Goals:

- ★ Perform data cleaning through two methods:
  - Iterative Regression Imputation
  - Mean Value Imputation
- ★ Perform exploratory data analysis
- ★ Evaluate the performance our predictive models through two criteria:
  - Test Error Rate
  - ROC Curve and Area Under the Curve
- ★ Implications

# Itinerary:

- ★ Description of data set
- ★ Methods used during the analysis
- ★ Visualize the exploratory data analysis
- ★ Evaluate the results
- ★ Discuss the findings

# Data Description

- 14 Variables total, 13 potential predictors 1 binary response

**Response**:

heartdisease, Heart disease (True or False)

**Quantitative Predictors (5 Total)**:

- age, Age in years
- trestbps, Resting blood pressure
- chol, Cholesterol
- thalach, Maximum heart rate obtained
- oldpeak, ST depression induced by exercise

# Data Description Part II

**Qualitative Predictors (5 Total)**:

- cp, Type of chest pain **4 levels**
- restecg, Resting electrocardiographic results **3 levels**
- slope, Slope of peak exercise segment **3 levels**
- ca, # of major vessels **4 levels**
- thal, normal, fixed defect, or reversible defect **3 levels**

**Binary Predictors (3 Total)**:

- sex, Male **1** or Female **0**
- fbs, Fasting blood sugar > 120ml/dl **1** for true and **0** for false
- exang, Exercise induced angina, **1** for true and **0** for false

# Overview of All Possible Methods

**Logistic Regression**

- No hyperparameters

**Bagging**

- One hyperparameter

**Random Forest**

- Two hyperparameters

**Gradient Boosting**

- Three hyperparameters

**KNN**

- One hyperparameter

**LDA**

- No hyperparameters

**QDA**

- No hyperparameters

**Support Vector Machine**

- Depends…

# Why These Methods?

- Logistic Regression
  - Specifically used in binary classification settings
- Random Forest
  - Random forest is a non-parametric approach
- Support Vector Machine
  - No distribution assumptions
  - Performance does not deteriorate with large number of predictors

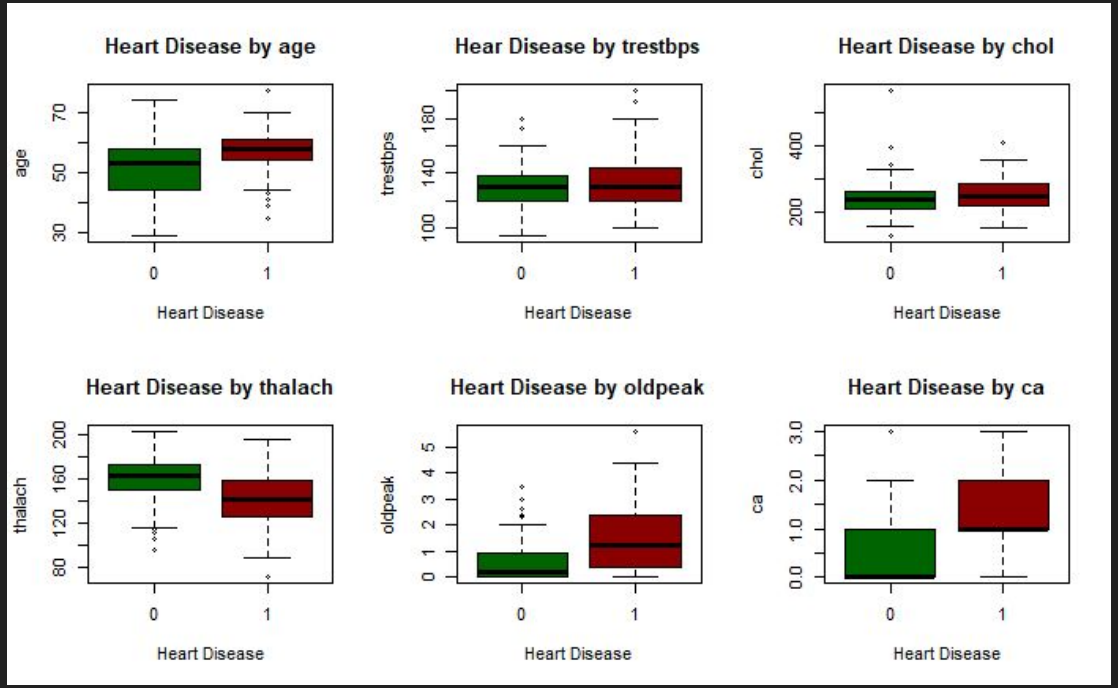
# Exploratory Data Analysis (EDA)

Sections:

- ★ EDA of mean imputed quantitative variables
- ★ EDA of mean imputed qualitative variables
- ★ EDA of iterative regression quantitative variables
- ★ EDA of iterative regression qualitative variables

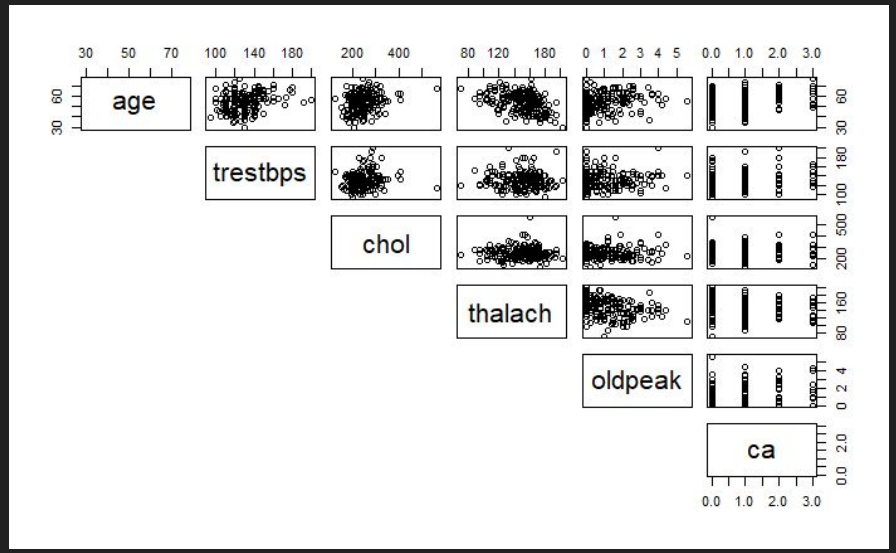
# EDA Mean Imputed Quantitative: I

Boxplots of the six quantitative variables

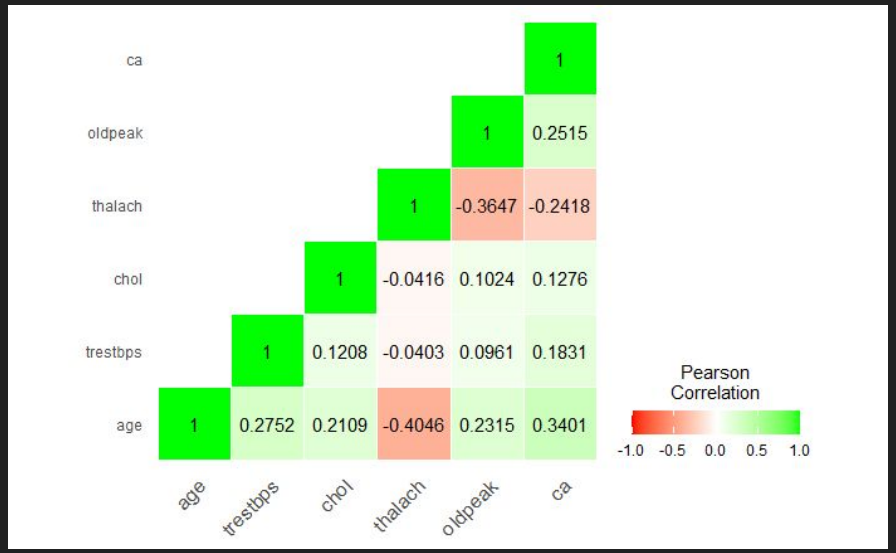


# EDA Mean Imputed Quantitative: II

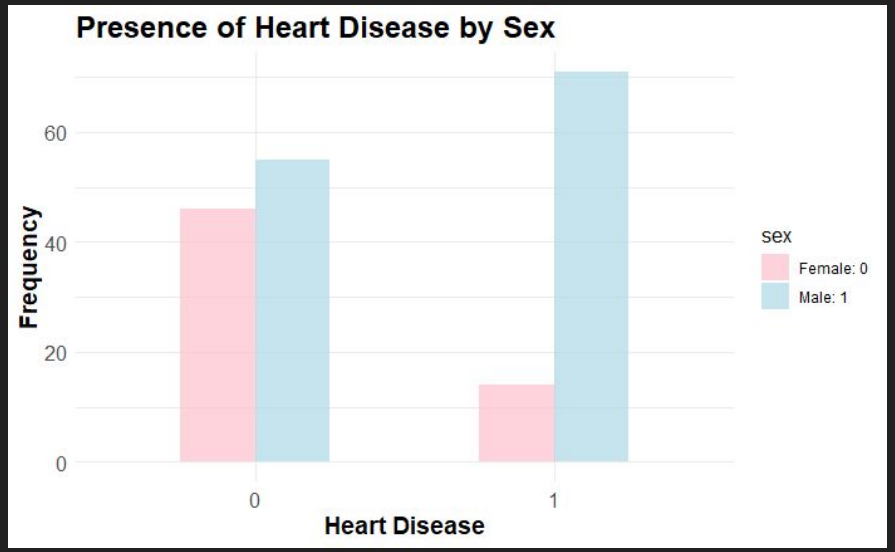
Scatterplot Matrix

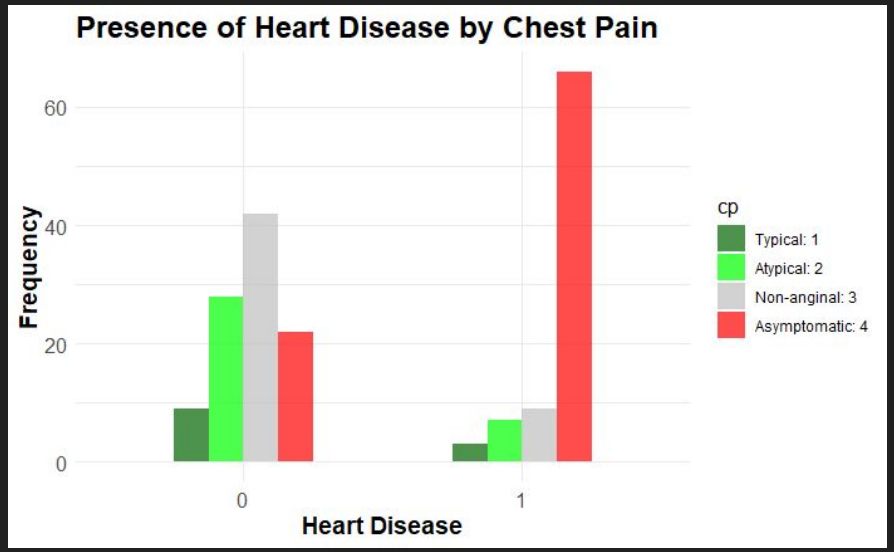



Pearson Correlation Matrix

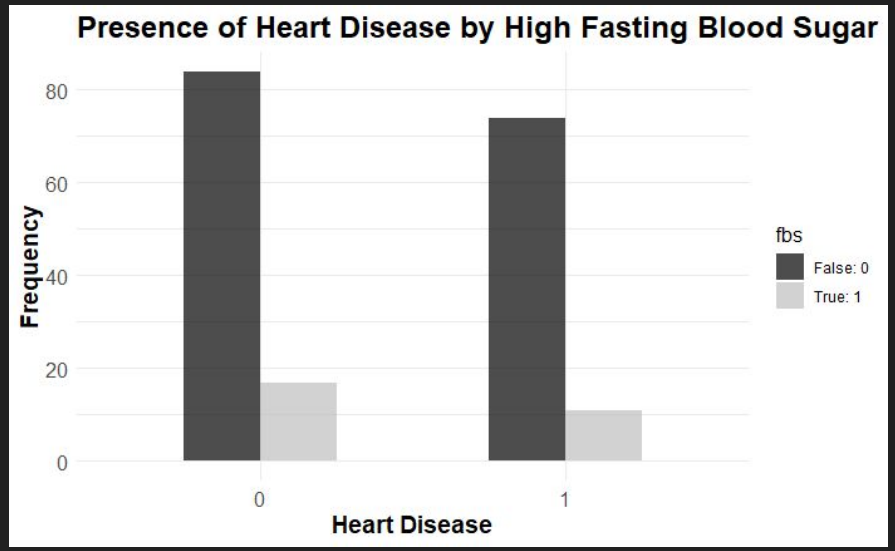


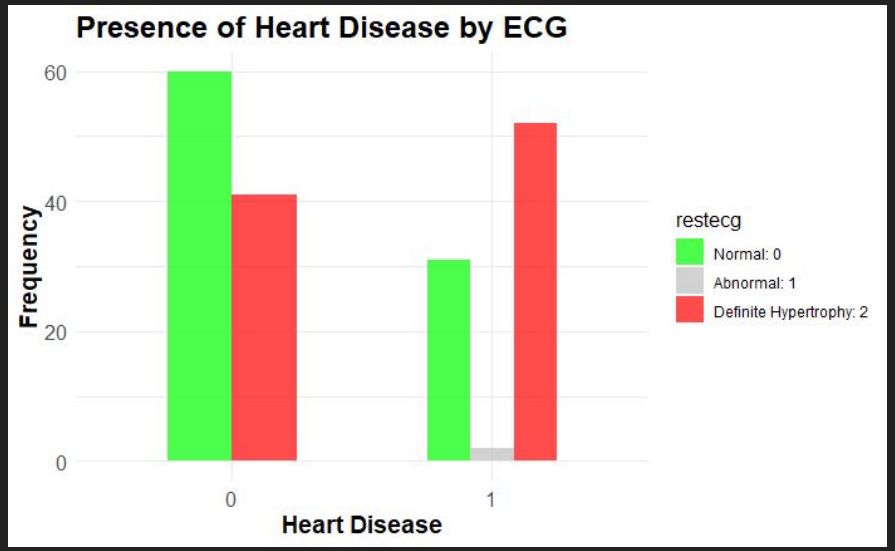
# EDA Mean Imputed Qualitative: I



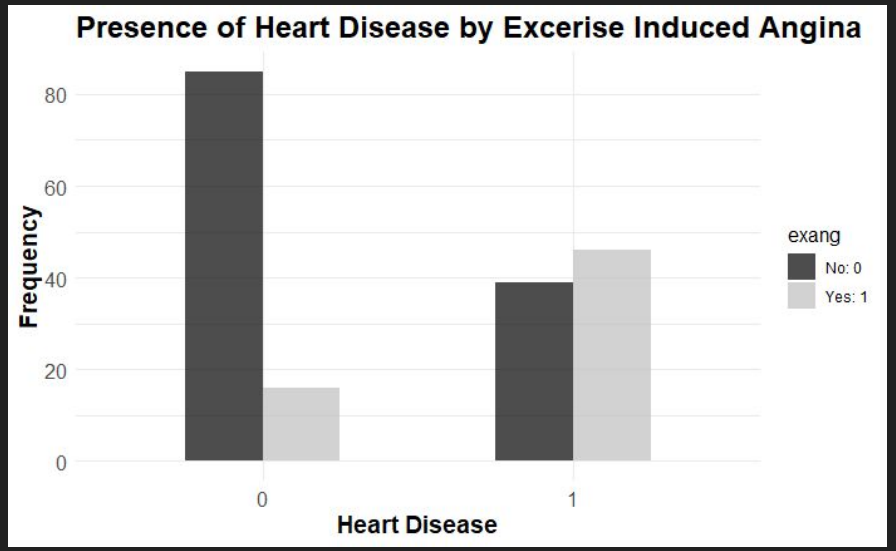



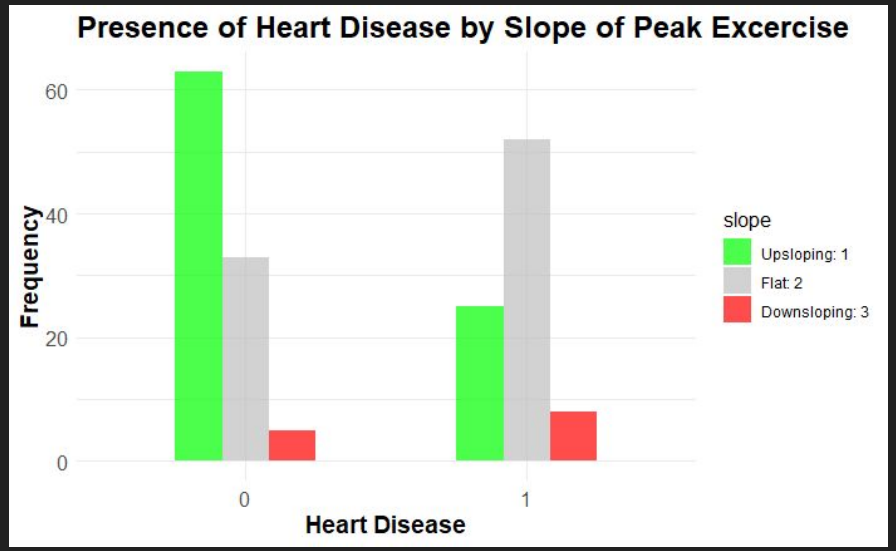
# EDA Mean Imputed Qualitative: II



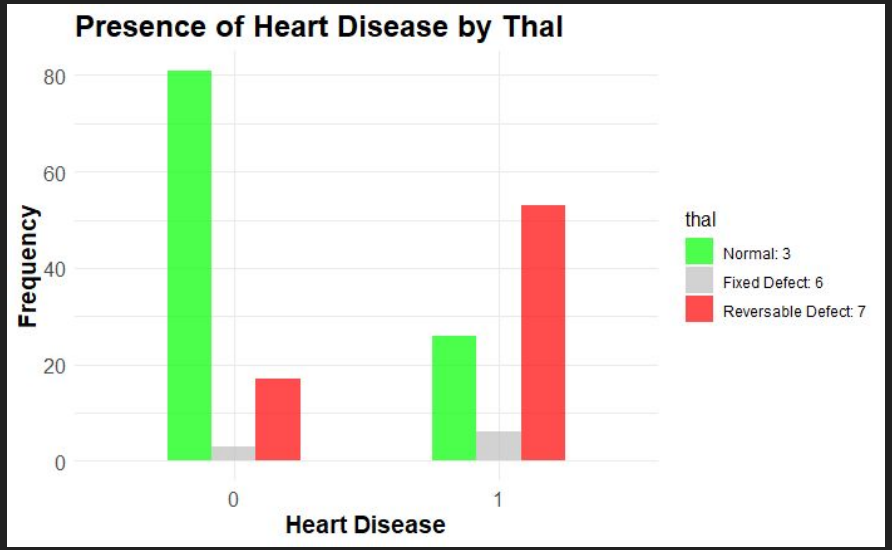



# EDA Mean Imputed Qualitative: III



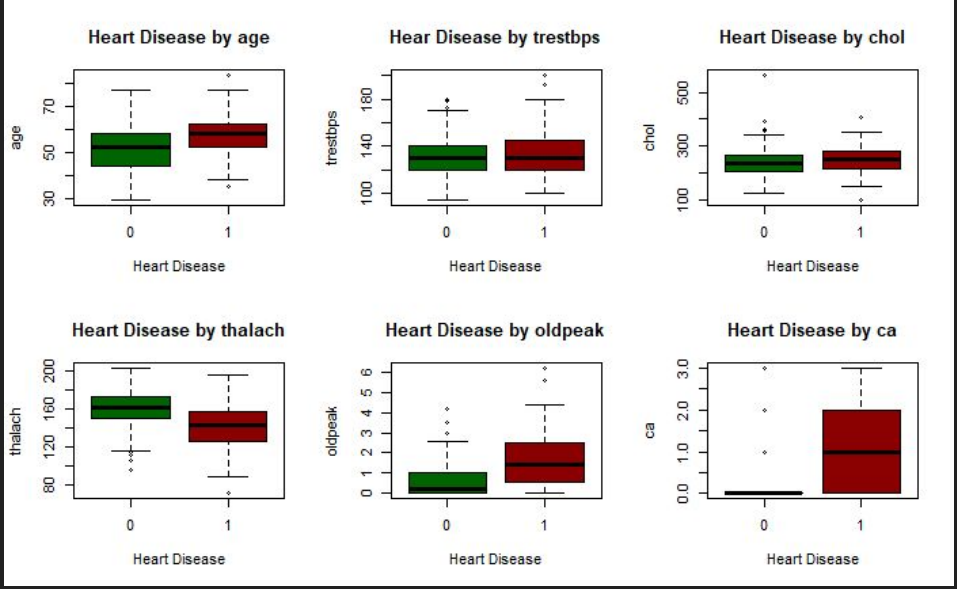



# EDA Mean Imputed Qualitative: IV



# EDA Iterative Regression Quantitative: I

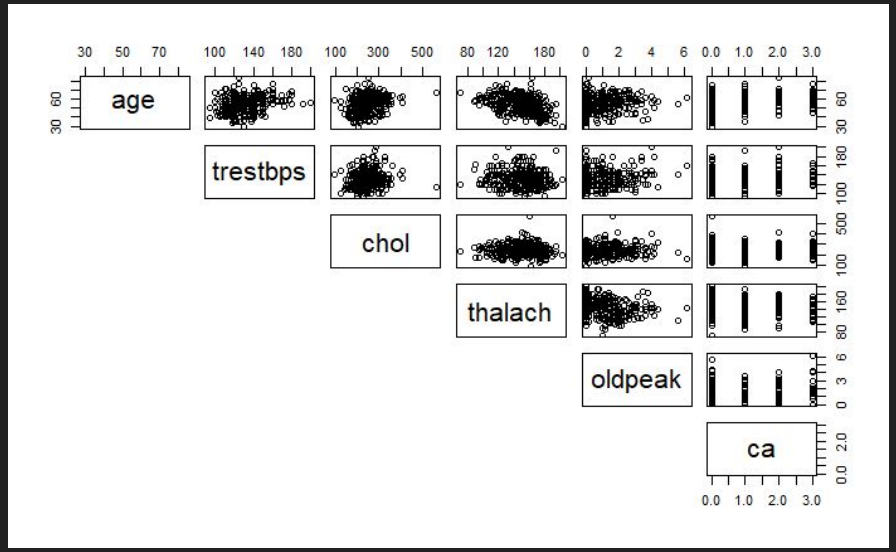
Boxplots of Iterative Regression



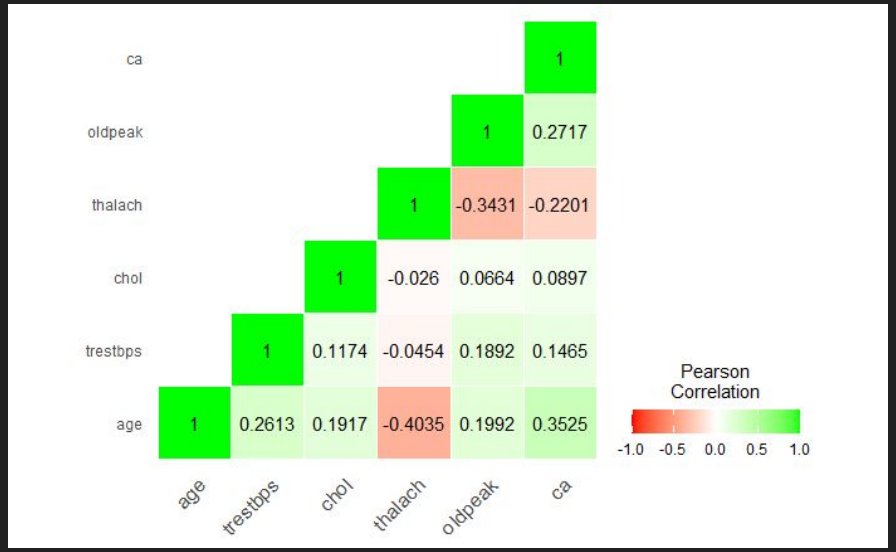

- For imputed data set, run iterative regression to replace NA values with values from a function

# EDA Iterative Regression Quantitative: II

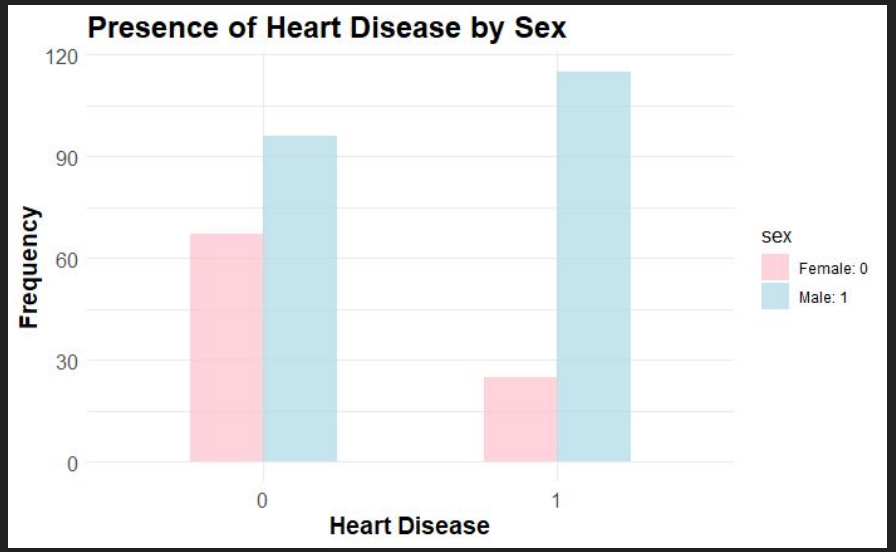
Scatterplot Matrix

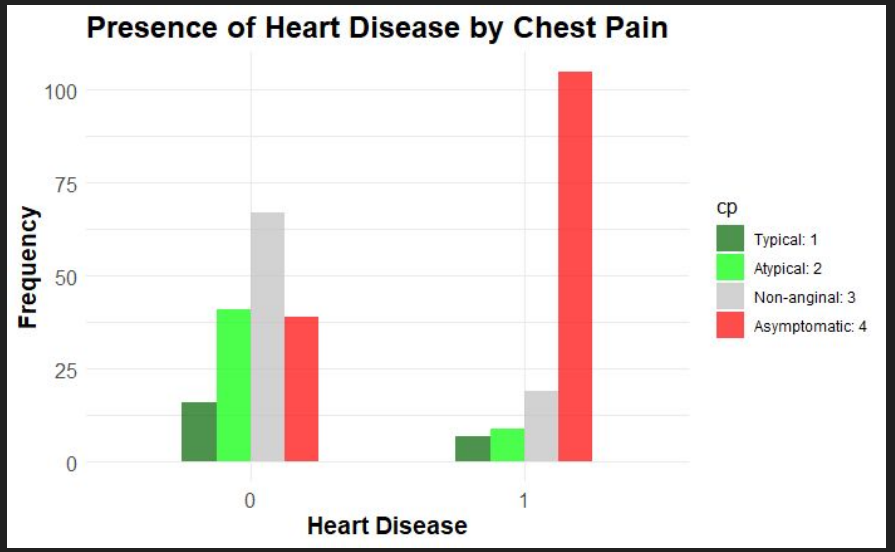



Pearson Correlation Matrix

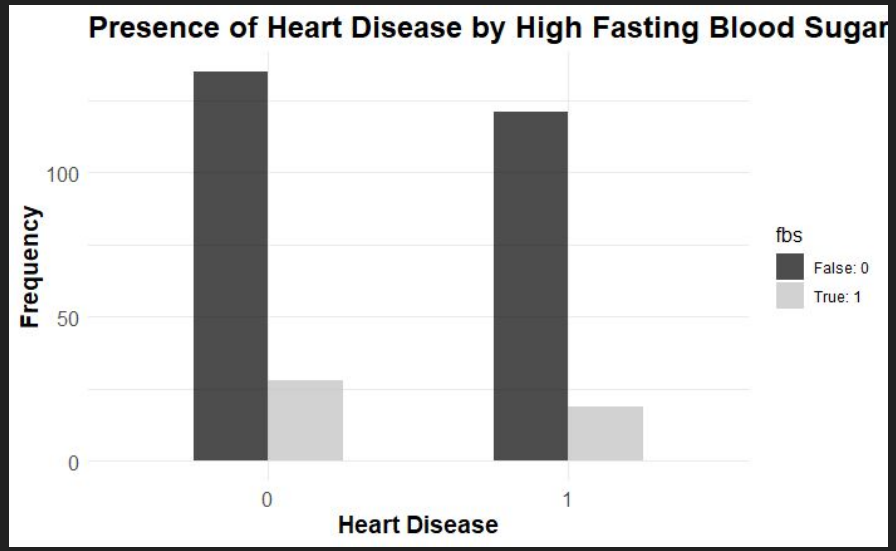


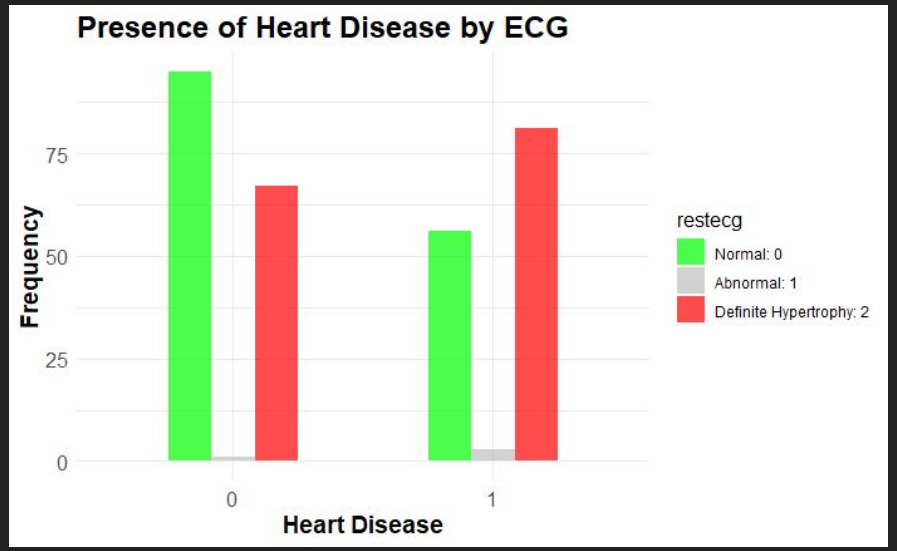
# EDA IR Imputed Qualitative: I



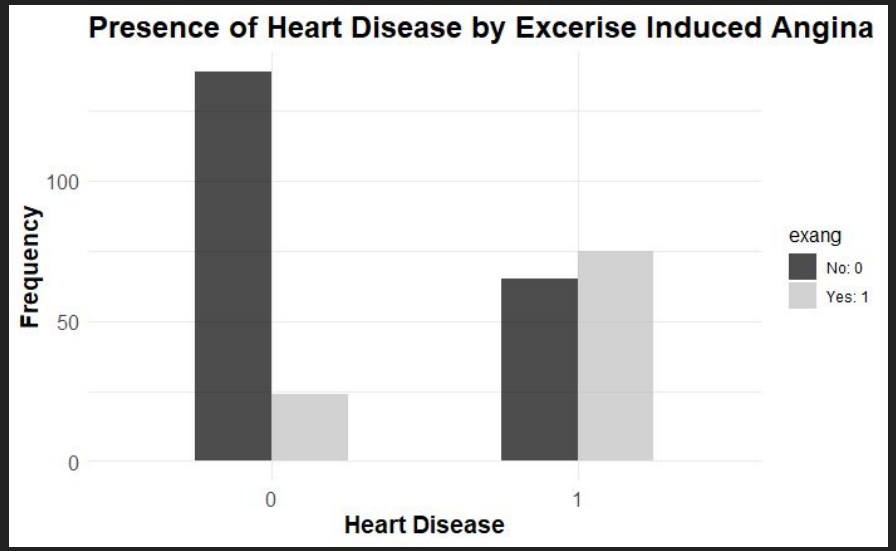



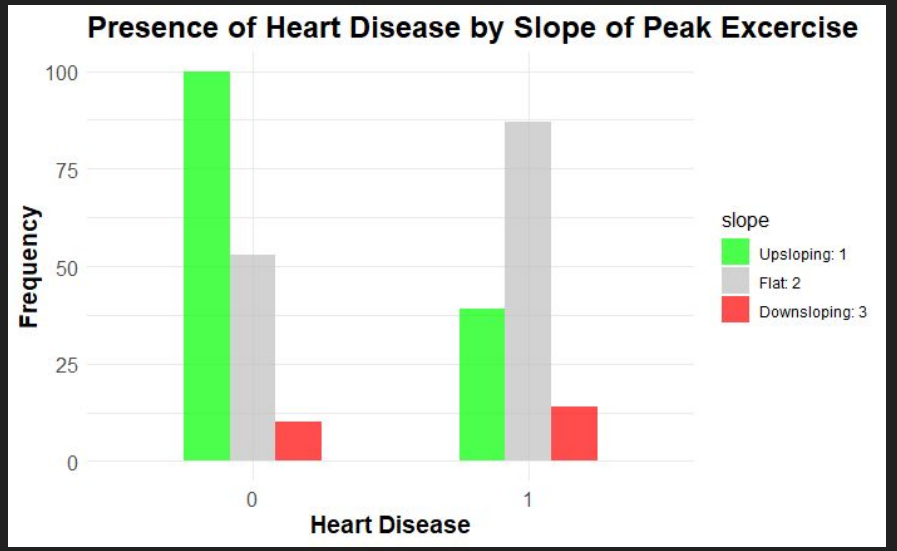
# EDA IRImputed Qualitative: II



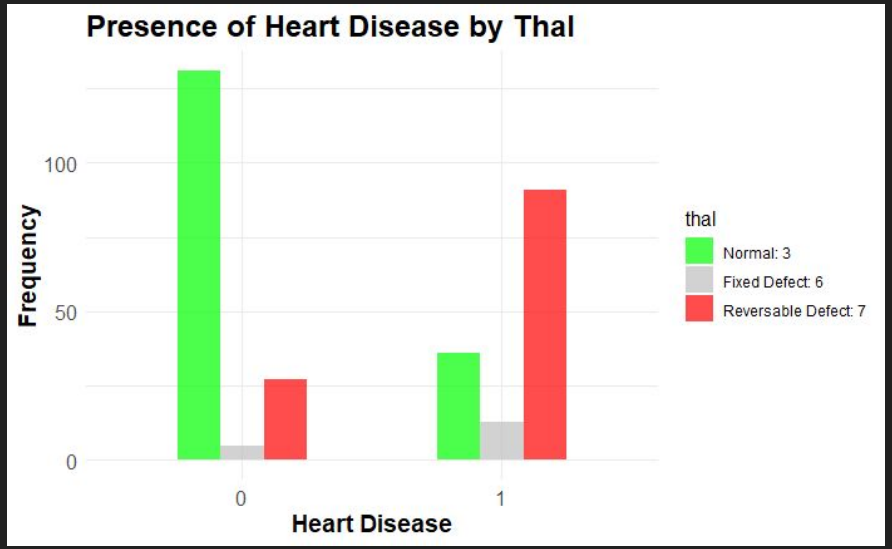



# EDA IR Imputed Qualitative: III



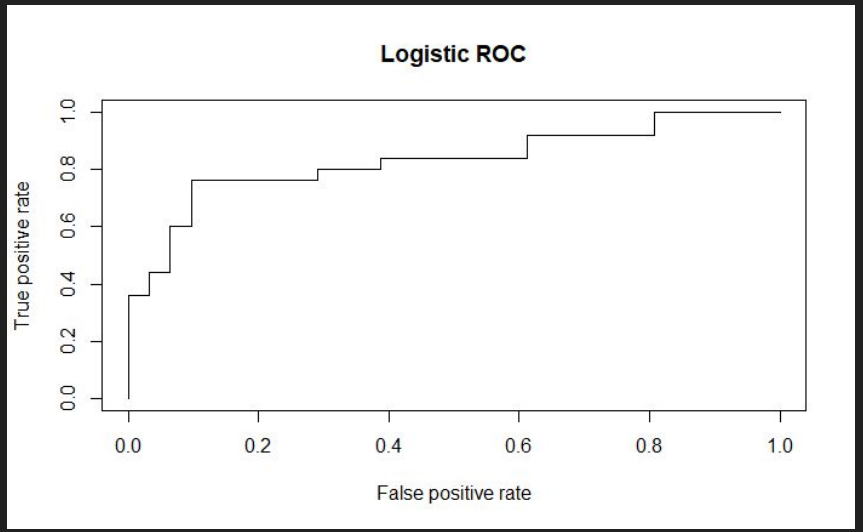



# EDA IR Imputed Qualitative: IV



# Logistic Regression (Mean Imputation)

ROC




Confusion Matrix

| Pred \ Test | 0 | 1 |
|---|---|---|
| 0 | 27 | 6 |
| 1 | 4 | 19 |

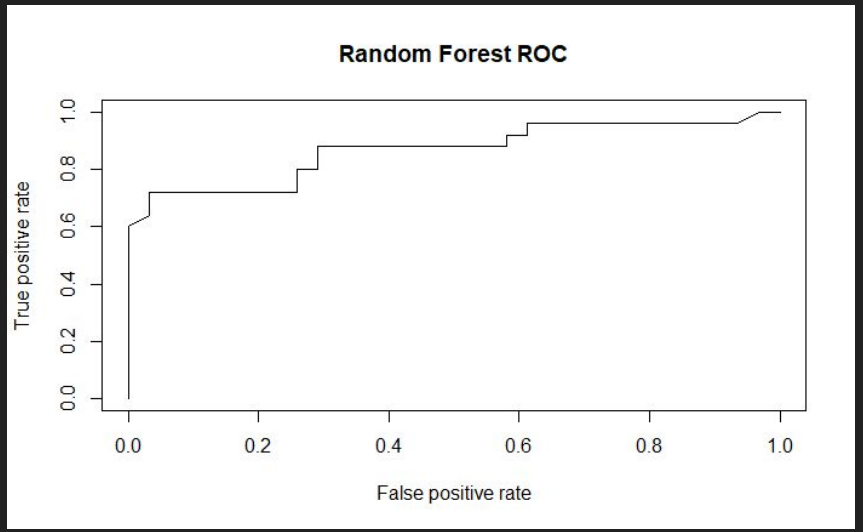
AUC: 0.8310 = 83.10%

Test Error Rate: 0.1786… ~ 17.86%

Tuning Parameters: mtry = 11, ntree = 200

# Random Forest (Mean Imputation)

## ROC




## Confusion Matrix

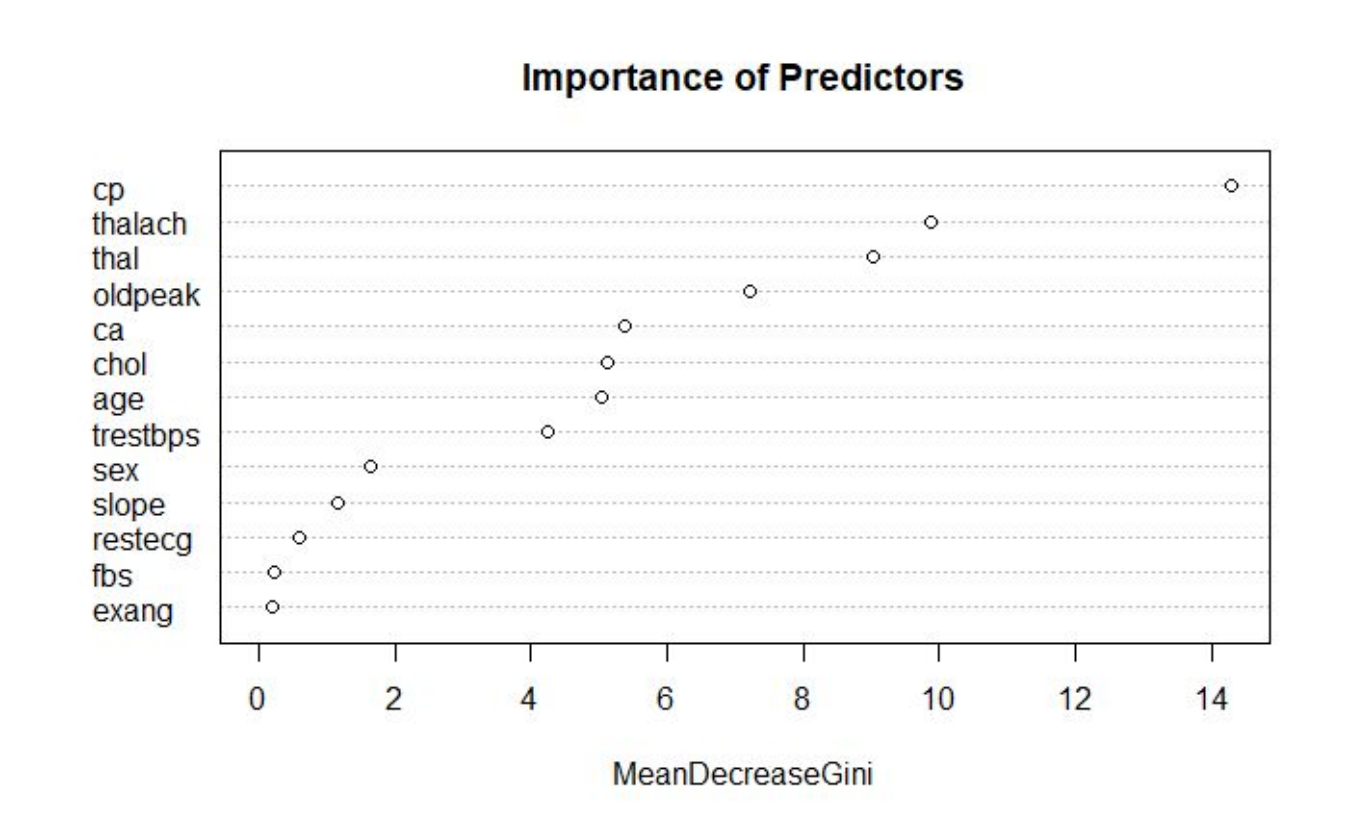
| | | Test | |
|---|---|---|---|
| | | **0** | **1** |
| Pred | **0** | 23 | 6 |
| | **1** | 1 | 13 |

AUC: 0.8671 = 86.71%

Test Error Rate: 0.2321= 23.21%

Tuning Parameters: mtry = 1, ntree = 200

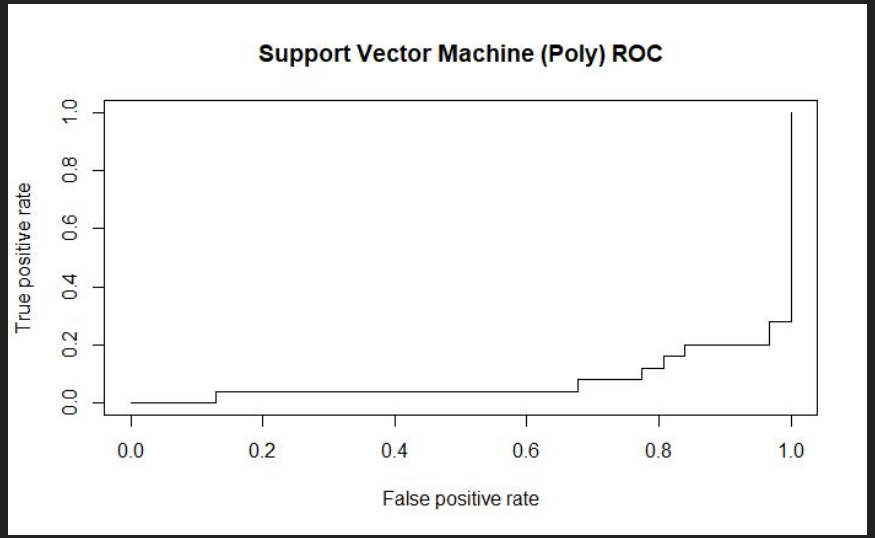
# Random Forest (Mean Imputation)Part II



Tuning Parameters: degree = 1, cost = 1

# Support Vector Machine (Mean Imputation)

## ROC




## Confusion Matrix

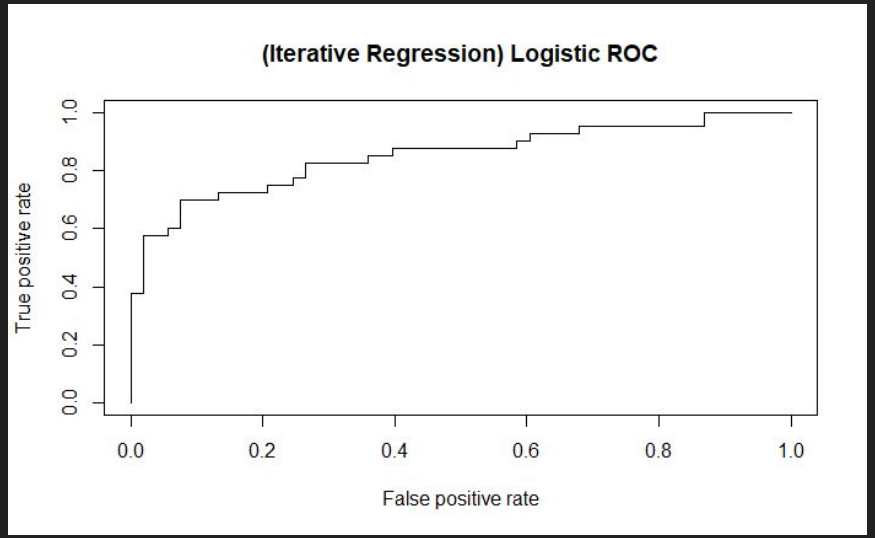
| Pred \ Test | 0 | 1 |
|---|---|---|
| 0 | 22 | 6 |
| 1 | 2 | 13 |

AUC: 0.0735 = 7.35%

Test Error Rate: 0.125 = 12.5%

# Logistic Regression (IR Imputed)

ROC




Confusion Matrix

| | | Test | |
|---|---|---|---|
| | | **0** | **1** |
| Pred | **0** | 40 | 9 |
| | **1** | 13 | 31 |

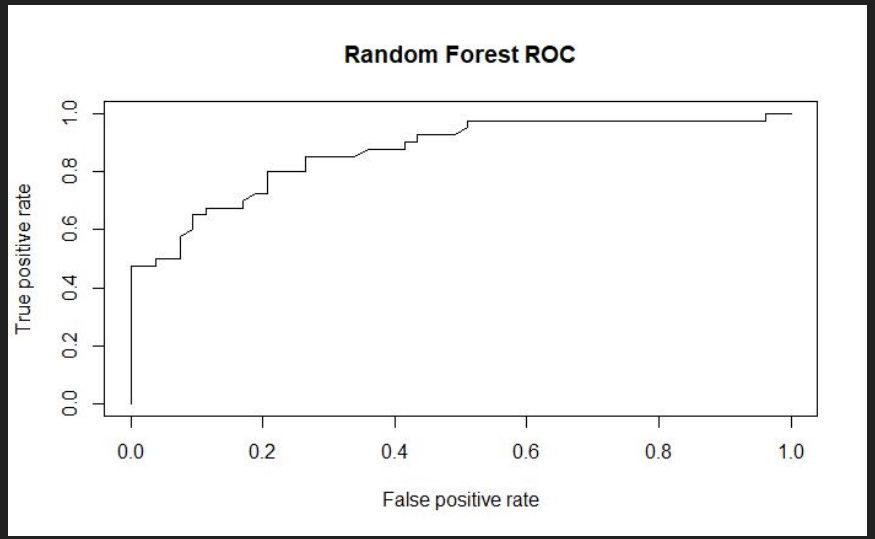
AUC: 0.8505… ~ 85.05%

Test Error Rate: 0.2366 = 23.66%

Tuning Parameters: mtry = 4, ntree = 300

# Random Forest (IR Imputed)

## ROC




## Confusion Matrix

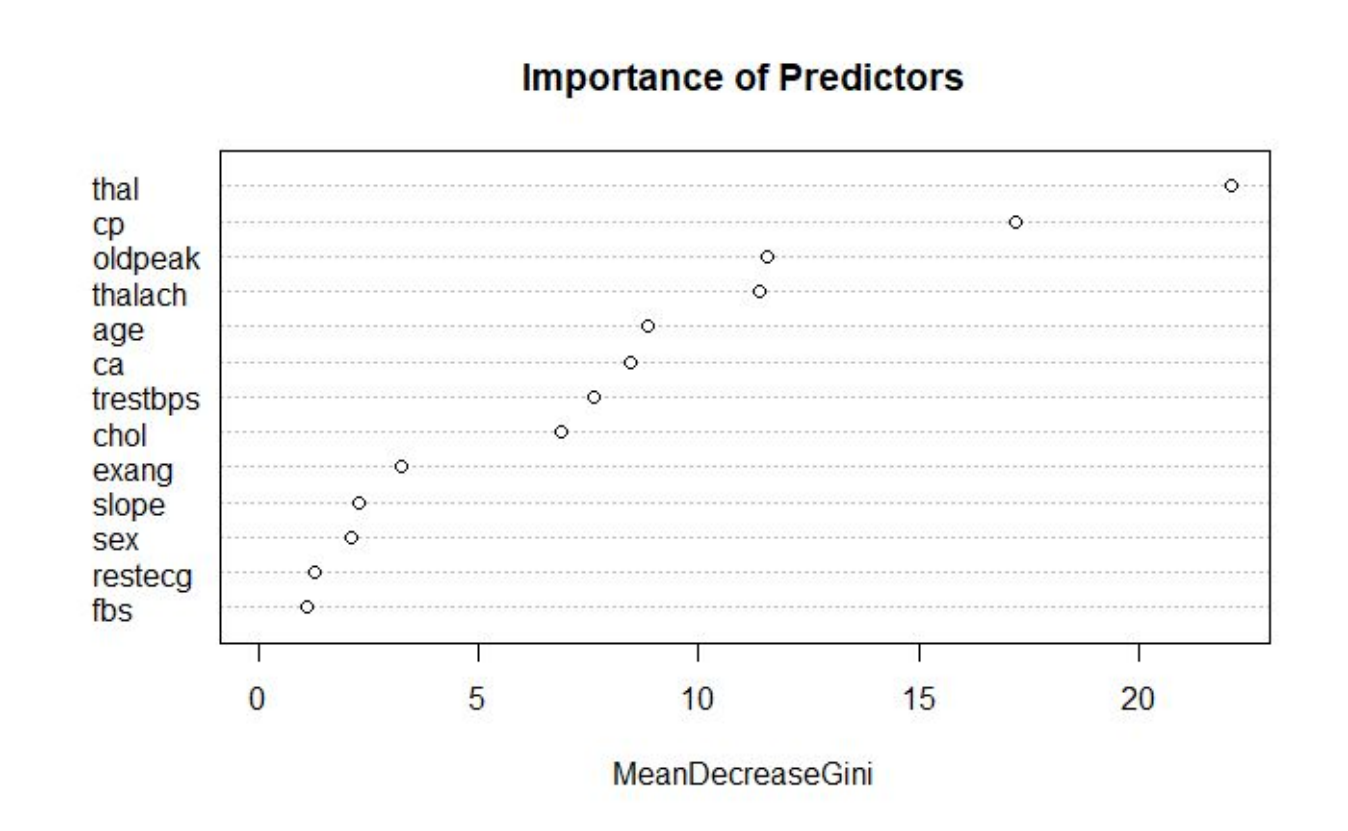
| Pred \ Test | 0 | 1 |
|---|---|---|
| 0 | 39 | 8 |
| 1 | 14 | 32 |

AUC: 0.8670… ~ 86.70%

Test Error Rate: 0.2366 = 23.66%

Tuning Parameters: mtry = 4, ntree = 500

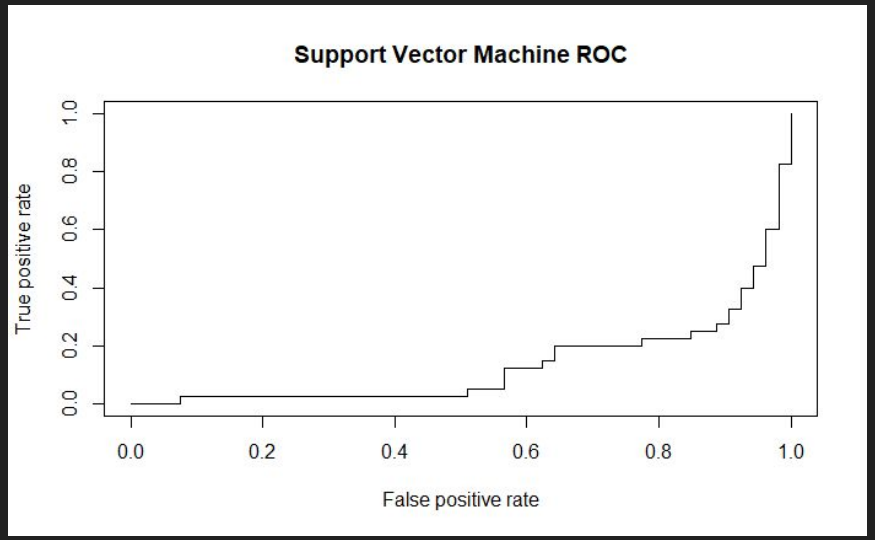
# Random Forest (IR Imputed) Part II



Tuning Parameters: degree = 1, cost = 10

# Support Vector Machine (IR Imputed)

## ROC




AUC: 0.1311… ~ 13.11%

## Confusion Matrix

| | | Test | |
|---|---|---|---|
| | | **0** | **1** |
| Pred | **0** | 40 | 9 |
| | **1** | 13 | 31 |

Test Error Rate: 0.2151=21.51%

# Main Differences With/Without Iterative Regression (IR)

- Number of observations available
  - Without IR: 145 of 303 total observations
  - With IR: 303 of 303 total observations

- Train/Test Data Split (70/30)
  - Without IR: For the test error rate, 43 test responses to compare to prediction.
  - With IR: For the test error rate, 90 test responses to compare to prediction.

# Relative Performance: AUC & Test Error Rate (TER)

Logistic Regression

- Both Imputed & Non-Imputed AUC performance was the ok

Random Forest

- Both Imputed & Non-Imputed AUC performance was best

Support Vector Machine

- Both Imputed & Non-Imputed AUC performance was the worst

# Thank you for your time!

## References


“Heart Disease Data Set.” UCI Machine Learning Repository: Heart Disease Data Set, https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R. Springer, 2022.