Capstone Project 1 Report
Heart Disease Predictor

Brian Lubeck

November 24, 2017

Introduction

Heart disease is the build up of plaque in a person's arteries and is the leading cause of death in the United States. More than 600,000 Americans die from it each year, with one out of four deaths each year being from heart disease. As the plaque builds up, the arteries narrow, reducing blood flow to the heart. A doctor can perform several tests to diagnose heart disease, including chest X-rays, coronary angiograms, electrocardiograms (ECG or EKG) and exercise stress tests. Fortunately, heart disease is treatable. Eating a healthy diet, exercising regularly, maintaining a healthy weight and taking medications are four ways to treat it and to reduce the risk of developing it in the first place. The goal of this analysis is to build a classifier that accurately diagnoses if a person has heart disease based on personal characteristics, symptoms and test results. The classifier will inform doctors which tests to perform on a patient and what personal information to collect. Doctors can then input the test results and personal information into the classifier to accurately determine the probability the person has heart disease.

Data and Initial Exploration

The Cleveland heart disease dataset from the University of California, Irvine's Machine Learning Data repository[1] will be used to build and test the classifier. The dataset consists of medical information collected from patients at the Cleveland Clinic Foundation. The dataset has 14 fields and 303 rows. Each row represents a different patient. The field of primary interest is *num*. It indicates the degree to which the diameter of a person's arteries has narrowed. It has five possible values 0 to 4 with 0 being less than 50% narrowing and $1, 2, 3$ and 4 being greater than 50% narrowing. As in previous studies, we take a value of 1 or greater to mean that the patient has heart disease. We create the field HD which equals 1 if $NUM \geq 1$ and 0 otherwise. We use the other 13 fields in the dataset to predict the value of HD.

The first steps in the analysis were to read in the data, assess its quality, and clean it as needed. The data was read into a pandas data frame object. Based on their descriptions, the fields were each given the numpy data type np.float_ or np.int_. Two of the fields CA and THAL were unable to be read in as integers because they had missing values, which had been marked as question marks in the data. Four records had missing values for CA, and two records had missing values for THAL. To handles the missing values, we simply replaced the question marks with -1.

Next, we looked at the distributions of the 14 data fields to assess if the data appeared reasonable and to detect any outliers. Table 1 displays quantiles for the continuous variables, and Figure 1 displays histograms for the discrete variables. The distributions for AGE and GENDER appeared reasonable. The minimum age in the dataset is 29 and the 10^{th} percentile is 42. People younger than 40 tend not to have heart disease, and more men than women are more affected by heart disease. There were not any

[1] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

outliers for AGE either. Assessing the distributions of the remaining variables required medical knowledge. For example, a cardiologist would know what the results of an electrocardiographic exam typically are for a person with heart disease. In general, having an understanding of the data is required to assess the reasonableness of the data at a high-level and low-level.

Table 1: 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

The next step is to develop an understanding of the relationships among the data fields. Since we are interested in predicting if a patient has heart disease, we look at the relationships between heart disease and the other data fields. Chest pain is often a sign of a heart attack and is usually the first concrete indication a person receives that something is wrong. Table 2 shows the relationship between heart disease (HD) and chest pain (CP). As can be seen from Table 2, there is a 21% chance a person has heart disease if chest pain is 1,2 or 3 but a 73% if chest is 4. Based on this, we should include chest pain in our prediction models. It also raises the question if it is worth the cost to educate people about chest pain symptoms and how to recognize the difference betweeen asymptomatic pain (4) and other types of pain (1,2 or 3).

Next, we look at how well a single medical test does at determining if a person has heart disease. Flouroscopy is medical imagining technique that uses X-rays to visualize parts of the body. One of its application is to visualize

Figure 1: Frequency Charts of Discrete Data Fields

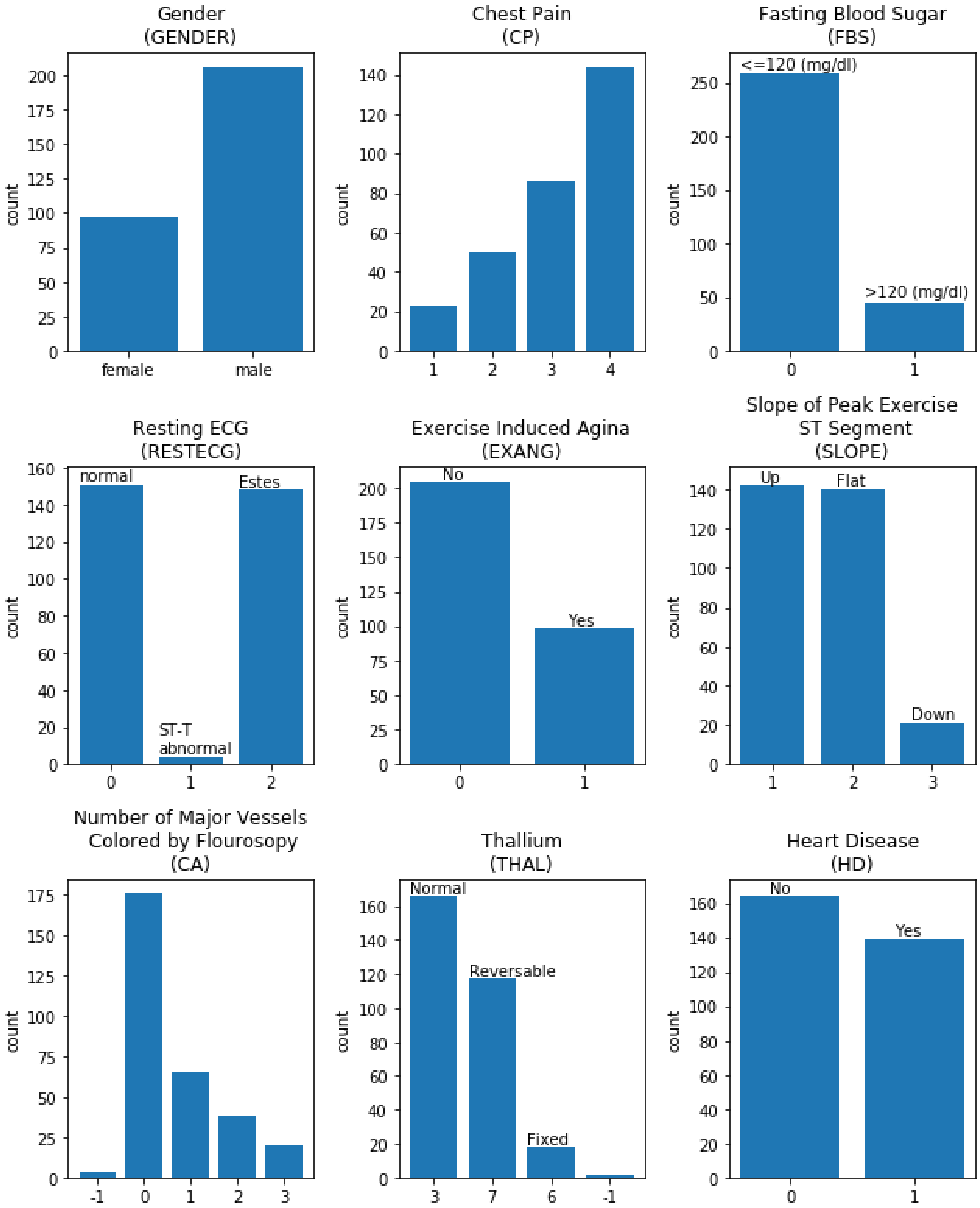
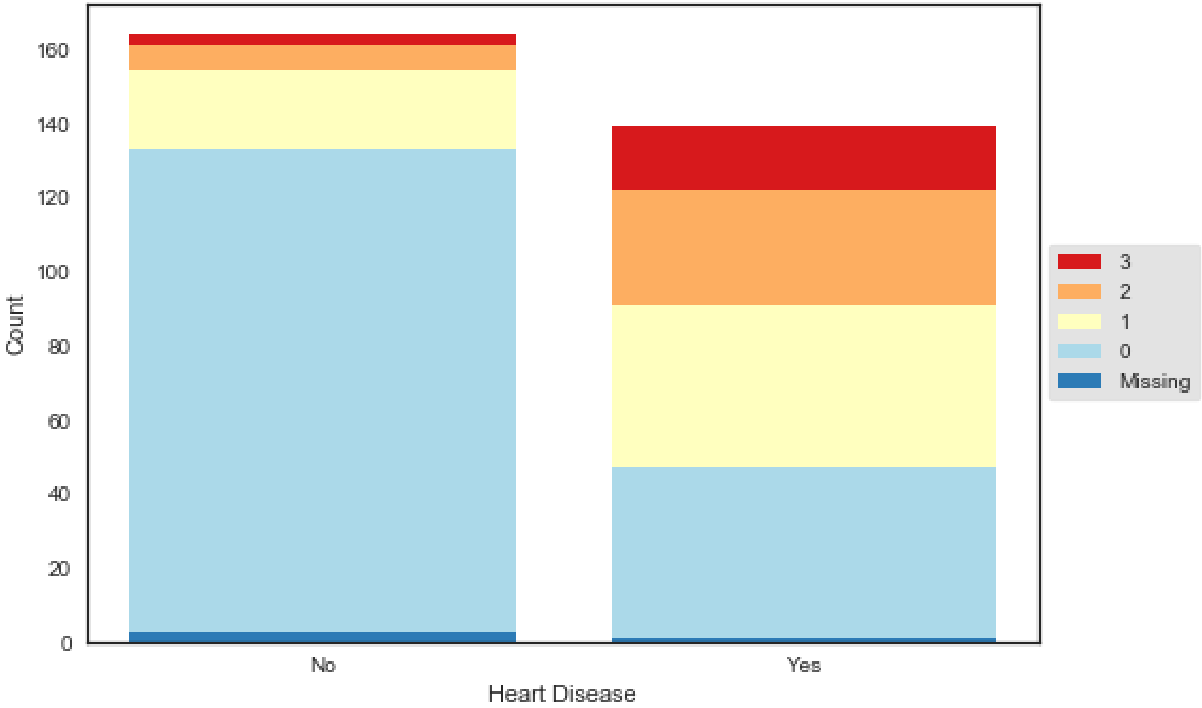


Table 2: Contingency Table of Heart Disease and Chest Pain

	Heart Disease (HD)		
Chest Pain (CP)	No	Yes	Total
1 (typical angina)	16	7	23
2 (atypical angina)	41	9	50
3 (non-anginal pain)	68	18	39
4 (asymptomatic)	39	105	144
Total	164	139	303

blood vessels and organs. Figure 2 shows the relationship between heart disease (HD) and the number of major vessels colored by flouroscopy (CA).

Figure 2: Flouroscopy versus Heart Disease



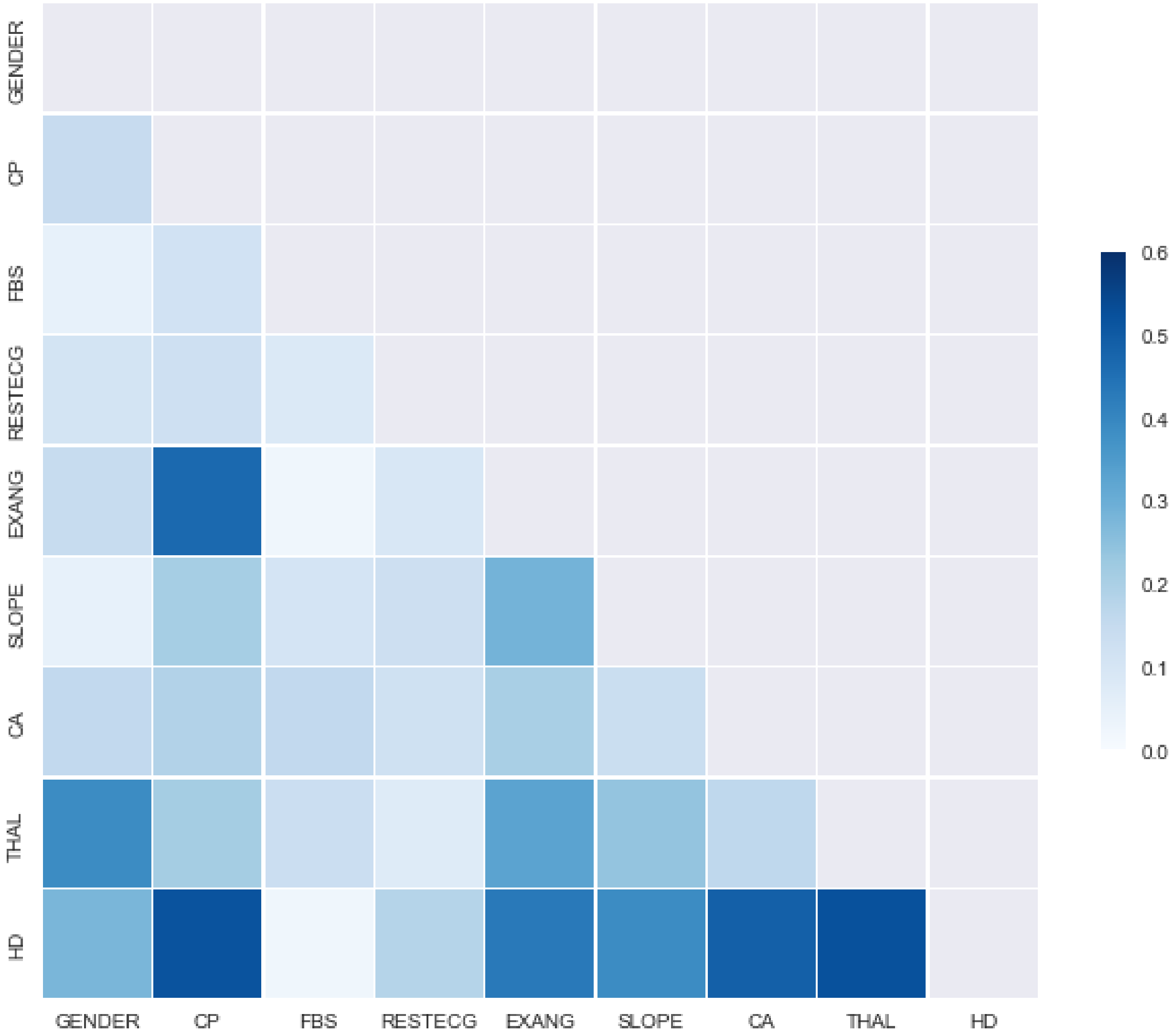
From Figure 2, we can see that having one or more major vessels colored by flouroscopy is indicative of heart disease. Thus flouroscopy is a useful test for a doctor to help diagnose heart disease.

Having explored the relationship between heart disease and two other data fields, we next explore how correlated all of the discrete data fields are with heart disease and each other. To answer this question, we calculate Cramer's V for each combination of discrete fields.[2] Figure 3 is a correlation matrix plot that shows the correlations compared to each other. As can be seen from Figure 3, chest pain (CP), thallium stress test (THAL) and flouroscopy (CA) have the strongest associations with heart disease among discrete fields. A thallium stress test is a nuclear imaging test that shows how well blood flows into the heart while exercising or resting. In addition, it appears that chest pain, thallium stress test and flouroscopy have a weakly moderate relationship. When building a model it is desirable to have predictor variables that are not strongly correlated with each other so that their individual effects can be more easily determined. On the other hand, a person's fasting blood sugar (FBS) has essentially no relationship with heart disease. Thus, this field can be safely excluded from our heart disease prediction model. Knowing how correlated each medical test is with heart disease can help doctors and patients determine if the additional information is worth the additional cost of performing the medical test.

Lastly, before we build any formal prediction models, we explore the association among the continuous variables. To do this, we calculate the sample Spearman rank correlation for each pair of continuous variables. Like, Pearson correlation, the Spearman correlation is a number between -1 and 1. A

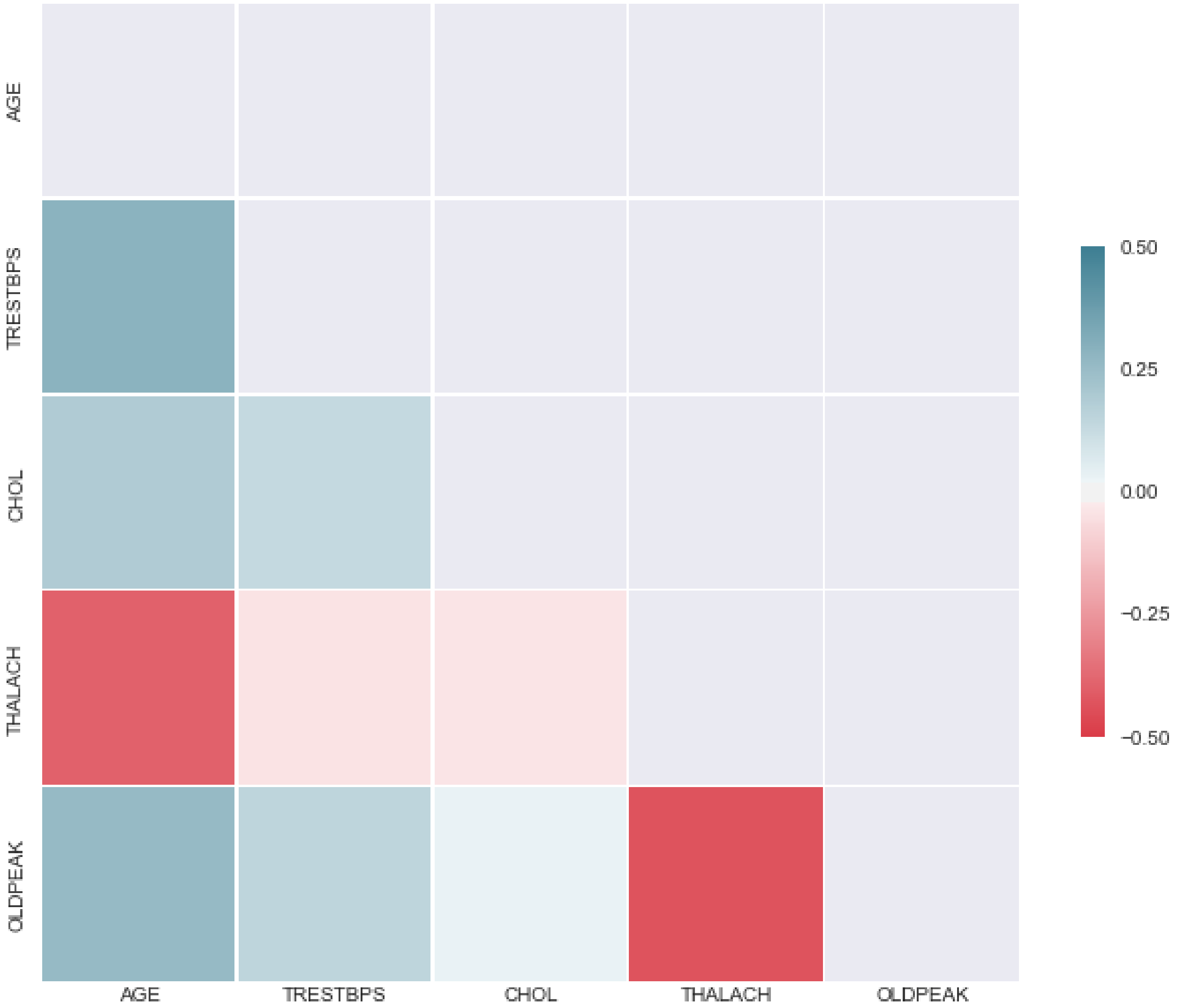
[2] Cramer's V is the χ^2 statistic normalized to lie between 0 and 1. A value of 0 means no relationship, while a value of 1 means a perfect relationship. One rule of thumb is that below 0.1 indicates a weak relationship, between 0.1 to 0.3 indicates a moderate relationship and above 0.3 indicates a strong relationship

Figure 3: Correlation Matrix - Discrete Variables



value of 1 indicates the two variables are an increasing function of each other, and a value of -1 indicates the two variables are a decreasing function of each other. As can be seen from Figure 4, there is a moderate decreasing relationship between age (AGE) and maximum heart rate achieved (THALACH), which is to be expected.

Figure 4: Correlation Matrix - Continuous Variables



Methodology

To build the classifier, we will use common machine learning binary classification models and evaluation metrics. The four models we will use are logistic regression, neural networks, support vector machines and random forests. We will evaluate the predictive accuracy of the models based on their F-scores and classification accuracy. The F-score is a commonly used evaluation metric for binary classification problems when the data contains very few zeros or ones.

A Description of Data Fields

N	Field Name	Description
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 on admission to the hospital
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
8.	THALACH	Maximum heart rate achieved
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	3 = normal; 6 = fixed defect; 7 = reversable defect
14.	NUM	Diagnosis of heart disease (angiographic disease status) 0 = less than 50% diameter narrowing 1,2,3,4 = greater than 50% diameter narrowing
15.	HD	1 if NUM $\geq$ 1, 0 otherwise