Capstone Project – The Machine Learning Model of Heart Failure Prediction

Alec Xu
29 Aug 2020

1. Introduction of Background

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and it means the heart is unable to pump sufficiently to maintain blood flow to meet the body's needs.

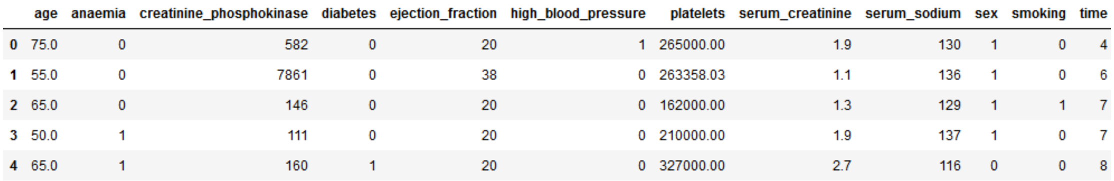
Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help. That is the starting point that I will use different modelling methods to estimate the risk that a person may get heart failure as an effective reference to support the media examination of hospital.

2. Data Cleaning and Splitting

This dataset contains 12 features that can be used to predict mortality by heart failure is obtained from Kaggle with the link below.
https://www.kaggle.com/andrewmvd/heart-failure-clinical-data?select=heart_failure_clinical_records_dataset.csv

The 12 attributes are age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, death event. And the total sample scope is 299.



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8

After loading data, firstly I will check the data types of each column and also whether there are null values in the dataset to ensure there is no blank entry.

Then I will drop the death event column, which means the person had heart failure (1) or not (0), from the dataset and treat it as y, the dependent variable, while the 12 factors will be independent variables X. And all the 299 data will be split referring to the ratio of 7:3 as training data and test data to be prepared for the modelling in the next stage.