**GROUP** Pietro Ferrazzi, Margarita Hernández

**TITLE** Analysis and survival prediction of heart attacks

**DATASET DESCRIPTION**

The data set has been downloaded at <https://archive.ics.uci.edu/ml/index.php>. It contains **299 observations** of **13 variables**. Each statistic unit is one patient. What follows is a brief description of the variables:

* TIME: time under observation for the patients, i.e., follow-up time during which the patient was constantly monitored. If the event occurred after that time, there is not track of that in the data;
* Event: it is set to 1 is the heart attack occurred, 0 otherwise;
* Gender, Smoking, Diabetes, BP and Anaemia are **dichotomous variables**;
* Age, Ejection.Fraction, Sodium, Creatinine, Pletelets and CPK are **continuous variables**.

**WHAT TO BE DONE**

* Classic explorative analysis to identify important variables and possible correlations in order to estimate the survival ratio.
* Explorative analysis to identify which variables can be useful to provide forecasts about the variable *Event* treated as a *class.*
* Modelling throw ML techniques to perform classification of the data.
* Comparing the results with previously performed analyses

**WHY**

Integrate classical statistical analysis with ML techniques to see if they can perform well in an *adverse* scenario: few variables, few observations, and the fact that the data are of a survival type and then should be treated following strict procedures. In addition, we are really interested in data coming from the sanitary field.

**PREVIOUS AVAILABLE WORK**

There are previous studies conducted with this dataset, two research papers with different approaches, as well as different results. Both tried to find significant predictors, one implementing Cox regression for survival analysis, the other using rather modern techniques, biostatistics and machine learning. The first study [1] determined that age, serum creatinine, blood pressure, anaemia and ejection fraction were contributors to the risk of mortality. The second study [2] concluded that only serum creatinine and ejection fraction were relevant features, and that those two alone led to more accurate predictions. Furthermore, the more recent study, showed that random forest was the top performing classifier among 10 different prediction models.

After conducting our own analysis, we would like to see which of these studies have results most similar to ours and analyse why this might be the case.

[1] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA (2017) *Survival analysis of heart failure patients: A case study*. PLoS ONE 12(7): e0181001. <https://doi.org/10.1371/journal.pone.0181001>

[2] Chicco and Jurman BMC Medical Informatics and Decision Making (2020), *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*, 20:16 <https://doi.org/10.1186/s12911-020-1023-5>