Capstone Project 1
Heart Disease Classifier

Brian Lubeck

December 9, 2017

Table of Contents

Introduction

- Heart disease is the build up of plaque in a person's arteries.
- It is the leading cause of death in the United States.
- More than 600,000 Americans die from it each year.
- Fortunately, heart disease is treatable.
- Four ways to treat it:
  - Eating a healthy diet
  - Exercising regularly
  - Maintaining a healthy weight
  - Taking medication

Introduction

- Goal: To accurately diagnose if a person has heart disease (0 = No, 1 = Yes)
- Built three classifiers to diagnose heart disease:
  - Naive Bayes
  - Logistic Regression
  - Random Forest
- Dataset: Cleveland heart disease dataset from the UC, Irvine's Machine Learning Data Repository

Data

Overview

- Dataset consists of medical information collected from patients at the Cleveland Clinic Foundation from 1988.
- 14 fields and 303 observations.
- Field of primary interest was NUM
- It is the degree to which the diameter of a person's arteries has narrowed. Five values 0 to 4:
  - $0 =$ less than 50% narrowing
  - $1, 2, 3$ and $4 =$ greater than 50% narrowing.
- Dependent variable was heart disease HD $= 1$ if NUM ≥ 1.

Data

Independent Variables I

1. AGE = Age in years
2. GENDER = Gender (1 = male; 0 = female)
3. CP = Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4. TRESTBPS = Resting blood pressure in mm Hg
5. CHOL = Serum cholestoral in mg/dl
6. FBS = Fasting blood sugar (1 if $>$ 120 mg/dl; 0 otherwise)
7. RESTECG = Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

Data

Independent Variables II

8. THALACH = Maximum heart rate achieved during thallium stress test
9. EXANG = Exercise induced angina (1 = yes; 0 = no)
10. OLDPEAK = ST depression induced by exercise relative to rest
11. SLOPE = Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
12. CA = Number of major vessels (0-3) colored by flourosopy
13. THAL = thallium stress test results (3 = normal; 6 = fixed defect; 7 = reversible defect)

Data

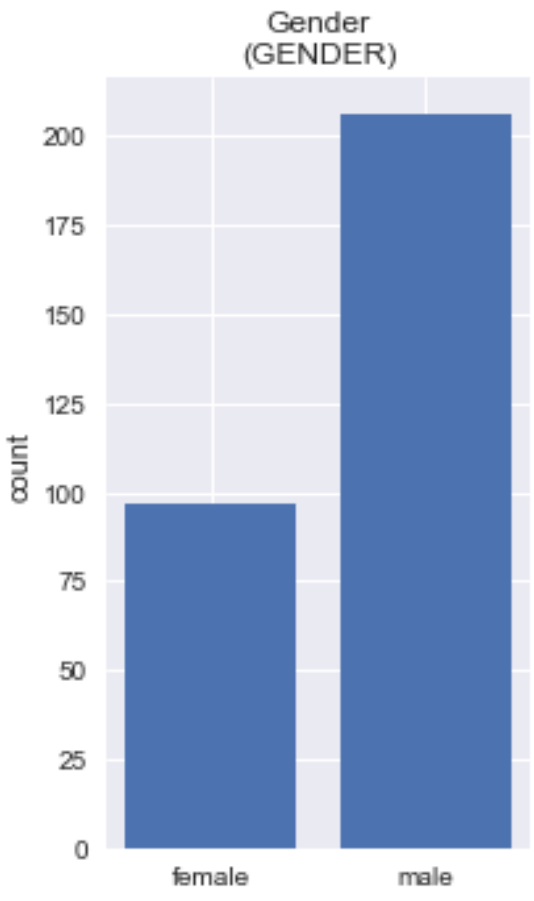
Continuous Distributions

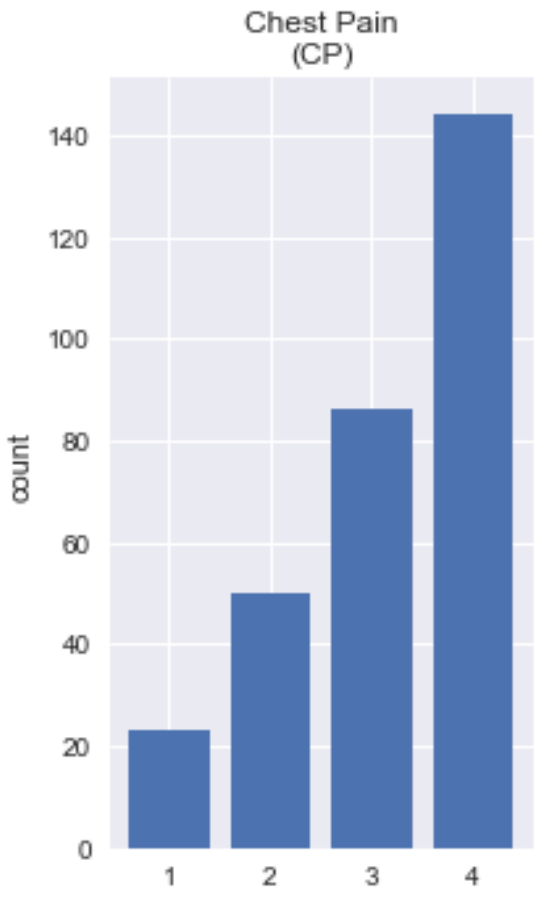
Quantiles of Continuous Data Fields

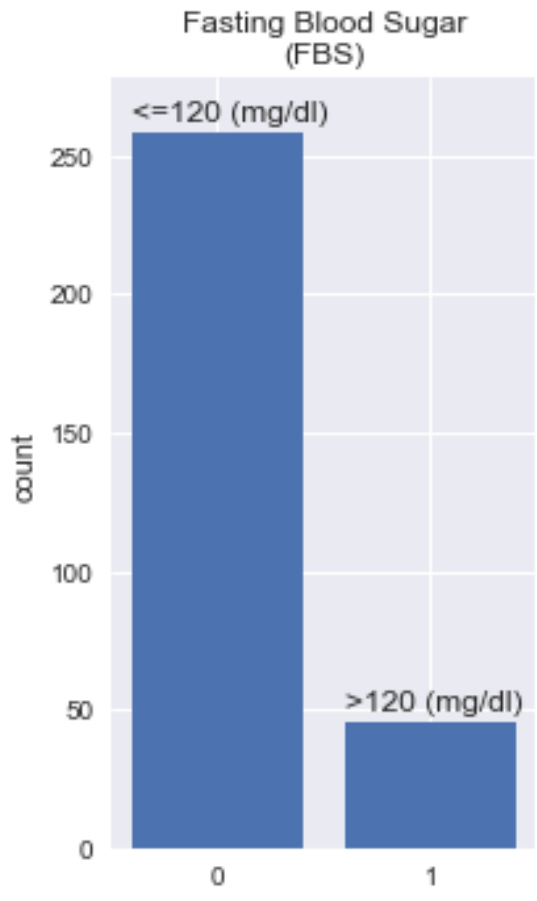
Field Name	Min	10th	25th	50th	75th	90th	Max
AGE	29.00	42.00	48.00	56.00	61.00	66.00	77.00
TRESTBPS	94.00	110.00	120.00	130.00	140.00	152.00	200.00
CHOL	126.00	188.80	211.00	241.00	275.00	308.80	564.00
THALACH	71.00	116.00	133.50	153.00	166.00	176.60	202.00
OLDPEAK	0.00	0.00	0.00	0.80	1.60	2.80	6.20

Data

Discrete Distributions

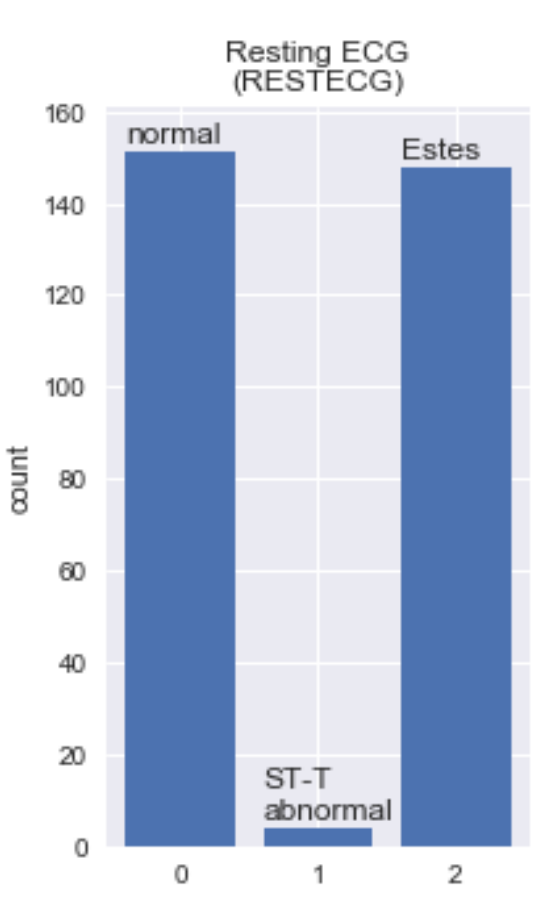


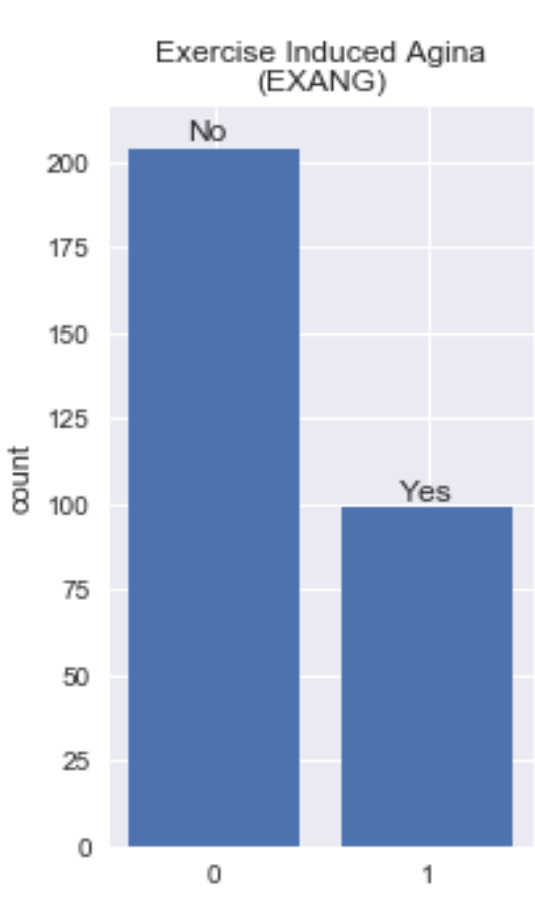


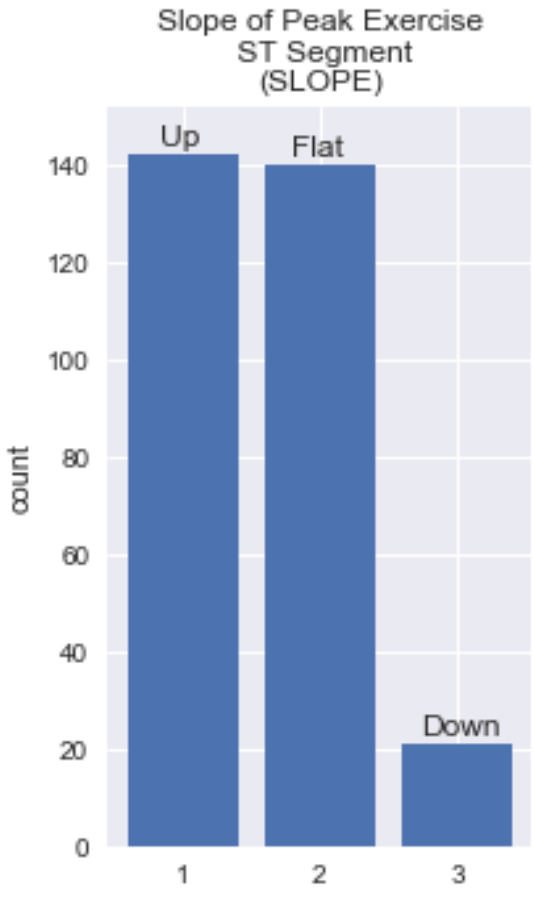


Data

Discrete Distributions

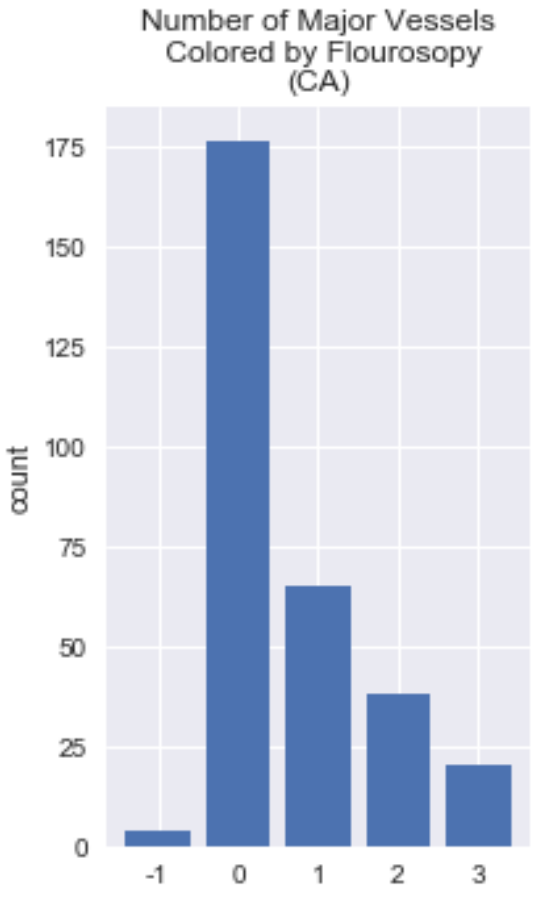


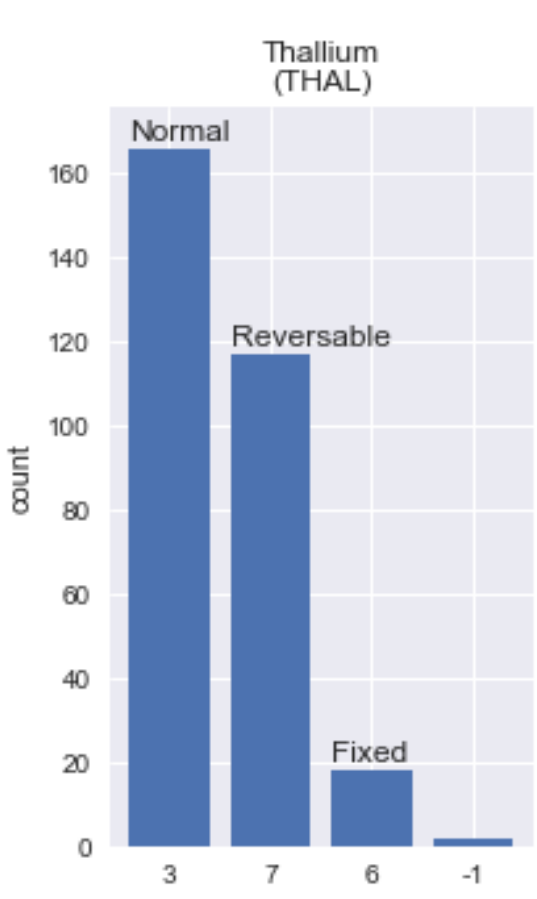


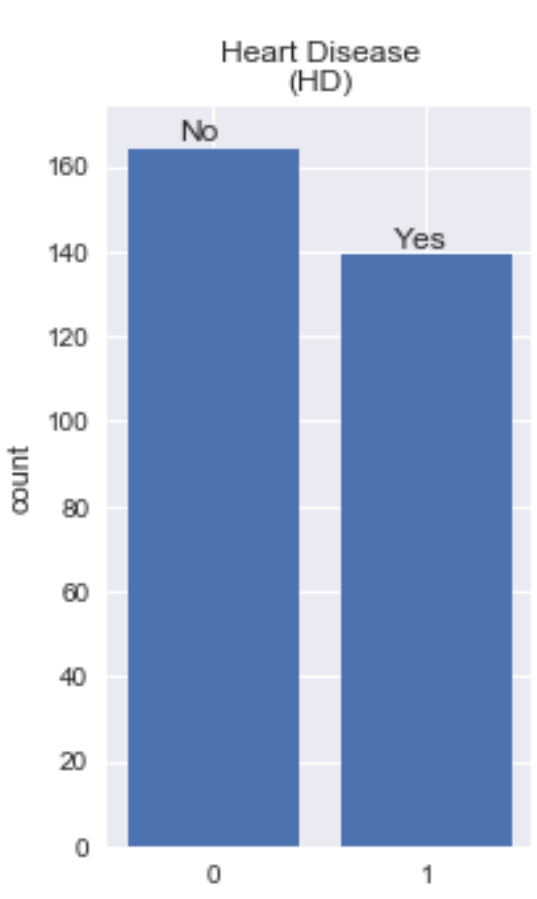


Data

Discrete Distributions







Classifiers

Naive Bayes

- Baseline Classifier
- Chose variables that are easy to understand and collect
- Model: $\hat{y} = \underset{y \in \{0,1\}}{\arg\min}\, P(Y = y) \prod_{i=1}^{n} P(X = x_i | Y = y)$
- X = AGE, GENDER, CP and TRESTBPS
- Y = HD (0 = No, 1 = Yes)

Classifiers

Logistic Regression

- Middle Classifier
- Incorporates medical tests
- Model: $P(Y = 1|X = x, \beta) = \frac{1}{1+e^{-(\beta^t x)}} + \lambda \frac{1}{2} \beta^t \beta$
- $X =$ CP, THAL, THALACH, EXANG and CHOL
- $Y =$ HD (0 = No, 1 = Yes)

Classifiers

Random Forest

- Most Complex Classifier
- Uses all of the fields in the data
- Model: Ensemble of 40 classification trees. Each tree grown with 2 features randomly selected for each split.
- X = AGE, GENDER, ..., CA, THAL
- Y = HD (0 = No, 1 = Yes)

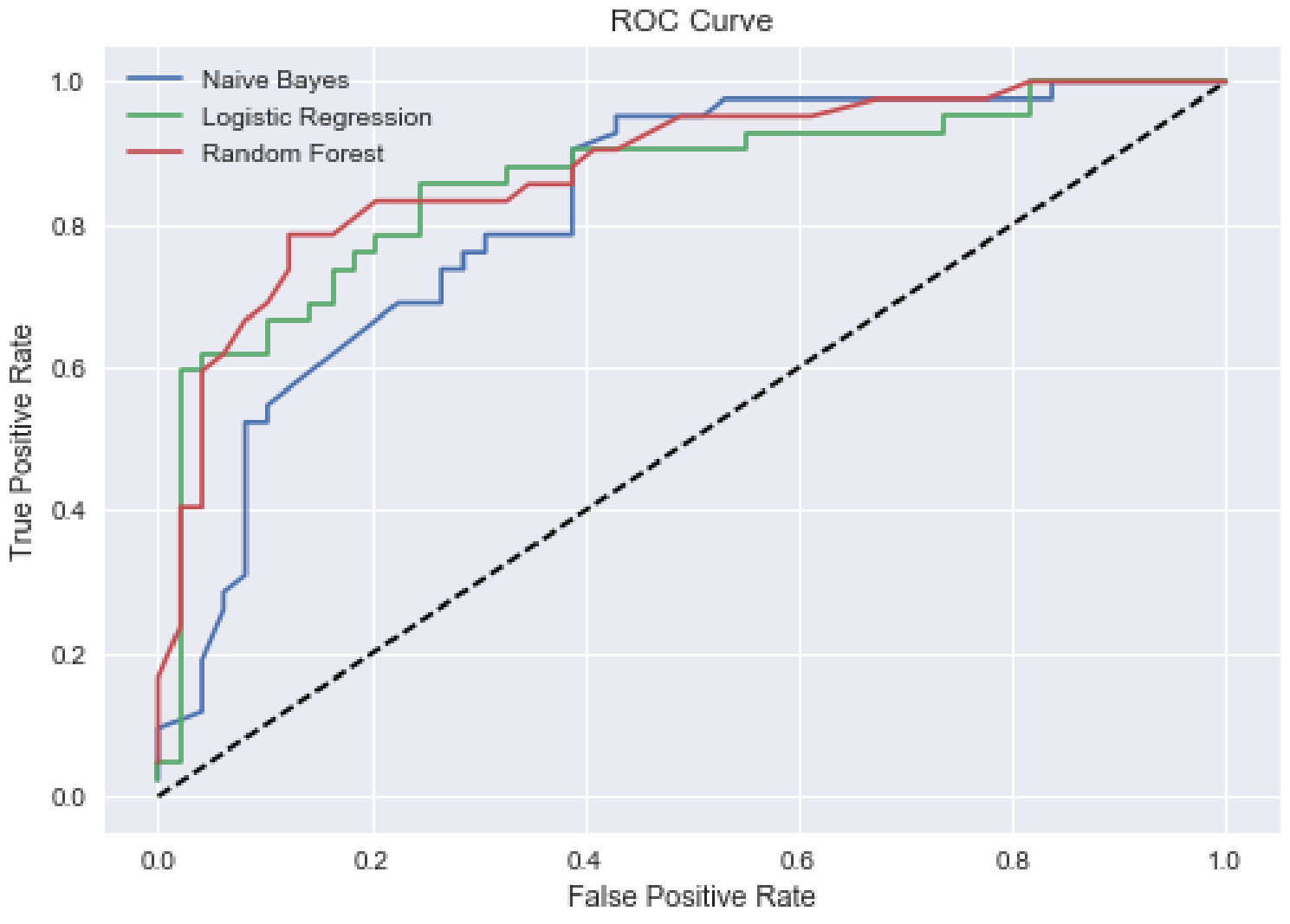
Results & Discussion

Evaluation Metrics - Testing Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC
Naive Bayes	0.736264	0.695652	0.761905	0.727273	0.821429
Logistic Regression	0.769231	0.733333	0.785714	0.758621	0.858601
Random Forest	0.835165	0.846154	0.785714	0.814815	0.878523

Results & Discussion

ROC Graph



Results & Discussion

Discussion

- Accuracy of the classifier increases as model complexity increases and number of features increases.
- Naive Bayes model = simple model but decent performance
- Random forest classifier = best classifier in terms of all of the evaluation metrics.
- Random forest has an accuracy of 83.52%, which is close to the best accuracy others have achieved using this dataset.
- Random forest classifier is relatively expensive. Cost to the patient of obtaining the information is $322.97. Naive Bayes is $4.00 and logistic regression is $199.47

Conclusion I

- Potential public policy implications.
  - Naive Bayes model shows AGE, GENDER, CP and TRESTBPS are good indicators of heart disease.
  - Simple enough model it can be explained to the general public.
  - Create chart of all possible combinations of AGE, GENDER, CP and TRESTBPT and associated probability of heart disease.
  - Educate people about these risk factors and treatments.

Conclusion II

- Shortcomings of Analysis
  - Only 303 observations in dataset. Compare with ~ 7 billion people and 600,000 deaths from heart disease per year.
  - Data taken from only one hospital.
  - Data from 1988. Many medical advances since then, e.g. genetic testing.