**Cardiovascular Disease**

**The Likely Causes of Cardiovascular Disease and How to Lower the Risk**

Mithil Patel

Bellevue University

DSC530-T302 Exploration and Analysis (2227 – 1)

Prof. Catie Williams

August 13, 2022

According to the World Health Organization (WHO), cardiovascular disease (CD) is the leading cause of death in the United States. I wanted to find a solution to this arduous problem using applications of data science. Due to the lack of progression in the cardiovascular research field, I wanted to produce a hypothesis using data science that could analyze certain health factors that affect a patient’s risk for heart disease. The answer I wanted to find is how we can take these certain risk factors and use a predictive model that would find a strong correlation between each variable. This will be beneficial as it will allow doctors and patients to regulate these risk factors promoting a healthier lifestyle.

Taking the results of the finds, I was surprised at certain aspects of which variables were strongly correlated and which others were not. Shockingly, age had a weak negative correlation with individuals with cardiovascular diseases, indicating an inverse relationship between age and heart disease. Additionally, the correlation coefficient between the variables hovered between -0.45 to 0.45. Using the logistic model, it suggested that the cp (chest pain) predictor had the greatest effect on heart disease. The logistic model is categorical which ends up raising or lowering the predicted probability to a discrete value, this in turn lowers the overall accuracy of the model. However, given additional time for research and in-depth analysis for the project would require alternative tools and strategies such as a machine learning model for better accuracy. In essence, I was unable to provide patients and doctors, with overwhelming statistical evidence, several lifestyle changes to potentially lower the risk of cardiovascular disease.

The results of the data analysis could have been made more accurate if I had not run into limitations with finding the right dataset. The most impactful and crucial tool I was finding during my analysis was finding a great data source. The additional dataset would have been a pivotal element in finding additional risk factors to improve the overall accuracy of the model. In addition to facing issues with datasets, there were some variables I felt could have also been significant in improving the accuracy of the model. Cholesterol, a key factor in providing strain to the heart, and diastolic and systolic which measure the blood pressure in the arteries when the heart rests and when the heart beats, respectively, are both important risk factors that would be correlated with heart disease.

Over the years, I have heard on numerous occasions that high cholesterol and blood pressure can increase the risk of heart disease. Therefore, I assumed that there will be a strong correlation between them; however, my analysis suggests contrary results. Perhaps it is indeed true that there is no correlation, or the dataset is not great. As mentioned previously, some of the challenges I faced were finding good datasets along with precisely associated variables that would be the best predictors of heart disease with a higher correlation coefficient. I also felt a challenge in trying to answer the hypothesis test. While I prove with the model and the analysis that some of the variables were strongly correlated, it did not amount to being fully conclusive since I was unable to fully comprehend the implications of logistic regression coefficients or what statistically significant implies.