Kaggle Cardiac Data

Sam Loyd

Bellevue University

The goal of my analysis in this dataset was to determine if being overweight, as measured by BMI, tended to increase the likelihood of having a cardiovascular disease. I also wanted to see if gender had any practical significance on the hypothesis above. I created a calculated variable from weight and height using the standard formula for determining body mass index or BMI. I started with a distribution analysis of the data. Most variables were found to have a non-normal distribution. Examples of this were provided in my power point and more fully explored in the notebook attached to the project. Extreme cases were removed using domain knowledge.

The non-normal distribution across most variables, led to me choosing Spearman’s method for correlation analysis. I found that systolic blood pressure had the highest correlation to the binary variable for cardiac disease. BMI, while statistically significant and confirmed by a Chai-squared hypothesis test and others had a weak correlation. Two different models using multiple variables achieved a 72 percent accuracy at predicting cardiac disease. Gender did not indicate a correlation with cardiovascular disease. A Chai-squared test also failed to prove a correlation for gender and cardio.

Unfortunately, the data set provided by Kaggle, did not provide much in the way of describing that population. The data was collected in a medical setting where measurements were taken by the examiner in addition to recording survey answers provided by participants. Age ranges would lead me to believe that the population is not strongly representative of the general population. That information is more detailed in the notebook attached to the project and beyond the scope of my summary.

In performing this analysis, I confirmed what I learned from the program on removing bias from my work. I was often surprised that the data was telling me something that showed less practical significance than what I expected going into the process. I learned that there are many facets to the data. A low pseudo r-squared value led to a higher accuracy from a model than I would have expected. The categorical nature of the cardiovascular feature proved a bit more challenging. I had to rely on a few methods for visuals and testing not always provided for in the course. I would like to feel more comfortable spending less time verifying findings and looking up ranges. For example, I ran some testing that I knew would show strong correlation just to prove that I was performing tests correctly. I am new to this and that comfort should come with time.

In summary, my analysis proved a significant, but weak correlation for BMI correlating to heart disease. Two different regression models using systolic pressure alone in the form of simple regression had a 71 percent accuracy rating for predicting cardio disease with less risk of overfit. The educational value of providing this analysis was invaluable to me.