# Predicting death events from heart failures

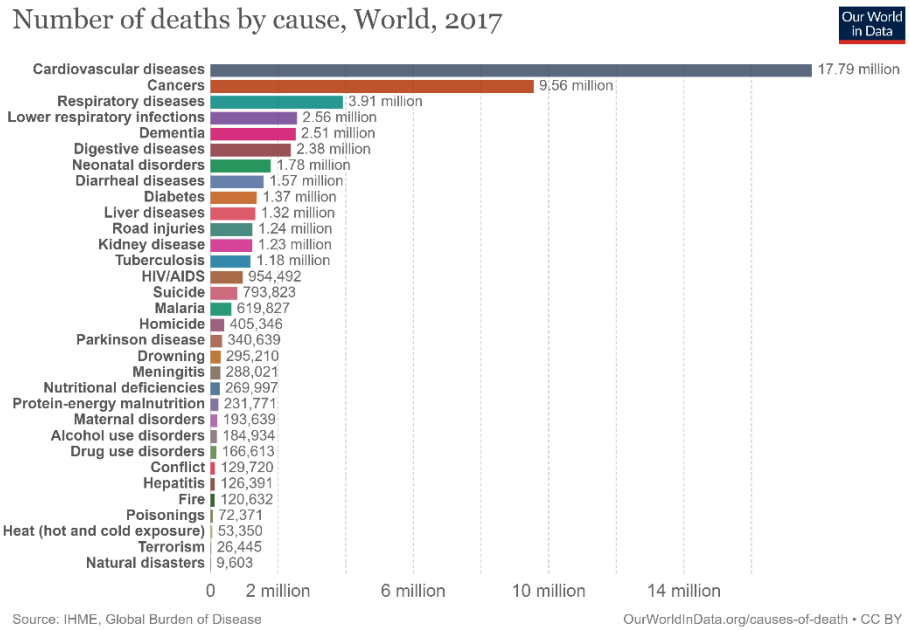
Brandon Yilmaz Castillo

27/09/2020

## 1.Introduction

### 1.1Background

Heart failures or heart diseases are the major cause of death worldwide. In order to handle this problem, healthcare systems are providing progressively more solutions based on technological improvements. One of many of these improvements is the application of Machine Learning modelling. Heart disease (or cardiovascular disease) is being the top disease worldwide, as it's shown in the table below from OurWorldInData database. Understanding properly and in-time what can cause this phenomena is of a vital importance to tackle it and therefore minimize its effects among the society.




### 1.2Problem

In order to increase efficiency when healthcare resources are in risk or are scarce, stakeholders of healthcare enterprises would choose better systems of diagnosis and prevention. Machine Learning models are suitable for this purpose and have demonstrated several positive results. In times of COVID-19 pandemic different healthcare systems have systematically failed, sometimes showing a lack of coordination or bad Database Management Systems (DBMS).

## 2.Data acquisition and cleansing

### 2.1Data source

The data was gotten from Kaggle.com, but originally the data was provided by UCI (University of California, Irvine) from the paper *"Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone"* (Chicco & Jurman, 2020).

### 2.2Data cleansing

The data had originally 13 features, with no missing data points. These features are: age of the patient, anaemia (dummy variable), high blood pressure (dummy variable), creatinine phosphokinase (CPK) : level of the CPK enzyme in the blood (mcg/L), diabetes: if the patient has diabetes (dummy variable), ejection fraction: percentage of blood leaving the heart at each contraction (percentage), platelets: platelets in the blood (kiloplatelets/mL), sex: woman or man (dummy variable), serum creatinine: level of serum creatinine in the blood (mg/dL), serum sodium: level of serum sodium in the blood (mEq/L), smoking: if the patient smokes or not (dummy variable), time: follow-up period (days), [target] death event: if the patient deceased during the follow-up period (dummy variable). It has to be said that this is a small dataset with just 300 observations, nevertheless, the features are relevant and with no missing values as pointed out earlier.

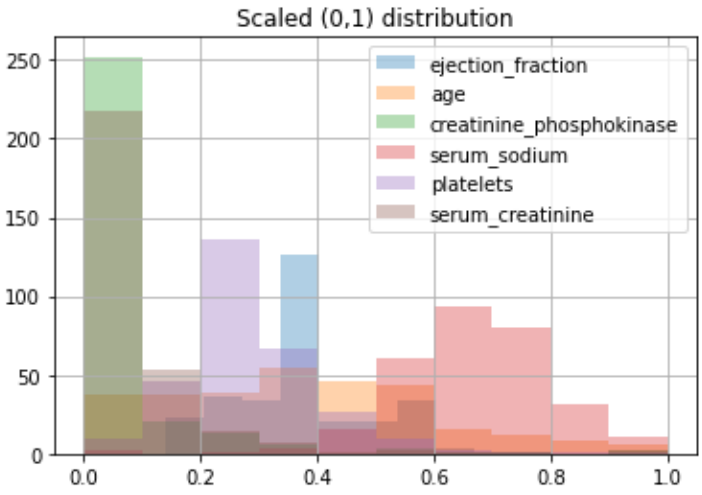
The main problems when data cleansing techniques are applied is the different measurement of the features. This can be a problem when Classification Algorithms are applied. Dummy variables are already 'scaled' in a sense, but continuous variables as platelets in the blood or serum creatinine are not. Therefore, Scaling technique was applied to the following variables:

- creatinine phosphokinase (CPK).
- serum sodium
- platelets: platelets in the blood (kiloplatelets/mL)
- serum creatinine
- age

The Scaling formula applied was:

$$X' = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}}$$

Ejection fraction was in percentage, so it has been divided by 100 in order to get the correct scale (0,1). The plot below shows the correct scaling applied to all continuous variables:




No Pearson correlations higher than 0.9 were found, so there is no Multicollinearity between features.