

Heart Disease Prediction

By Jacob Hurley and Sofia Gray

Introduction:

According to the CDC, heart disease is the leading cause of death for people of most racial and ethnic groups in the United States. About 655,000 Americans die from heart disease every year and between the years of 2014 and 2015, it cost the United States about \$219 billion each year. Heart attacks can happen with no warning and also be silent (meaning that the person is not aware of it happening). Additionally, about 805,000 Americans every year have heart attacks and about 605,000 of these are their first heart attack. Some known factors that can put people at risk for heart disease are high blood pressure, high blood cholesterol, overweight and obesity, an unhealthy diet, and physical inactivity. Being able to predict if you have developed heart disease, or are going to develop heart disease is important in stopping it while you can. Since things like blood pressure, cholesterol, weight, and more are affected by daily habits such as if you workout (including cardio), eat healthy foods consistently, and stress, it is important to know which factors may lead to heart disease, you may change certain habits if you are able to predict if they will lead to heart disease.

Methods:

The data we will be using for this project was collected from already existing individual datasets that were compiled into one single dataset. In this singular compiled dataset, we observe five different heart datasets from Cleveland, Hungarian, Switzerland, Long Beach, and Stalog, which include 303, 294, 123, 200, and 270 observations respectively. These individual datasets were created at medical institutions by Andras Janosi, M.D. at Hungarian Institute of Cardiology in Budapest, William Steinbrunn, M.D. at University Hospital, Zurich, Switzerland, Matthias Pfisterer, M.D. at University Hospital, Basel, Switzerland, and Robert Detrano, M.D., Ph.D. at

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. We can assume that the data collected at each institution for its respective dataset was derived from patients at those institutions but it is unclear whether or not this is true. There should be no sources of bias except from the measurement instruments which would be negligible. This is because all of the data is objective statistics about the individuals.