Heart-Failure

Paul Gis

12/30/2020

1. Introduction

1.1 Preface

As a physician and epidemiologist, I wanted to have knowledge in the data science area in order to improve my profile and to take the additional step towards the use of the data as nowadays, these skills are crucial for integrating and analysing the great deal of information gathered from different sources. In this sense, and for completing the "choose your own" submission in the capstone course, I looked for a project related with health.

1.2 Project Overview

The project is based on a dataset published in kaggle (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data) with information collected in 2015 of 299 patients with heart failure (HF), a cardiovascular disease (CVD) where the heart progressively is not capable of pumping enough blood to meet the body needs and could be life-threatening if not treatetd properly. The dataset contains 13 features, 12 of them can be used to predict the 13th one: death event; so in general, we are in front of a surpervised machine learning project to predict mortality caused by heart failure.

1.3 Executive Summary

1.3.1 Workflow

The present project applys several machine learning classifiers to both predict the patients survival, and rank the features corresponding to the most important risk factors. Despite I am following the work made by Davide Chicco and Giuseppe Jurman (please find bellow the credicts to their work in the Acknowledgements section), they only described the processes and analysis they made but they did not share any of their algorithms, so the workflow I present here represents my own coding skills and the results may vary from theirs.

After reviewing the structure of the dataset, and arranging the cathegories of each feature, I proceeded to make a general review to look for NAs or any other empty slots

Moving on, a more in depth analysis was done to see if the numerical features had normal distributions, and after adjusting them by taking out the extreme values, a t-test was preformed to compare the means of ecah feature with the death event. For cathegorical features, a chi squared test was performed to see whether they were independent or not of the death event.

finally, a desition tree and random forest models were trained for prediction and find out the most relevant features in the dataset.

1.3.2 Results

Based in the proposed model, we found out that in the dataset 5 of 7 numerical features did not have normal distributions, which led to find several outlayer values. Not to mention the cathegorical values, where was demonstrated that all of them were independent of the death event; taking out the extreme values in the numerical features resulted in the similarity of the means between each feature and the death event, which supported the non statistical differences. Following with the models, the desition tree and the random forest, these considerations were not taken into account and the algorithms used the whole dataset including the outlayer values, resulting in similar accuracy despite different features took relevance in each model.

1.3.3 Conclusions

- There is no information about the previous medical history of the patients and for how much time they had the diagnosis of the HF before entering the registry, so it is difficult to find correlation between variables or stablish any clinical significance.
- There is no information about when the tests or mesurements in the dataset were taken once the patients entered the registry, which makes it impossible to know if they were taken at the begining or any other moment during the patinet´s follow up.
- There is no information about any treatments taken before or during the follow up period, so it is no clear the clinical baseline status of the patients included in the registry and whether the HF was controlled or not, or if the death event was attibutable to the HF or any other cause.
- The results of the models presented in this project are only for academic purposes and cannot be considered for clinical desitions, as more detailed information should be included for deeper analysis.

1.4 Acknowledgements

Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5. (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5)

2. Methods and Analysis

As mentioned before, heart failure (HF) is a cardiovascular disease (CVD) where the heart is unable of pumping the blood to meet the organs and tisues needs, therefore, the patients need an appropriate and early treatment for controling the progression of the disease and avoiding further complications.

The medical literature has already described several risks factors for CVD, such as hypertension, diabetes or hyperlipidemia among others; but there are also some behavioural risks factors that contribute to the disease and are completly preventable, such as such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol. The dataset used for this project, combines a mix of both clinical and behavioural risks.

2.1 Comprehensive Analysis of the Dataset

The dataset presented for this project contains the medical records of 299 HF patients (rows) 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. It contains 13 columns (features): Age, Anaemia, High blood preasure, Creatine phosphokinase

(CPK), Diabetes, Ejection fraction, Sex, Platelets, Serum creatinine, Serum sodium, Smocking, Time and death event, being this last one the target feature for the prediction model. A short description for each variable can be found in Table 1.

Table 1 Meanings, measurement units, and intervals of each feature of the dataset

(https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5/tables/1)

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
(CPK)			
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving	Percentage	[14,..., 80]
	the heart at each contraction		
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,..., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,...,285]
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre

For a better understanding of the features, we can found a quick description of them in the paper written by Davide Chicco and Giuseppe Jurman: *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.*

As is shown in table 1, some features are numeric: years, Creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium and days; an others are boolean (binary), so they ranges are 0 or 1: anaemia, high blood pressure, diabetes, sex, smoking and death event; however, taking a closer look to the dataset in R shows all features as numeric:

```
## tibble [299 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age                     : num [1:299] 75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : num [1:299] 0 0 0 1 1 1 1 1 0 1 ...
##  $ creatinine_phosphokinase: num [1:299] 582 7861 146 111 160 ...
##  $ diabetes                : num [1:299] 0 0 0 0 1 0 0 1 0 0 ...
##  $ ejection_fraction       : num [1:299] 20 38 20 20 20 40 15 60 65 35 ...
##  $ high_blood_pressure     : num [1:299] 1 0 0 0 0 1 0 0 0 1 ...
##  $ platelets               : num [1:299] 265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num [1:299] 1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : num [1:299] 130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : num [1:299] 1 1 1 1 0 1 1 1 0 1 ...
##  $ smoking                 : num [1:299] 0 0 1 0 0 1 0 1 0 1 ...
```

```
##  $ time                    : num [1:299] 4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : num [1:299] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   anaemia = col_double(),
##   ..   creatinine_phosphokinase = col_double(),
##   ..   diabetes = col_double(),
##   ..   ejection_fraction = col_double(),
##   ..   high_blood_pressure = col_double(),
##   ..   platelets = col_double(),
##   ..   serum_creatinine = col_double(),
##   ..   serum_sodium = col_double(),
##   ..   sex = col_double(),
##   ..   smoking = col_double(),
##   ..   time = col_double(),
##   ..   DEATH_EVENT = col_double()
##   .. )
```

So binary features should be changed to factors, and where 0 represents "no" and 1 "yes" in anaemia, Hihg blood pressure, diabetes, smocking and death event; and in sex feature "female" or "male" respectively. we can also re name some of them for having short and more easy to use names:

```
# changing to factor bolean features
hf <- hf %>% mutate(anaemia = as.factor(if_else(anaemia == 0, "no", "yes")),
              diabetes = as.factor(if_else(diabetes == 0, "no", "yes")),
              high_blood_pressure = as.factor(if_else(high_blood_pressure == 0, "no", "yes")),
              sex = as.factor(if_else(sex == 0, "female", "male")),
              smoking = as.factor(if_else(smoking == 0, "no", "yes")),
              DEATH_EVENT = as.factor(if_else(DEATH_EVENT == 0, "no", "yes")))
# renaming long name features
hf <- rename(hf, cpk = creatinine_phosphokinase,
               eject_f = ejection_fraction,
               hbp = high_blood_pressure,
               serum_c = serum_creatinine,
               serum_s = serum_sodium,
               death_event = DEATH_EVENT)
```

Now we can explore whether there is any NAs, missing values or both:

```
# exploring the data set for NAs
sum(is.na(hf))
```

```
## [1] 0
```

Fortunately this is not the case, sowe can now move forward to have a general view of the information:

```
summary(hf)
```

```
##       age        anaemia        cpk         diabetes     eject_f       hbp
##  Min.   :40.00   no :170   Min.   :  23.0   no :174   Min.   :14.00   no :194
##  1st Qu.:51.00   yes:129   1st Qu.: 116.5   yes:125   1st Qu.:30.00   yes:105
```

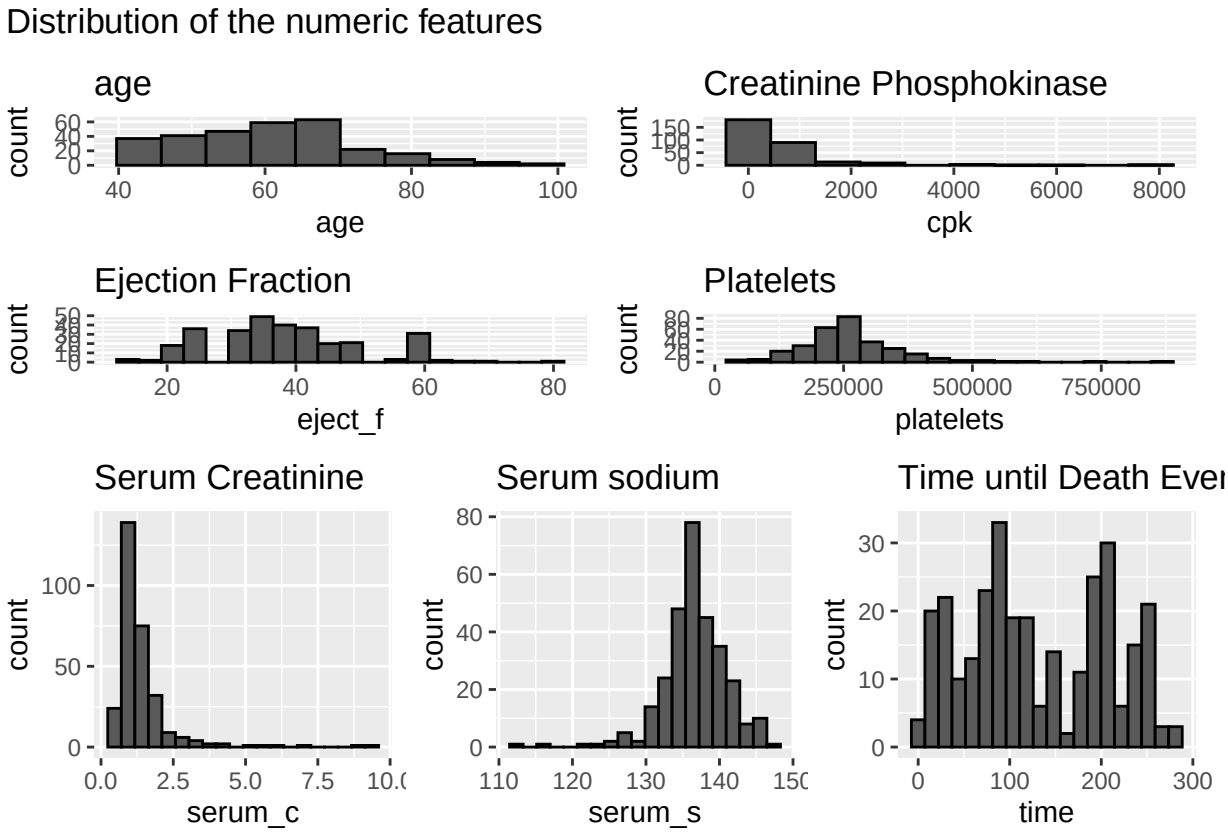
```
##  Median :60.00            Median : 250.0            Median :38.00
##  Mean   :60.83            Mean   : 581.8            Mean   :38.08
##  3rd Qu.:70.00            3rd Qu.: 582.0            3rd Qu.:45.00
##  Max.   :95.00            Max.   :7861.0            Max.   :80.00
##    platelets         serum_c         serum_s          sex      smoking
##  Min.   : 25100   Min.   :0.500   Min.   :113.0   female:105   no :203
##  1st Qu.:212500   1st Qu.:0.900   1st Qu.:134.0   male  :194   yes: 96
##  Median :262000   Median :1.100   Median :137.0
##  Mean   :263358   Mean   :1.394   Mean   :136.6
##  3rd Qu.:303500   3rd Qu.:1.400   3rd Qu.:140.0
##  Max.   :850000   Max.   :9.400   Max.   :148.0
##       time       death_event
##  Min.   :  4.0   no :203
##  1st Qu.: 73.0   yes: 96
##  Median :115.0
##  Mean   :130.3
##  3rd Qu.:203.0
##  Max.   :285.0
```

It shows that the cohort studied has 105 women and 194 men. Other binary features apparently shows that there are more proportion of patients without anaemia, diabetes and high blood preasure, and also not smockers, which is curious considering that some of these are well known risks factors for HF and CVD, nevertheless, right now is to premature to have any conclusions as more in depth analysis should be preformed.

For numeric features, we see that the age range is from 40 to 95 years old. The creatinine phosphokinase shows a very broad range, going form a minimum of 23 through a maximum of 7861 mcg/L, which could be due to some outlayer values that must be analysed considering that the mean and 3rd quartile are close (581.8 and 582.0 respectively). For ejection fraction feature, the clinical community groups heart failure into two types based on the ejection fraction value, that is the proportion of blood pumped out of the heart during a single contraction, given as a percentage with physiological values ranging between 50% and 75%. The former is heart failure due to reduced ejection fraction (HFrEF), known as systolic dysfunction or systolic heart failure and characterized by an ejection fraction smaller than 40. The latter is heart failure with preserved ejection fraction (HFpEF), called diastolic heart failure or heart failure with normal ejection fraction. In this case, the left ventricle contracts normally during systole, but the ventricle is stiff and fails to relax normally during diastole, thus impairing filling. We see in the dataset that the range goes from 14 to 80, with a mean of 38, which probably we have patients both with HFrEF and HFpEF. For platelets, normal counts go from 150.000 to 450.000, and in the dataset we see that the mean is in the midle of that range but there are some extreme values too. For serum creatinine, normal values are among 0.7 and 1.2 mg/dL, but in our cohort we see a little higher mean with 1.394 mg/dL with an incredible maximum of 9.4 mg/dL, wich clearly states that there is some renal impairment in this patients consistent with the HF condicion. in sodium serum values, we see that the mean is 136.6 mEq/L and with minumum of 113.0 mEq/L, which is low of the normal values of 135 to 145 mEq/L and again, it is consistent with the clinical of a heart failure in an advanced stage. It is also important to note that the maximum time before a death event was 285 days, however, the minimum noted was 4 days, which gives a broad range with a mean of 130.3 days. This information could be useful to take into account for our model if we want to see if there is any correlation among other features and the time in days of a death event to ocurr. Last, but not least, there were 96 death events in total, which is a high number considering that the follow up period is less than a year and tells us that probably these patients had suffered of HF from long time before they entered the registry.

2.2 Numeric Feature Distribution and General Information

Visualy, we can see how the features are distributed in the following graphs:



So, taking a look in the summary and the visual of the distribution of the features, we have this information so far:

Numeric:

1. age: patients form 40 to 95 years old
2. creatinine phosphokinase: the less the better, but we found a mean of 581 with huge extreme high values
3. ejection fraction: mean of 38.0.8, let´s say by know that we have balanced HFrEF and HFpEF
4. platelets: in average 263.000, but with extreme values
5. serum creatinine: normal values are among 0.7 and 1.2 mg/dL but also with some extreme values
6. serum sodium: normal values range from 135 mEq/L to 145 mEq/L, but with a mean in our cohort near the lower limit (136.6mEq/L) and more extreme lowe minimal values
7. time: mean around 130 days but broad spectrum with a minimum of 4 days and a maximum of 285. All death events in less than a year time.
8. death event: 96 deaths vs 203.

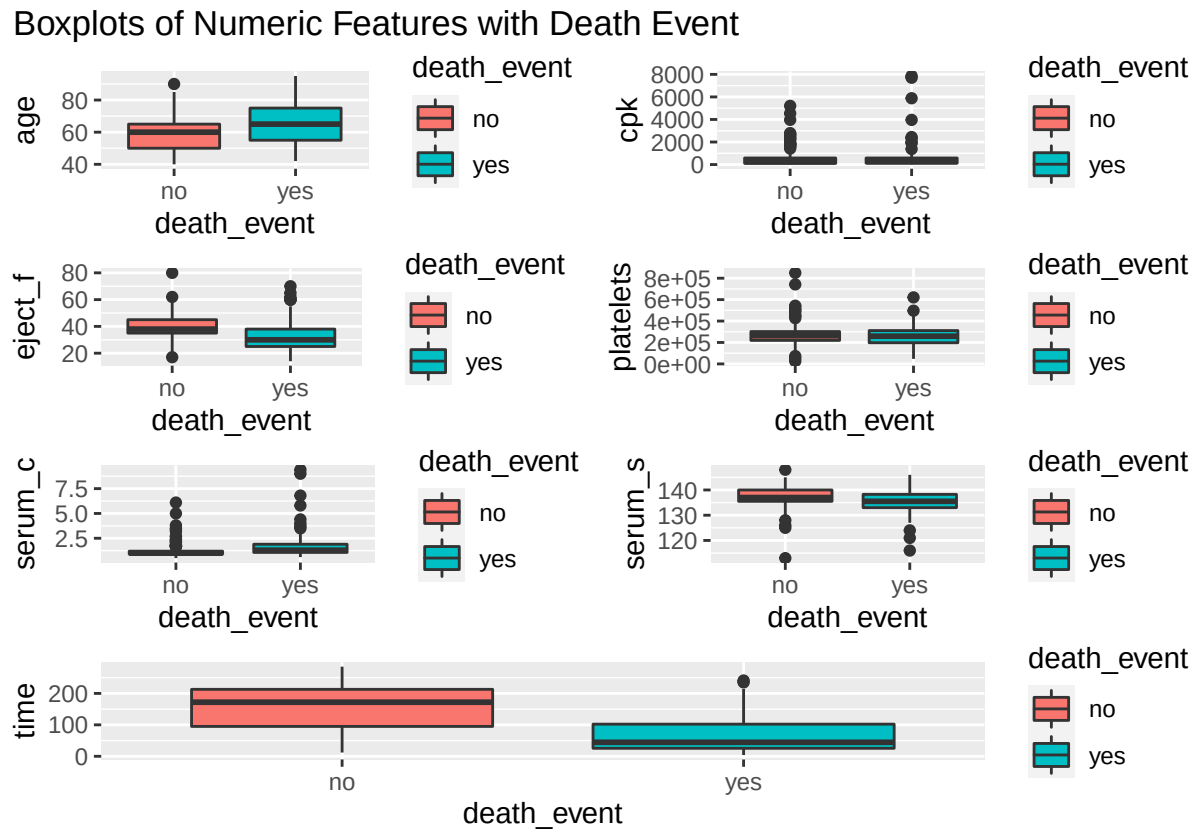
Cathegorical:

1. anaemia: more people without it (170 vs 129). Need to see correlation with death events
2. diabetes: more people without it (174 vs 125). Need to see correlation with death events
3. high blood preasure: more people without it (194 vs 105). Need to see correlation with death events
4. sex: 194 men and 105 women
5. smocking: more not smocking people (203 vs 96). Need to see correlation with death events.

we need now to take a closer look of the features in order to see if there is correlation among them

2.3 Correlations

Starting with the numeric features, we need to take a look to their correlation with the death event feature, and is it shown in the next graph, we can see that by grouping them by death event, some of them visualy do not change so much based on that they have similar results for both conditions (their means apparently are at the same level). This is the case in example, for creatinine phoshokinase or platelets; nevertheless, we see that some others apparently do have differences, as it´s seen in ejection fraction, serum creatinine, serum sodium and time:



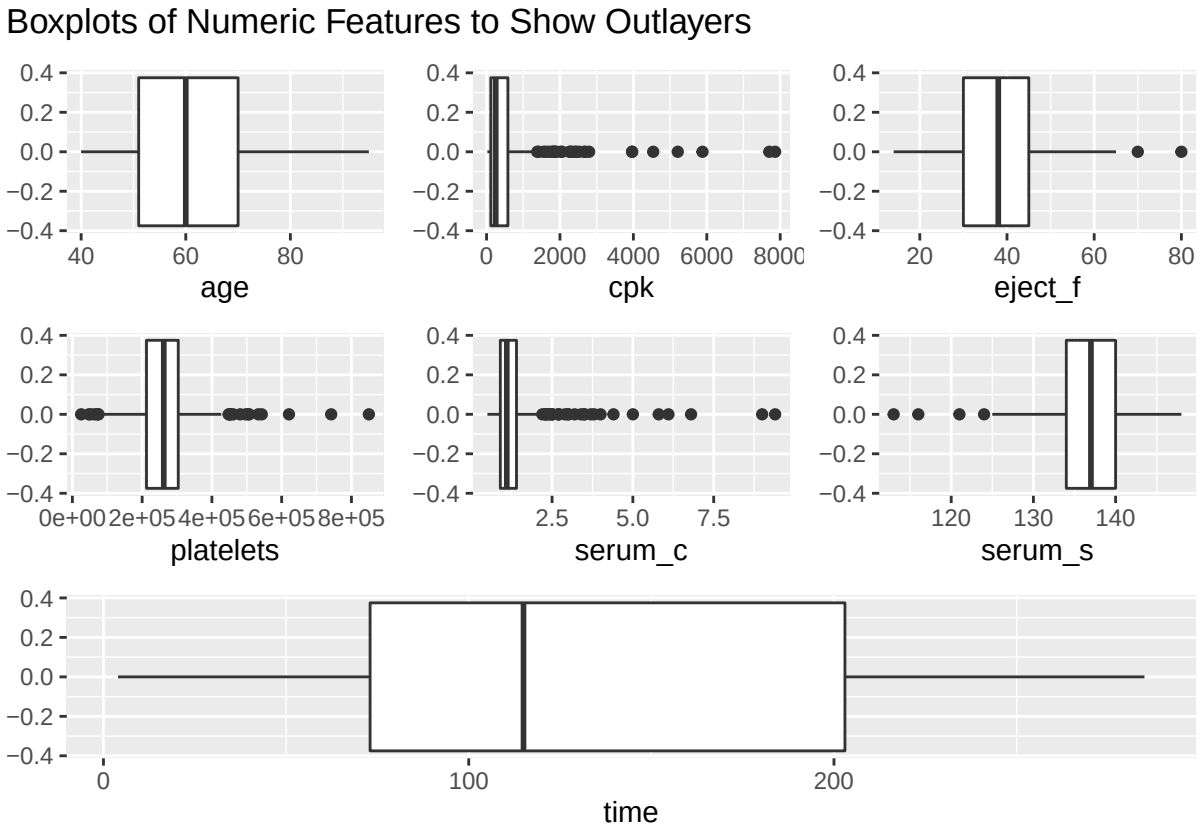
To have a formal confirmation of these assumtions, we can pass a normal distribution test to them.

2.3.1 Normal Distribution Test

Before we perform any test, we need to confirm first if the features are normal distributions. As we have more than 50 observations, a Kolmogorov-Smirnov test can help us:

```
##           p.value
## age        0.0013
## cpk        0.0000
## eject_f    0.0000
## platelets  0.0000
## serum_c    0.0000
## serum_s    0.0000
## time       0.0000
```

As none of them are normal distributions (all are < 0.05), we should take a look to the outlayers as probably they are messing all things up:



As we can see, only the age and time appear to be clear of extreme values, the rest of them have outlayers, so we need to take them out to see whether things get better in terms of normality for distribution. so after taking out the oulayers for each feature, we can see that now all of them are normally distributed:

```
##    Features      p.value
## 1       age 1.304201e-03
## 2       cpk 8.070210e-23
## 3   eject_f 4.214660e-21
## 4 platelets 4.296722e-03
## 5   serum_c 3.742716e-18
## 6   serum_s 3.903684e-61
## 7      time 2.009876e-08
```

2.3.2 Homogeneity of Variances Test

Now we have to see if we have homogeneity of variances:

```
##    Features p.value
## 1       age  0.0080
## 2       cpk  0.5691
## 3   eject_f  0.0630
## 4 platelets  0.1734
## 5   serum_c  0.0001
## 6   serum_s  0.0653
## 7      time  0.0051
```

These tests show that we have homogeneity of variances in the cpk, ejection fraction, platelets and serum sodium features. This is important to arrange the t-test function as we need to declare if the variances have or not homogeneity within the formula.

Finally, we are all set with the numeric features to see which of them doesn´t have statistical differences comparing their means by death event. This step will help us out to determine which numeric features are relevant to consider in our prediction model.

2.3.3 T of Student Test

```
##    Features p.value
## 1       age  0.0000
## 2       cpk  0.9826
## 3   eject_f  0.0000
## 4 platelets  0.5930
## 5   serum_c  0.0000
## 6   serum_s  0.0022
## 7      time  0.0000
```

2.3.4 Correlation of Cathegorical Features with Death Event

Moving forward, we need to see if there is correlation of the cathegorical features and death events, so we can use the chi squared test to see this.

```
##                        X_squared   p_value
## anaemia.X-squared  1.042175e+00 0.3073161
## diabetes.X-squared 2.161684e-30 1.0000000
## hbp.X-squared      1.543461e+00 0.2141034
## sex.X-squared      0.000000e+00 1.0000000
## smoking.X-squared  7.331474e-03 0.9317653
```

As we can see, none of them have a p value < 0.05, so based on this we can tell that none of the cathegorical features is related with the death event (they are independent of it) so we do not bother to include any of them in our model.

2.4 The Models

Unfortunately for the model, we saw that several features are independent (cathegoricals) or have the same mean (numerical) of the death event, which means that the data has not statistical significance in their results so we cannot assume the results for actual clinical desitions. Nevertheless, and for academic purposes, we can train some models in order to use the dataset to predict death events in new hf patients

First, we have to partitionate the dataset into training and test sets so we can build the models in the train set and then test them in the test set:

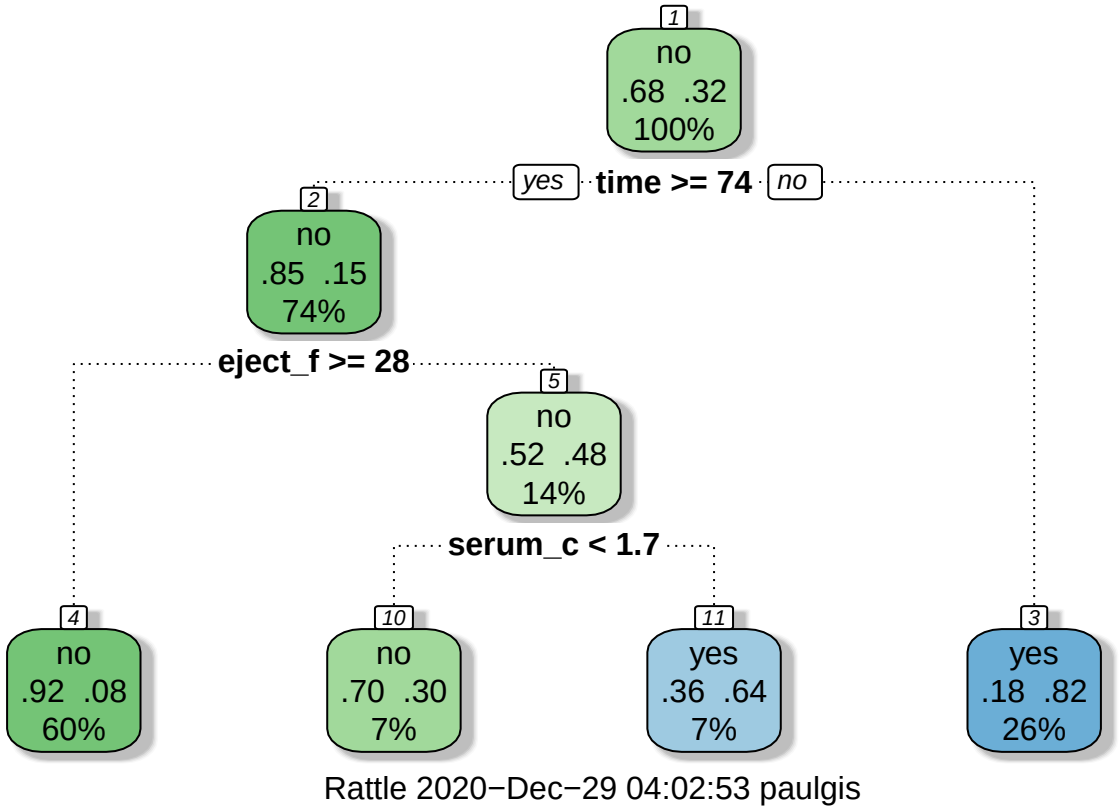
```
# creating training and test sets
set.seed(123)
test_index <-  createDataPartition(y =hf$death_event, times = 1, p = 0.5, list = FALSE)
train_set <- hf[-test_index, ]
test_set <- hf[test_index, ]
```

```
## Warning: The 'i' argument of ''['()' can't be a matrix as of tibble 3.0.0.
## Convert to a vector.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_warnings()' to see where this warning was generated.
```

2.4.1 Desition Tree

Using the train set, we can grow a desition tree based on the relevance of the features:

```
## Warning in set.seed(0, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```



As we can see, it only includes time and serum creatinine as key features for modeling, but despite that, we can obtain an accuracy of 81.33%:

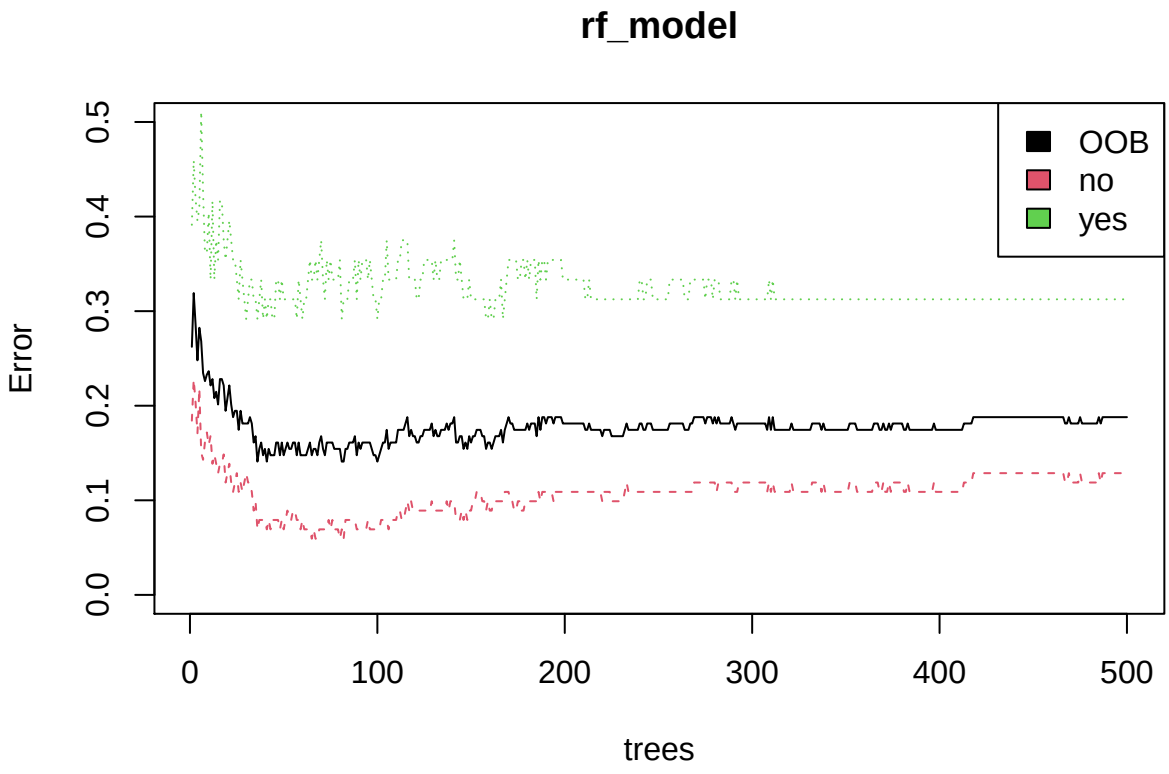
```
# Using desition tree for prediction
prediction_tree <- predict(tree_hf, test_set, type = "class")
# Save the solution to a dataframe with two columns: death event and prediction
solution <- data.frame(death_event = test_set$death_event, prediction = prediction_tree)
solution <- solution %>% mutate(comparison = if_else(death_event == prediction, 1, 0))
accuracy <- sum(solution$comparison)/nrow(solution)*100
accuracy
```

```
## [1] 85.33333
```

2.4.2 Random Forest

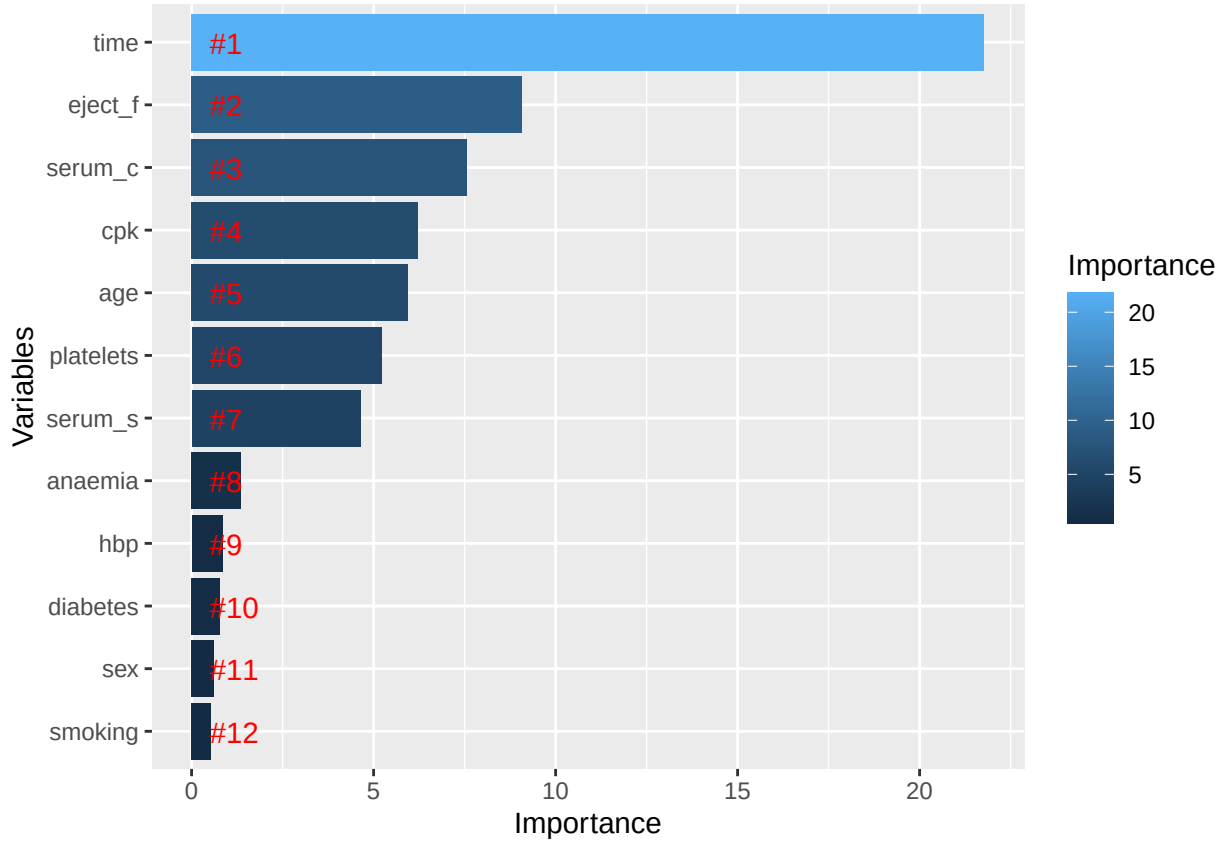
According to the course materials for the analysis, it stated that at least 2 models should be trained to accomplish the minimal requirements for this project, so based in the model of the desition tree, we can move forward and try a random forest model:

```
## Warning in set.seed(456, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```



We see that it gives an interesting result, where the OOB estimate oferror rate is 14.77%, which in other terms gives an accuracy of 85.23%, similar to the one obtained with the desition tree.

Finally, we can have the variables importance to see the most important features in the model:



Interestingly, we see that in this model, time is the most relevant feature, followed by serum creatinine and ejection fraction.

3. Results

Based in the proposed model, we found out that in the dataset 5 of 7 numerical features did not have normal distributions, which led to find several outlayer values. Not to mention the cathegorical values, where was demonstrated that all of them were independent of the death event; taking out the extreme values in the numerical features resulted in the similarity of the means between each feature and the death event, which supported the non statistical differences. Following with the models, the desition tree and the random forest, these considerations were not taken into account and the algorithms used the whole dataset including the outlayer values, resulting in similar accuracy despite different features took relevance in each model.

4. Conclusions

- There is no information about the previous medical history of the patients and for how much time they had the diagnosis of the HF before entering the registry, so it is difficult to find correlation between variables or stablish any clinical significance.
- There is no information about when the tests or mesurements in the dataset were taken once the patients entered the registry, which makes it impossible to know if they were taken at the begining or any other moment during the patinet´s follow up.
- There is no information about any treatments taken before or during the follow up period, so it is no clear the clinical baseline status of the patients included in the registry and whether the HF was

controlled or not, or if the death event was attibutable to the HF or any other cause.

- The results of the models presented in this project are only for academic purposes and cannot be considered for clinical desitions, as more detailed information should be included for deeper analysis.