**Introduction:**

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 worlwide. Heart failure is a common event caused by CVDs, which could be derivative of 12 different features that may be used to predict mortality by heart failure.

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

In this context, electronic health records (EHRs, also called medical records) can be considered a useful resource of information to unveil hidden and non-obvious correlations and relationships between patients’ data, not only for research but also for clinical practice and for debunking traditional myths on risk factors. To this aim, several screening studies have been conducted in the last years, covering different conditions and demographics and with different data sources, to deepen the knowledge on the risk factors. 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, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Machine learning applied to medical records, in particular, can be an effective tool both to predict the survival of each patient having heart failure symptoms and to detect the most important clinical features (or risk factors) that may lead to heart failure.

**DataSet:**

We analyzed a dataset containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All 299 patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure.

The dataset contains 13 features, which report clinical, body, and lifestyle, that we briefly describe here. Some features are binary: anaemia, high blood pressure, diabetes, sex, and smoking. The hospital physician considered a patient having anaemia if haematocrit levels were lower than 36%. Unfortunately, the original dataset manuscript provides no definition of high blood pressure.

Regarding the features, the creatinine phosphokinase (CPK) states the level of the CPK enzyme in blood. When a muscle tissue gets damaged, CPK flows into the blood. Therefore, high levels of CPK in the blood of a patient might indicate a heart failure or injury.

The ejection fraction states the percentage of how much blood the left ventricle pumps out with each contraction.

The serum creatinine is a waste product generated by creatine, when a muscle breaks down. Especially, doctors focus on serum creatinine in blood to check kidney function. If a patient has high levels of serum creatinine, it may indicate renal dysfunction.

Sodium is a mineral that serves for the correct functioning of muscles and nerves. The serum sodium test is a routine blood exam that indicates if a patient has normal levels of sodium in the blood. An abnormally low level of sodium in the blood might be caused by heart failure.

The death event feature, that we use as the target in our binary classification study, states if the patient died or survived before the end of the follow-up period, that was 130 days on average. Regarding the dataset imbalance, the survived patients (death event = 0) are 203, while the dead patients (death event = 1) are 96. In statistical terms, there are 32.11% positives and 67.89% negatives.