Sravani Korupolu

Winter 2020

Heart Disease Predictive Analysis

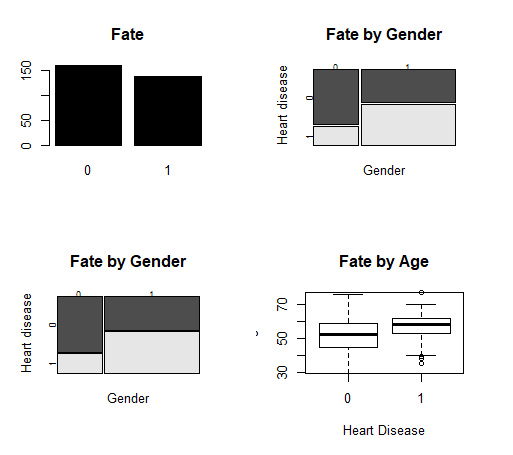
**Introduction**

Heart disease is one of the most prevalent health issues seen in America accounting for more than 1 in 4 deaths every year. Although it becomes increasingly more prevalent every year, the wide range of risk factors identified for it make it difficult to diagnose early on or even until it’s too late. Clinical data analysis provides the potential to identify which risk factors may be higher or what combinations of risk factors have an increased correlation for a heart disease diagnosis. This analysis in R will look utilize machine learning techniques to determine if a new patient should be given a diagnosis for heart disease. The data set used in this analysis for the Cleveland Heart Disease date set from the UCI repository.

The chosen data set has data from 303 patients who were tested for 14 different risk factors which were age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, max heart rate achieved, exercise enduced angina, ST depression induced by exercise, peak exercise ST segment, number of major blood vessels, thalassemia, and diagnosis of heart disease. For more information on how these categories were broken down, please visit the UCI website which details how they categorized this data into numbers for this data set.

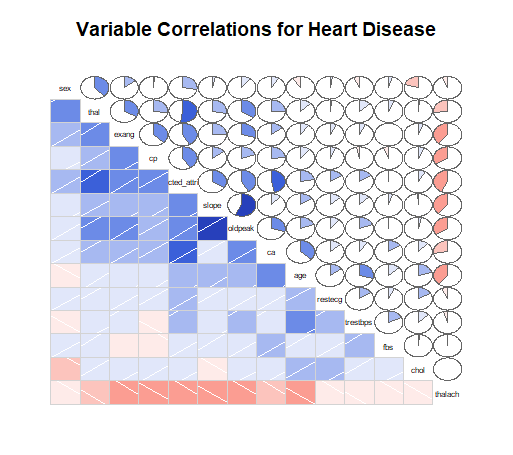
**Data Analysis**

First, we looked at how the most common variables had an affect on the likelihood of diagnosis. We first looked at how gender plays a factor:



From the first chart above, where 0 is representative of not having heart disease and 1 is representative of any kind of heart disease diagnosis. We are able to see that men are much more susceptible for heart disease than woman are. Next, we looked at how age can play a factor. In the box plot displayed above, we can see that the median age for developing heart disease is higher, close to 60%.

Next, we wanted to see how risk factors may be correlated not only with a heart disease diagnosis but other risk factors as well. Using the correlation tools in RStudio, the following graphic was obtained to show how they may be related:



The term labeled predicted attribute is representative of whether a person has heart disease or not while the rest of the terms are abbreviations for the 13 risk factors identified. We can see that some risk factors are more correlated with a diagnosis than others. For example, thalach (maximum heart rate reached) has fairly strong correlations with all of the risk factors and especially the heart disease diagnosis itself. In addition to thalach, cp (chest pain type), old peak (ST depression induced by exercise), and slope (slope of the peak exercise ST segment) were identified as having a higher correlation with a hear disease diagnosis. It is also interesting to see how these risk factors are correlated with each other and could help in assisting determing what other symptoms to test patients for an earlier diagnosis.

Having looked at some of the risk factors individually to gain a better understanding of our data, a linear regression was performed to see if using these risk factors and their correlations could help predict a diagnosis. The data set was separated into two categories with 70% of the data being used to train the algorithm and the remaining to test the algorithm against the known outcomes. The resulting accuracy using this data set and linear regression was determined to be 86.5% which sounds high, but in a clinical setting that’s missing 13.5% of your patients that should have been diagnosed or were wrongly diagnosed and given medication that could be potentially harmful. While linear regression seemed to have worked mathematically, for a clinical setting, there should be a much more accurate method of diagnosing patients with heart disease.