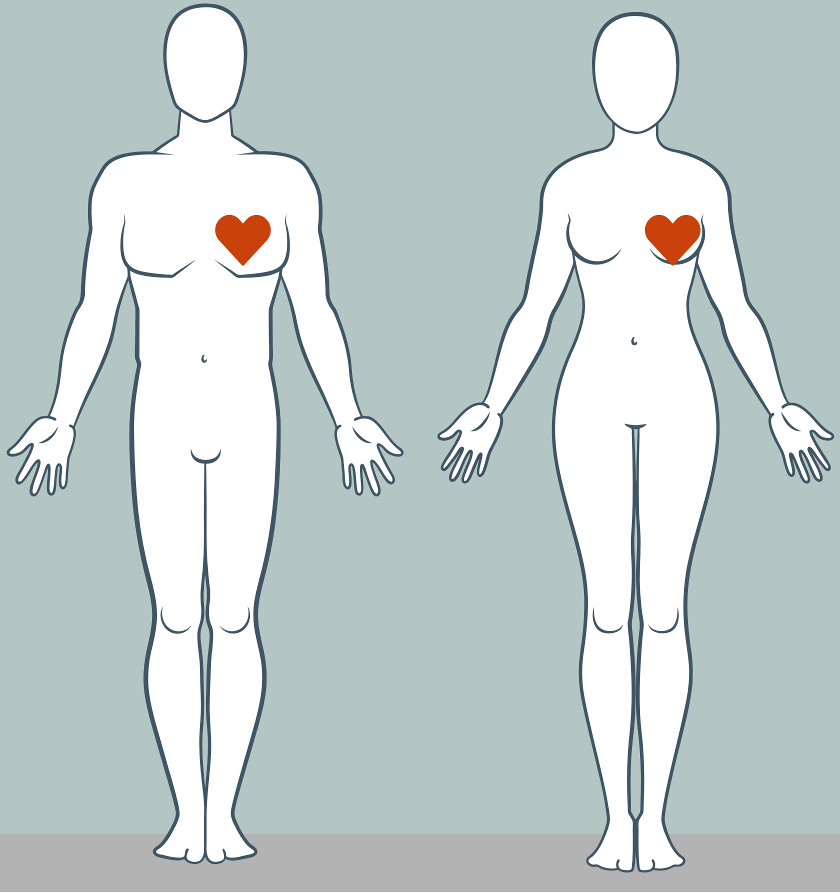
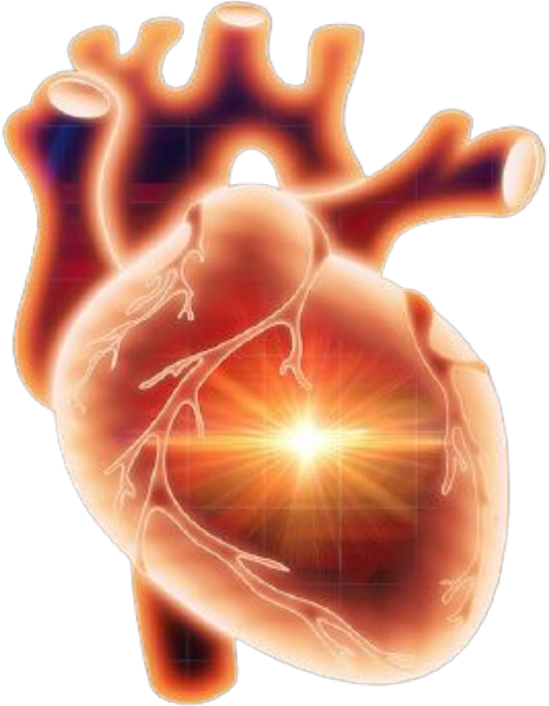
Predicting Heart Disease

Conner Byrd, Eric Han, Ki Hyun, Sara Shao, Alex Shen, Mona Su, Dani Trejo, Steven Yuan



Introduction



- Cardiovascular diseases (CVDs) are the #1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 32% of all global deaths.
- Currently, there are several different variables for physicians to diagnose patients that they believe to be at risk for CVDs.
- We set out to build two models, using the current variables that doctors take into consideration, in order to better understand and predict if a high-risk patient will develop CVDs.

Our Goals

Predictive Model

- Assess the likelihood of a possible heart disease event for potential at-risk patients

Interpretative Model

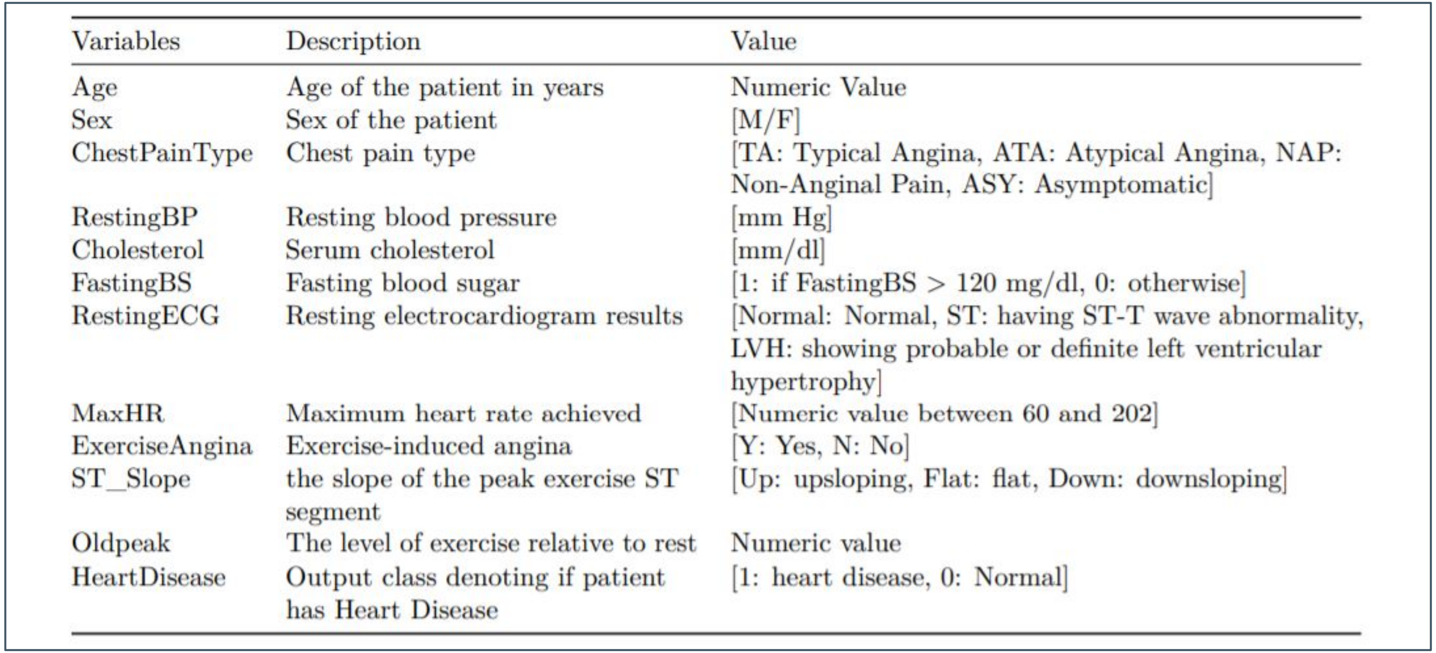
- Understand what factors increase a patient’s risk for CVDs
- Provide patients and doctors with a greater understanding of signs to look for if they are at high-risk for CVDs

Impact

- Our findings will be applicable to patients going to the hospital for check-ups on their hearts and who suspect they may be at high risk of developing CVDs

Data: Introduction

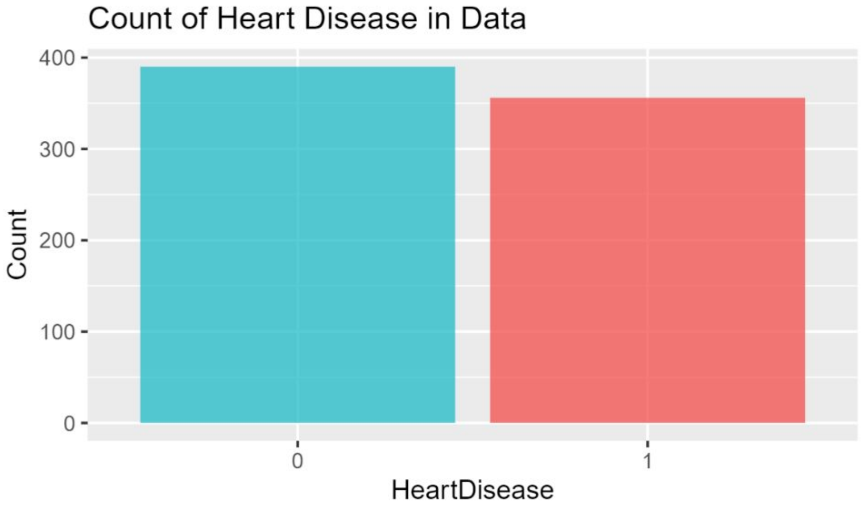
- Kaggle dataset, originally from US Naval Research Laboratory
- Patients from US, Switzerland, and Hungary



Variables	Description	Value
Age	Age of the patient in years	Numeric Value
Sex	Sex of the patient	[M/F]
ChestPainType	Chest pain type	[TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	Resting blood pressure	[mm Hg]
Cholesterol	Serum cholesterol	[mm/dl]
FastingBS	Fasting blood sugar	[1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	Resting electrocardiogram results	[Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy]
MaxHR	Maximum heart rate achieved	[Numeric value between 60 and 202]
ExerciseAngina	Exercise-induced angina	[Y: Yes, N: No]
ST_Slope	the slope of the peak exercise ST segment	[Up: upsloping, Flat: flat, Down: downsloping]
Oldpeak	The level of exercise relative to rest	Numeric value
HeartDisease	Output class denoting if patient has Heart Disease	[1: heart disease, 0: Normal]

Challenges

- Our data does not reflect a general population (roughly half have heart disease so our results can only be applicable to at-risk patients).
- We also see that there is 171 missing observations for the variable Cholesterol and 1 missing observation for Resting BP



Data: Missing Value Imputation

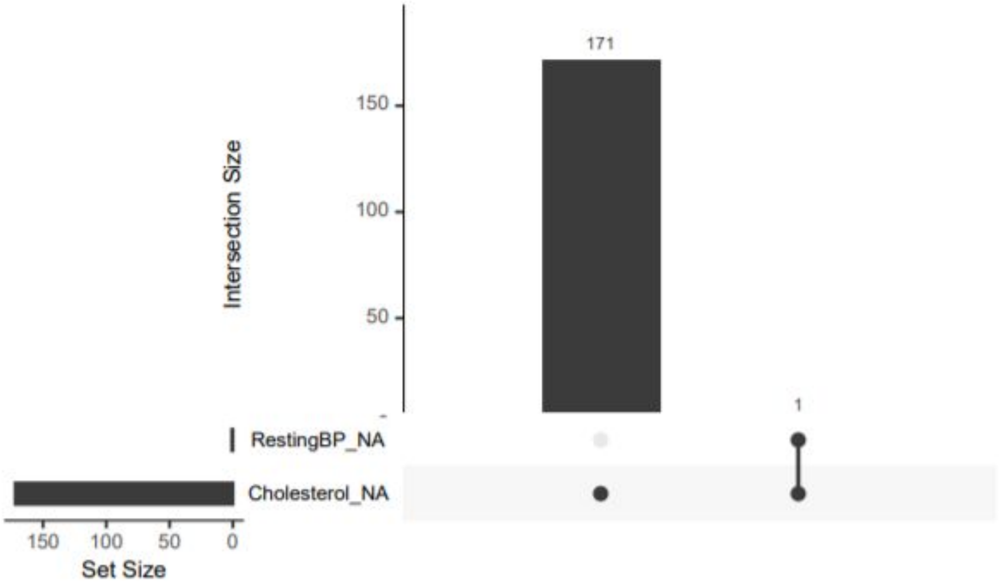
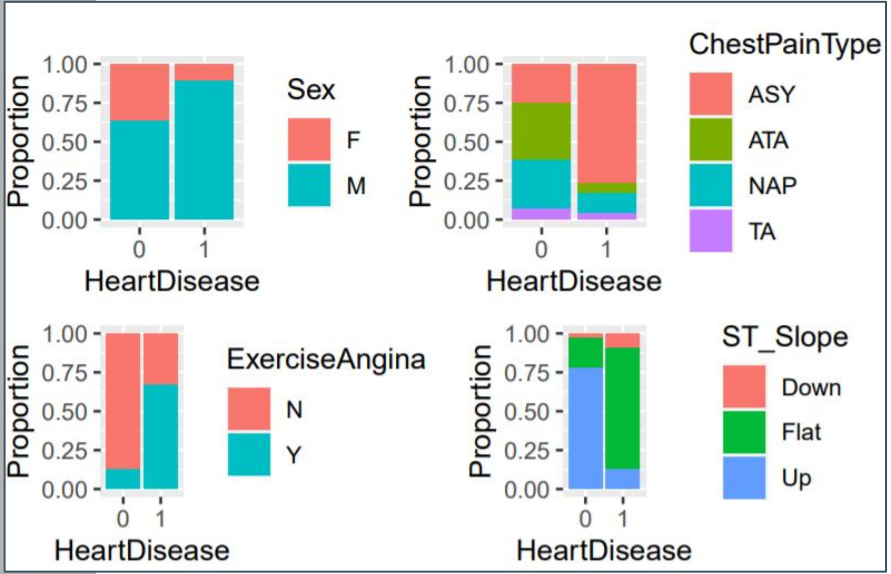


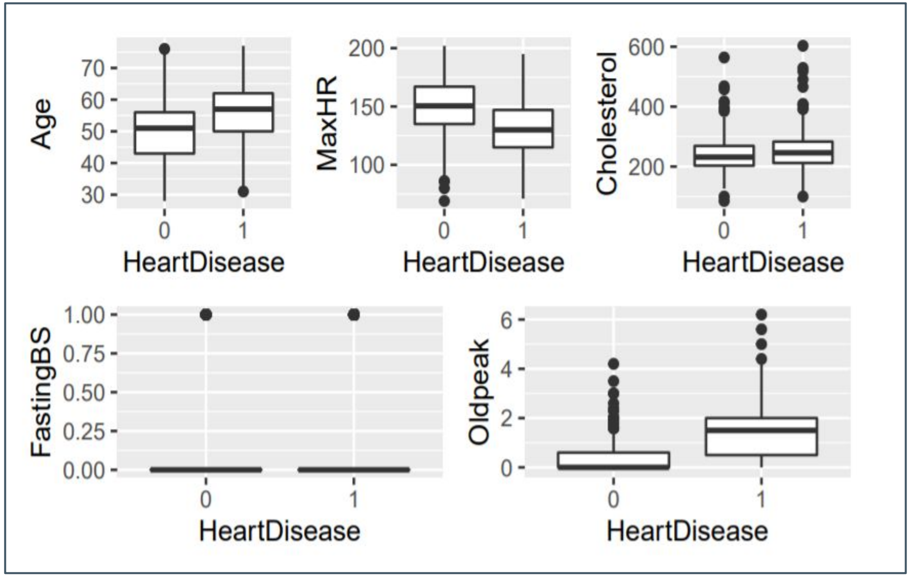
Figure 1: Missing Data Visualization

Data: Missing Value Imputation

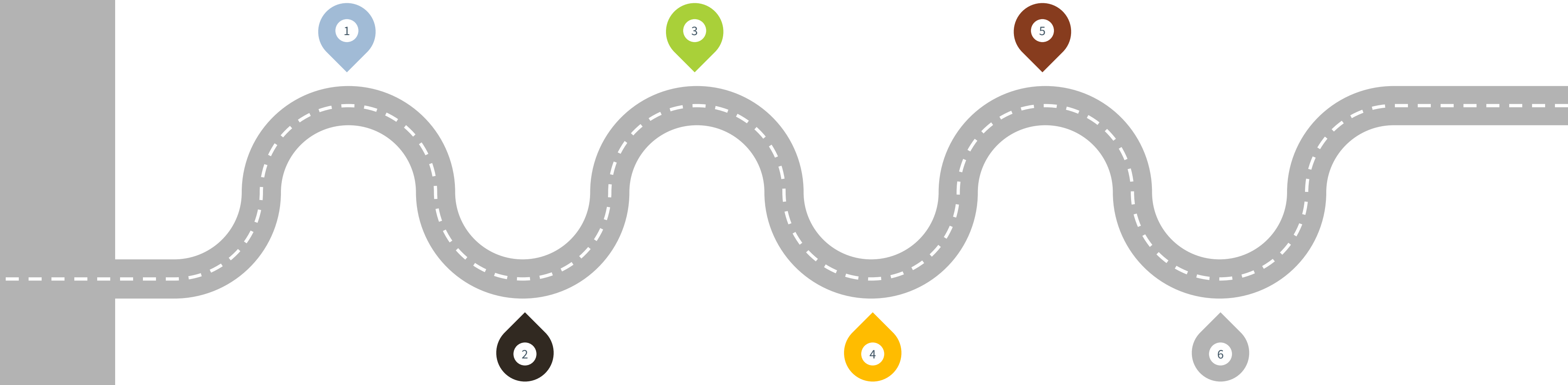
- Multiple Imputation using Chained Equation (MICE) in R was used to impute the values
- Five chains of imputations were conducted for Cholesterol with each chain using the default method of predictive mean matching (ppm) method since Cholesterol was a numerical variable.

Data: EDA


Proportion
1.00
0.75
0.50
0.25
0.00
0
1
HeartDisease
Sex
F
M
ChestPainType
ASY
ATA
NAP
TA
ExerciseAngina
N
Y
ST_Slope
Down
Flat
Up


Age
70
60
50
40
30
MaxHR
200
150
100
Cholesterol
600
400
200
FastingBS
1.00
0.75
0.50
0.25
0.00
Oldpeak
6
4
2
0
0
1
HeartDisease

Methodology Overview



Logistic Regression

Random Forest

Support Vector Machine (SVM)

Ensemble Model

Select Best Predictive Model

Compare and Analyze

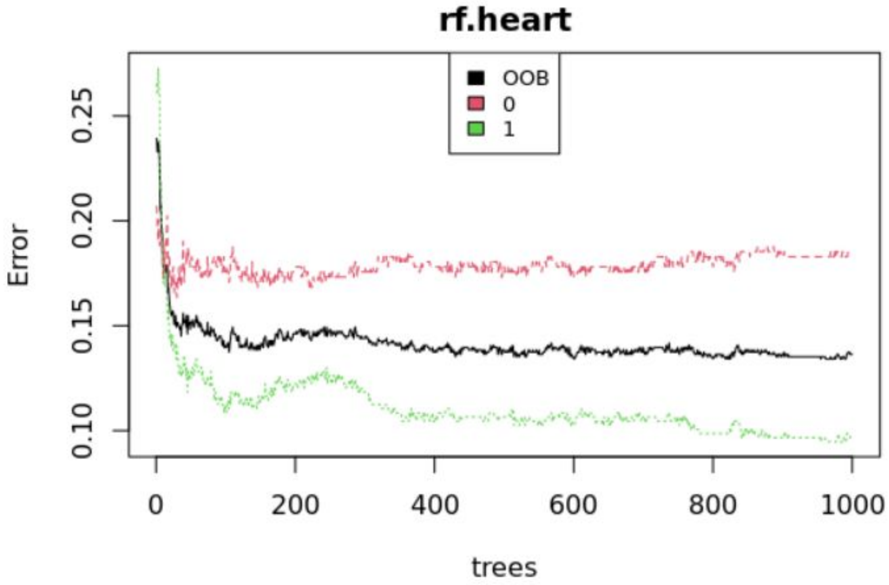
Logistic Regression Model

$$\log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$$

- Backwards stepwise selection using AIC
  - Starting full model included all main effects and all pairwise interactions
- Made sure that after selection, hierarchy principle is satisfied to aid in interpretation
- The threshold for prediction was set as 0.5

Random Forest

- Method: Average results across many trees
- Parameters:
  - subset size = 3
  - # of trees = 1000



SVM

- Method: Maximizes margin dist. while allowing for "budget" of misclassified points
- Tried 3 different kernels: linear, polynomial, and radial
  - Radial kernel performed the best
  - Cost = 1.78 (determined using 5-fold CV)

5-fold CV

- Each fold represents randomly selected 20% of testing data and 80% of training data (folds are independent)
- Using training data, get optimal models for each MICE chain for logistic, Random Forest, SVM and aggregate models
- Get prediction values from each MICE chain and optimal model
- Average the prediction values across 5 chains
- Get misclassification rate for each model
- Average the misclassification rate over 5 folds

Results: Best Predictive Model

Logistic Regression	0.141
Random Forest	0.125
SVM	0.129
Combined Model	0.129

The best predictive model was the random forest with an error rate of 12.5%.

Results: Logistic Regression and Important Factors

- Significant term*, Coefficient
  - SexM, 5.34
  - ChestPainTypeATA, -5.62
  - ChestPainTypeNAP, -3.74
  - ChestPainTypeATA:Cholesterol, 0.02
  - ChestPainTypeNAP:RestingECGST, 2.75

* only p-values < 0.01 included here, one chain of imputations used

Results: Random Forest

- Out-of-bag error rates:
  - Misclassification rate: 13.6%
  - False positive rate: 18.3%
  - False negative rate: 9.9%
- Limitations: Cannot discern important factors or likelihoods

Results: SVM

- Misclassification error rate: 12.9%

- Limitations: Cannot discern important factors or likelihoods

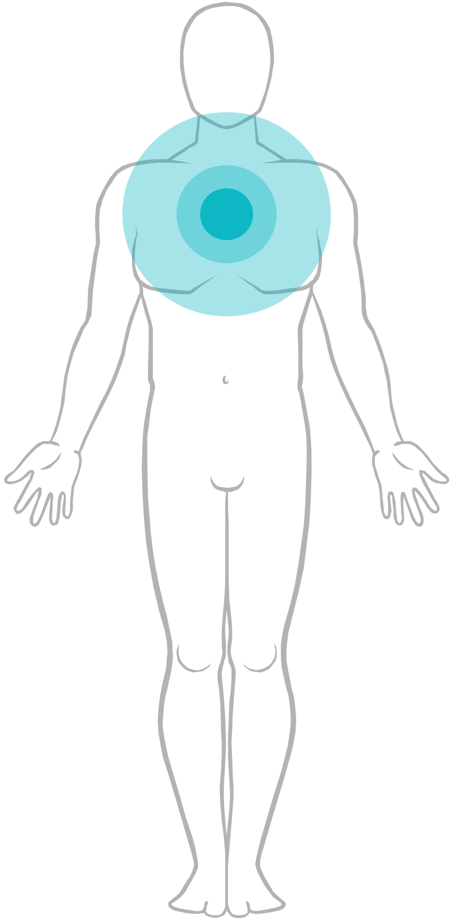
Conclusion

Prediction

Our methods yielded a heart disease prediction model with an 87.5% accuracy rate for at-risk patients. Although this model cannot replace a diagnosis or test, it can give doctors and patients an informal assessment of risk.

Inference

Patient sex and chest pain type are important predictors of CVD. Men are more likely to have CVD than women and people with type ASY chest pain are more likely to have CVD than people with ATA or NAP type chest pain.



Limitations and Further Work

- Dataset is not representative of the general population
  - Can only generalize to people already high risk for CVDs seeking medical testing
- Further research into Type I or Type II error
  - False negatives are considered more expensive.
  - Adjust cutoffs in logistic regression and trees to reflect “cost” of each type of error

References

“Cardiovascular Diseases (Cvds).” *World Health Organization*, World Health Organization, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

“Cardiovascular Diseases (Cvds).” *World Health Organization*, World Health Organization, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

“EKG Can Show False Positive Readings for Diagnosing Heart Condition.” *ScienceDaily*, ScienceDaily, 17 Nov. 2009, https://www.sciencedaily.com/releases/2009/11/091116103435.htm.

“Ensemble Modeling.” *Ensemble Modeling - an Overview | ScienceDirect Topics*, https://www.sciencedirect.com/topics/computer-science/ensemble-modeling.

Fedesoriano. “Heart Failure Prediction Dataset.” *Kaggle*, 10 Sept. 2021, https://www.kaggle.com/fedesoriano/heart-failure-prediction.

“Multivariate Imputation by Chained Equations [R Package Mice Version 3.14.0].” *The Comprehensive R Archive Network*, Comprehensive R Archive Network (CRAN), 24 Nov. 2021, https://cran.r-project.org/web/packages/mice/index.html.

NHS Choices, NHS, https://www.nhs.uk/conditions/cardiovascular-disease.