1. Project Description:

Cardiovascular disease (CVDs) is the number one cause of death globally, claiming an estimated 17.9 million lives each year, accounting for 31% of all deaths worldwide. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet, obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with or at risk for cardiovascular disease need early detection and management where a model can be Machine learning is very useful.

1. Dataset Brief Description:

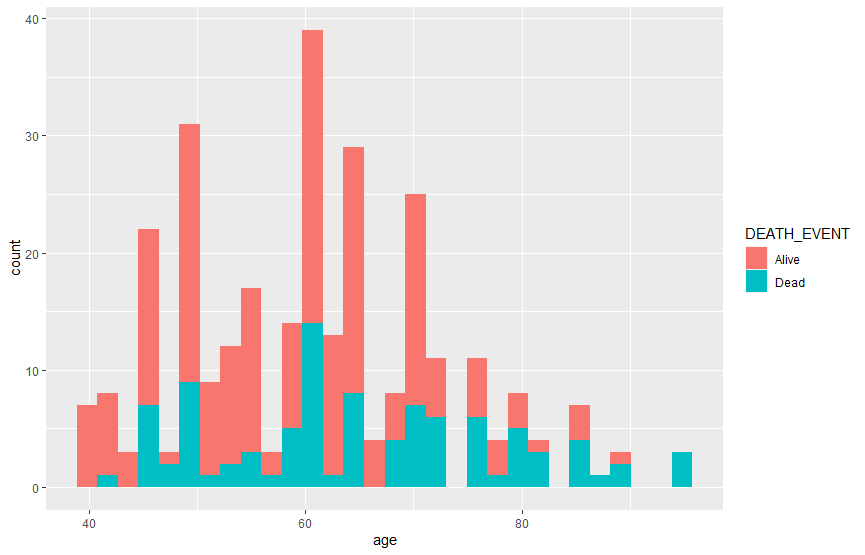
Heart failure is a common CVD event and this dataset contains 12 features (such as age, anaemia, creatinine\_phosphokinase, diabetes, ejection\_fraction, high\_blood\_pressure, platelets, serum\_creatinine, serum\_sodium, sex, smoking and time) that can be used to predict mortality from heart failure.

1. The Problem Definition and Objectives:

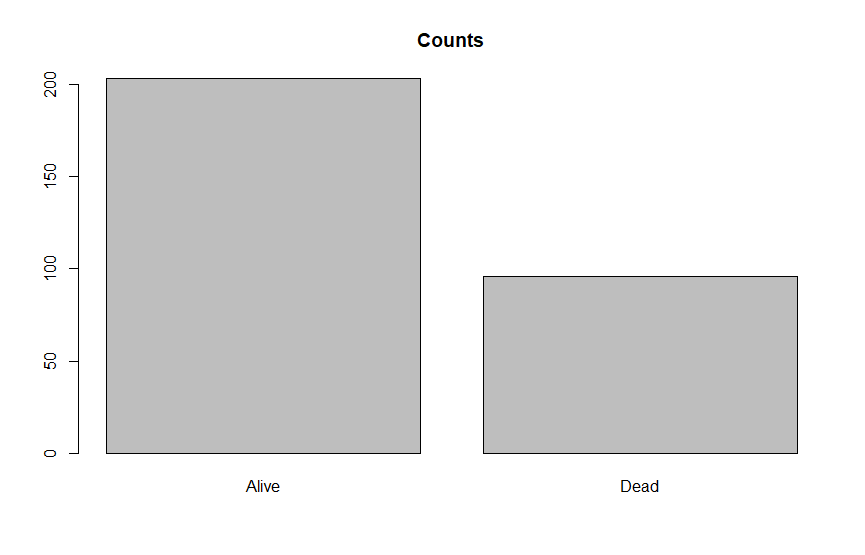
The Problem Definition: 12 features that contributes to the number of deaths due to heart failure.

Objective: Applying the Big Data Analytic project cycle on dataset. And using machine learning techniques such as Decision Tree on dataset model to predict probability of heart failure using the 12 features. And visualizing the data to find the relationships between them using data visualization charts.

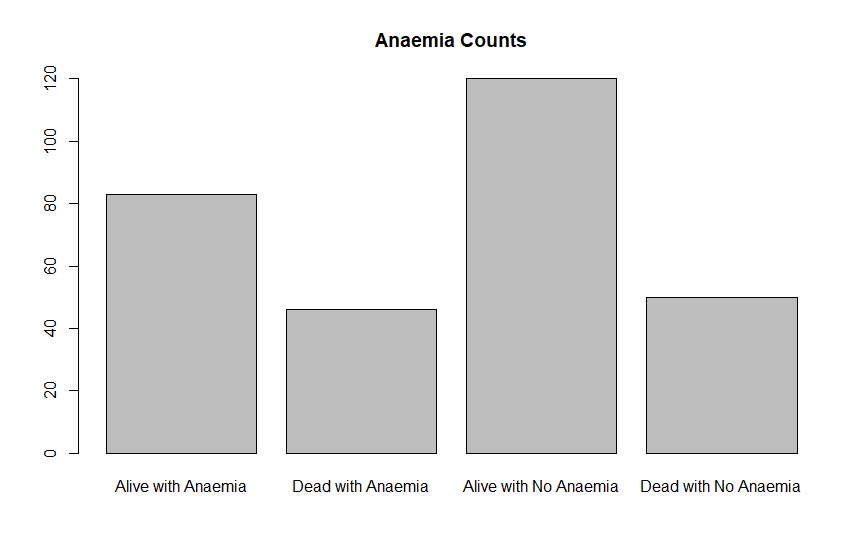
1. Screenshot:



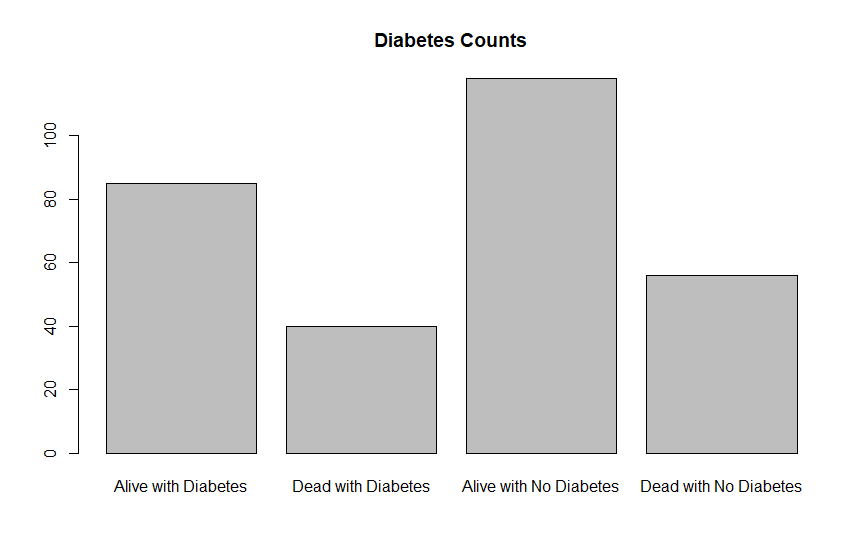
This chart shows that as people age increase, their chance of dying due to heart failure increases. It also shows that most people who are very old (about 75 years old and above) are dead due to heart failure since more that three-quarter of people who are about 75 years old and above are dead due to heart failure. It also shows that people who are about 40 years old have the lowest chance of dying due to heart failure since there is almost no death due to heart failure at the age of around 40. It also shows that people who are about 60 years old have the highest number of dying due to heart failures since the highest number shown in the chart of death due to heart failure is at the age of around 60.



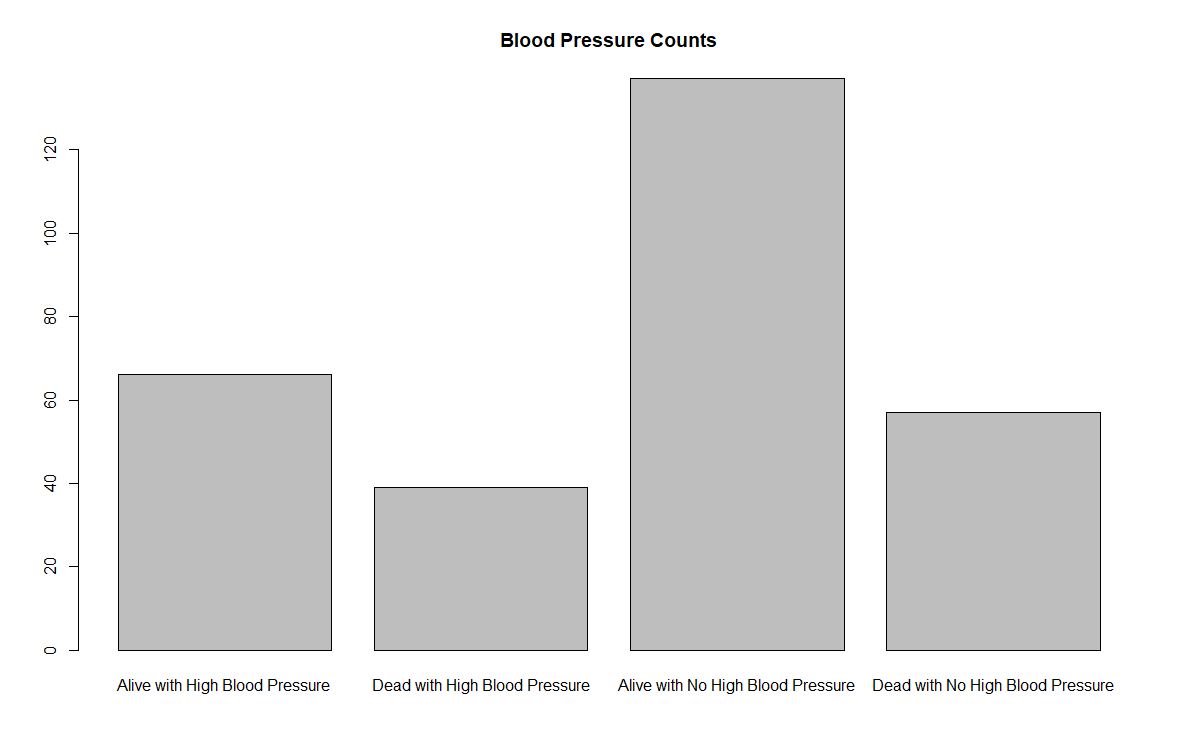
This chart shows that the number of people who are dead because of heart failure are half the number of people alive.



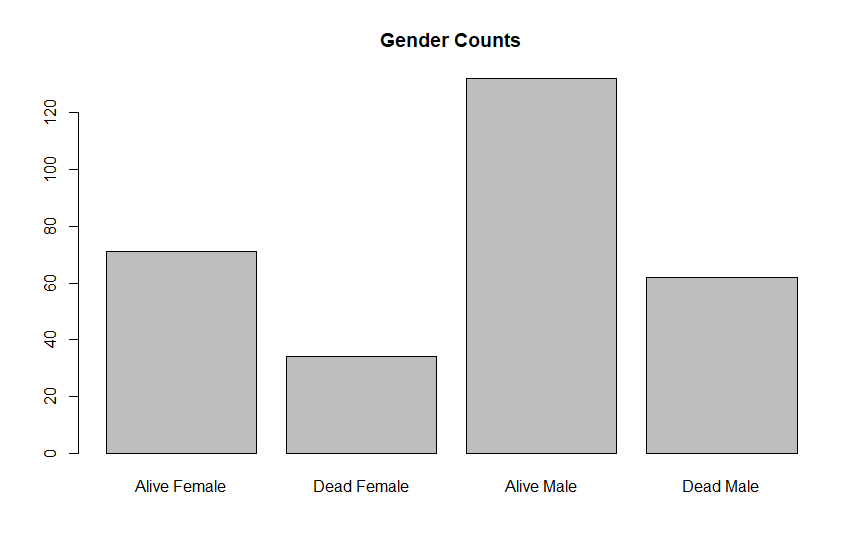
This chart shows that having anaemia doesn’t increase chances of heart failure since the dead with anaemia and the dead with no anaemia are almost equal on the chart. It also shows that the number of people who are alive but have anaemia are about two-third the number of people who are alive but don’t have anaemia.



This chart shows that having diabetes doesn’t increase chances of heart failure since the dead with diabetes are lower that the dead with no diabetes. It also shows that the number of people who are alive but have diabetes are lower that the number of people who are alive but don’t have diabetes.



This chart shows that having high blood pressure doesn’t increase chances of heart failure since the dead with high blood pressure are lower that the dead with no high blood pressure. It also shows that the number of people who are alive but have high blood pressure are about half the number of people who are alive but don’t have high blood pressure.



This chart shows that the number of males who are dead due to heart failure are almost double the number of females who are dead due to heart failure. It also shows that in both gender the number of death due to heart failure are half the number of people alive.

1. Applied Data Cleaning Methods:
2. Incorrect Values
3. Change Data Types
4. Dataset Preparation:

Shuffle Dataset.

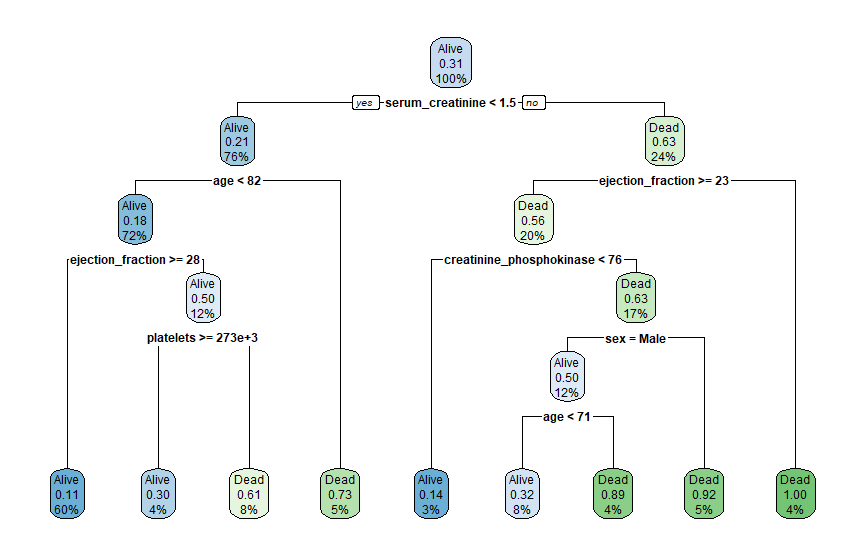
Create train/test set: - Split data 80/20, 80 percent of the data serves to train the model, and 20 percent to make predictions.

Build the model using training dataset.

Make a prediction using test dataset

1. Used Data Analytics Techniques:

Decision Tree. We used decision tree because output ( DEATH\_EVENT ) is discrete.



1. Performance Measure:

Measure performance by computing an accuracy measure for classification task with the confusion matrix:

The confusion matrix is a better choice to evaluate the classification performance.

Accuracy = 0.8 = 80%