Determining Extreme, High, Medium, and Low Risk Patients for Heart Failure

Cardiovascular disease is the leading cause of death in the United States. It has been proved through research that every 37 seconds a person dies from the deadly disease, that is about 674,000 Americans in one year, and 17 million people globally every year. Heart failure occurs when the heart cannot pump sufficient blood for the survival of the body. In order to detect the symptoms and risk factors, doctors have electronic medical records of patients with symptoms, body features, and results of clinical laboratory tests. These can be used to perform biostatistics analysis and correlations for detection of a potential cardiovascular disease. In this case study, we will be using the dataset from Davide Chicco and Giuseppe Jurman from the Krembil Research Institute. This dataset contains 13 different attributes from age, anemia, all the way to the death event.

The goal of this analysis is to detect high risk, medium risk, and low risk patients of heart failure by using unsupervised learning. We will achieve this by first determining the death events of patients within k means of 2 attributes. We will perform a correlation to make sure we are including graphs with the best high correlating attributes. Then within each graph, we will construct clustering of 3-4 clusters with high, medium, low, and extreme risk that does not include labeling of deaths. Finally, to double check that we selected the correct high-risk attributes, we can compare the high risks to the death event attribute. Using this information, we will be able to determine patients with certain attribute or attributes that are at extreme risk. This is important for the medical industry as it optimizes the number and types of patients in need of urgent treatment. In addition, will be able to clarify how these attributes influence heart failure and back up our results with current medical research.