Capstone Project 1 Report
Heart Disease Predictor

Brian Lubeck

October 9, 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 his or her 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. 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 outliers for AGE either. Assessing the distributions of the remaining variables required medical knowledge. For example, a cardiologist

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

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.

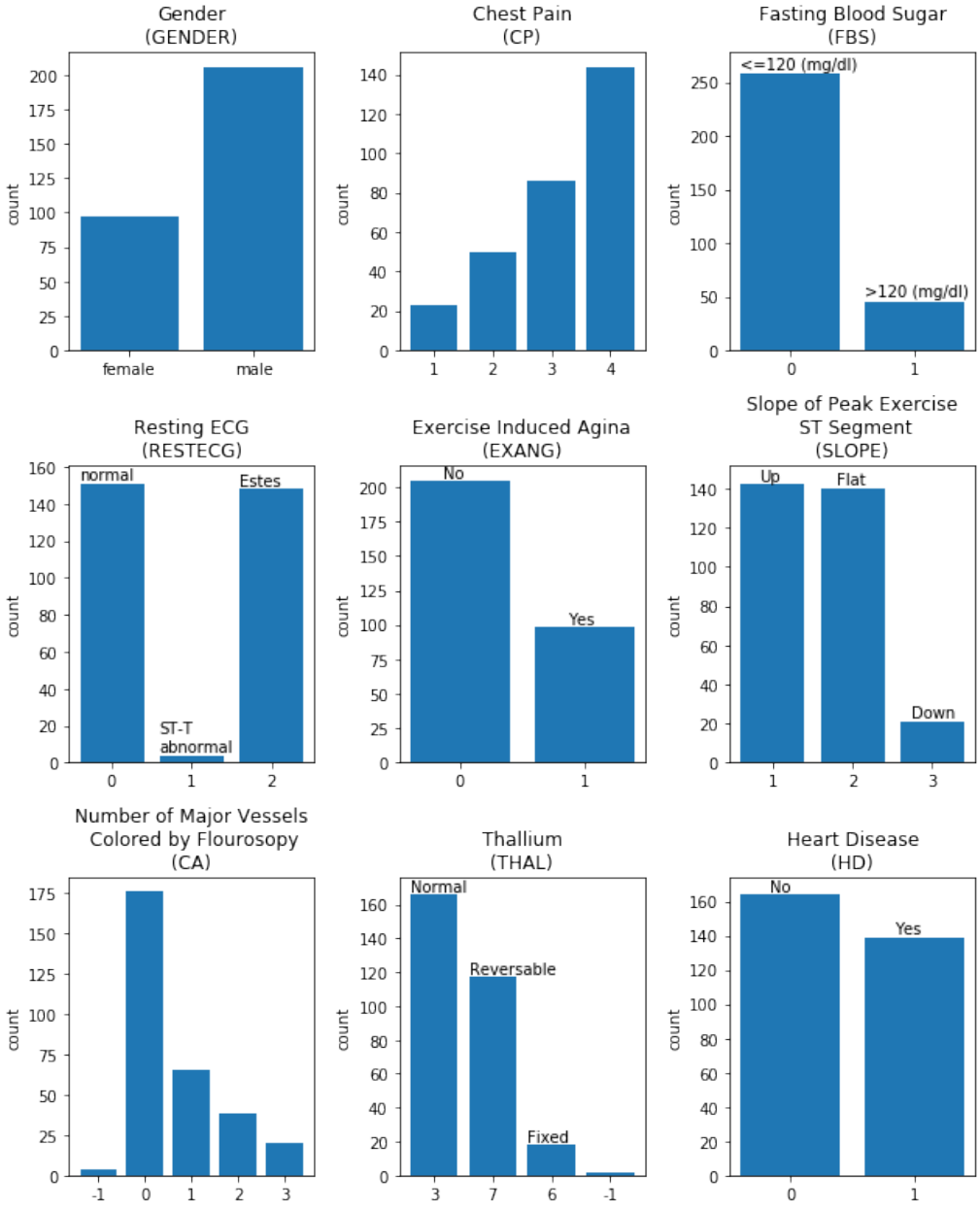
Table 1: Quantiles of Continuous Data Fields

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

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.

Figure 1: Frequency Charts of Discrete Data Fields



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