Predicting the Survival Chances of Patients with Heart Disease

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 predict the most influential factors in a patient’s death event. Based on these factors we can predict the patient’s survival chances. The data will first be pre-prepossessed and go through a pre-exploratory process. Furthermore, we will use different techniques such as Nonlinear SVM, Random Forest and Neural Networks to predict the survival chances of the patient. Since the data is imbalanced with 33% patients death and 67% survival from heart failure, we will try different techniques to to improve accuracy of prediction of the minority class which is patients that didn’t survive but were predicted as patients that survived and vice versa. Moreover, we will use feature ranking to identify which factors are the most and least influential for survival. 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.