UNIVERSITÀ DEGLI STUDI DI CATANIA

DIPARTIMENTO DI ECONOMIA E IMPRESA

CORSO DI LAUREA IN DATA SCIENCE FOR MANAGEMENT

HEART FAILURE PREDICTION

Casella Bruno
1000014143

Advanced Machine Learning

ACADEMIC YEAR 2020 – 2021

"HEART FAILURE CLINICAL RECORDS" DATASET

The dataset used in this report is the "Heart Failure Clinical Records" available at the following link:
https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
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 this dataset contains 12 features that can be used to predict mortality by heart failure.
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.
This dataset contains 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 dataset contains 13 features, which report clinical, body, and lifestyle information. The first 12 features are the input features, and the 13th is the target variable:

1. Age: Age of the patient
2. Anaemia: Decrease of red blood cells or hemoglobin (Boolean)-0=No, 1=Yes
3. Creatinine Phosphokinase: Level of the CPK enzyme in the blood (mcg/L)
4. Diabetes: If the patient has diabetes (Boolean)-0=No, 1=Yes
5. Ejection fraction: Percentage of blood leaving the heart at each contraction (percentage)
6. High blood pressure: If the patient has hypertension (Boolean)-0=No, 1=Yes
7. Platelets: Platelets in the blood (kiloplatelets/mL)
8. Serum creatinine: Level of serum creatinine in the blood (mg/dL)
9. Serum sodium: Level of serum sodium in the blood (mEq/L)
10. Sex: Woman or Man (binary) -0=Female, 1=Male
11. Smoking: If the patient smokes or not (Boolean)-0=No, 1=Yes
12. Time: Follow-up period (days)
13. DEATH EVENT: If the patient deceased during the follow-up period (Boolean)-0=No, 1=Yes. This is the target variable.

DATA ANALYSIS AND VISUALIZATION

The dataset contains 299 rows, that are the medical records of the patients, 194 males and 105 females, and their ages range between 40 and 95 years old. All the features are numerical. In the dataset there are not NULL values.
All the descriptive statistics are contained in the 2 pictures below, that represent the transposition of the original matrices, only for a better plotting.

	0	1	2	3	4	5	6	7	8	9	...	289	290	291	292	293
age	75	55	65	50	65	90	75	60	65	80	...	90	45	60	52	63
anaemia	0	0	0	1	1	1	1	1	0	1	...	1	0	0	0	1
creatinine_phosphokinase	582	7861	146	111	160	47	246	315	157	123	...	337	615	320	190	103
diabetes	0	0	0	0	1	0	0	1	0	0	...	0	1	0	1	1
ejection_fraction	20	38	20	20	20	40	15	60	65	35	...	38	55	35	38	35
high_blood_pressure	1	0	0	0	0	1	0	0	0	1	...	0	0	0	0	0
platelets	265000	263358	162000	210000	327000	204000	127000	454000	263358	388000	...	390000	222000	133000	382000	179000
serum_creatinine	1.9	1.1	1.3	1.9	2.7	2.1	1.2	1.1	1.5	9.4	...	0.9	0.8	1.4	1	0.9
serum_sodium	130	136	129	137	116	132	137	131	138	133	...	144	141	139	140	136
sex	1	1	1	1	0	1	1	1	0	1	...	0	0	1	1	1
smoking	0	0	1	0	0	1	0	1	0	1	...	0	0	0	1	1
time	4	6	7	7	8	8	10	10	10	10	...	256	257	258	258	270
DEATH_EVENT	1	1	1	1	1	1	1	1	1	1	...	0	0	0	0	0

Figure 1 - Some rows of the original dataset

	mean	std	min	25%	50%	75%	max
age	60.833893	11.894809	40.0	51.0	60.0	70.0	95.0
anaemia	0.431438	0.496107	0.0	0.0	0.0	1.0	1.0
creatinine_phosphokinase	581.839465	970.287881	23.0	116.5	250.0	582.0	7861.0
diabetes	0.418060	0.494067	0.0	0.0	0.0	1.0	1.0
ejection_fraction	38.083612	11.834841	14.0	30.0	38.0	45.0	80.0
high_blood_pressure	0.351171	0.478136	0.0	0.0	0.0	1.0	1.0
platelets	263358.029264	97804.236869	25100.0	212500.0	262000.0	303500.0	850000.0
serum_creatinine	1.393880	1.034510	0.5	0.9	1.1	1.4	9.4
serum_sodium	136.625418	4.412477	113.0	134.0	137.0	140.0	148.0
sex	0.648829	0.478136	0.0	0.0	1.0	1.0	1.0
smoking	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0
time	130.260870	77.614208	4.0	73.0	115.0	203.0	285.0
DEATH_EVENT	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0

Figure 2 - Descriptive statistics

I have plotted the boxplots of the features that are not Boolean, in order to detect outliers.

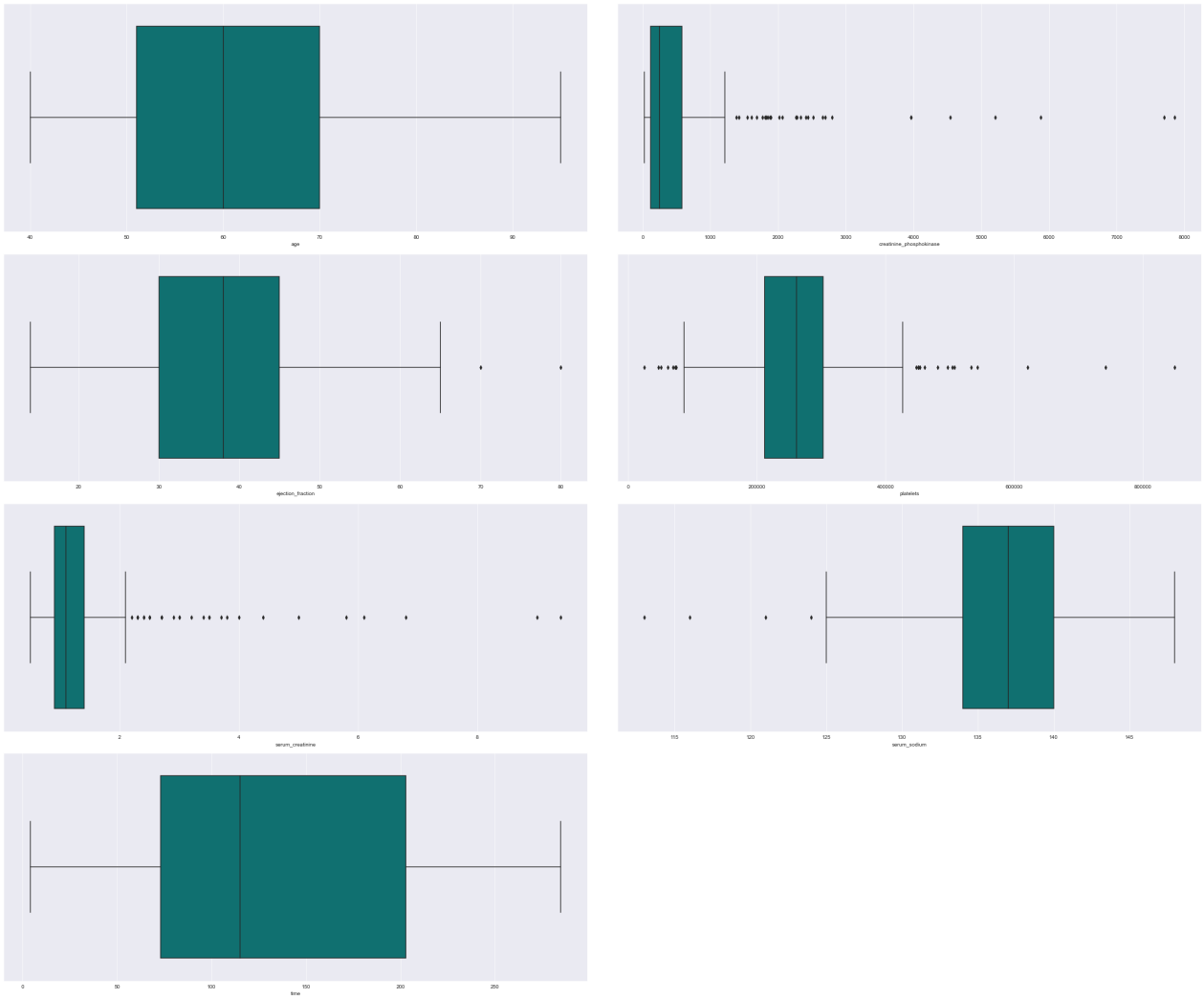


Table 1 – Boxplots

It is possible to see that in age and time there are not outliers, instead in all the other features there are outliers. However, before dealing with outliers we require knowledge about them, the dataset and possibly some domain knowledge. Removing outliers without a good reason will not always increase accuracy. Without a deep understanding of what are the possible ranges that exist within each feature, removing outliers becomes tricky.
So, for this work I will not remove the outliers because I do not have any knowledge about the possible ranges of the features, but for a better analysis with the help of a domain expert you could remove the outliers and try to obtain a better performance of the models.

Now, let's try to find some insights using visualizations, starting by asking us if age and sex are indicators for Death Event.

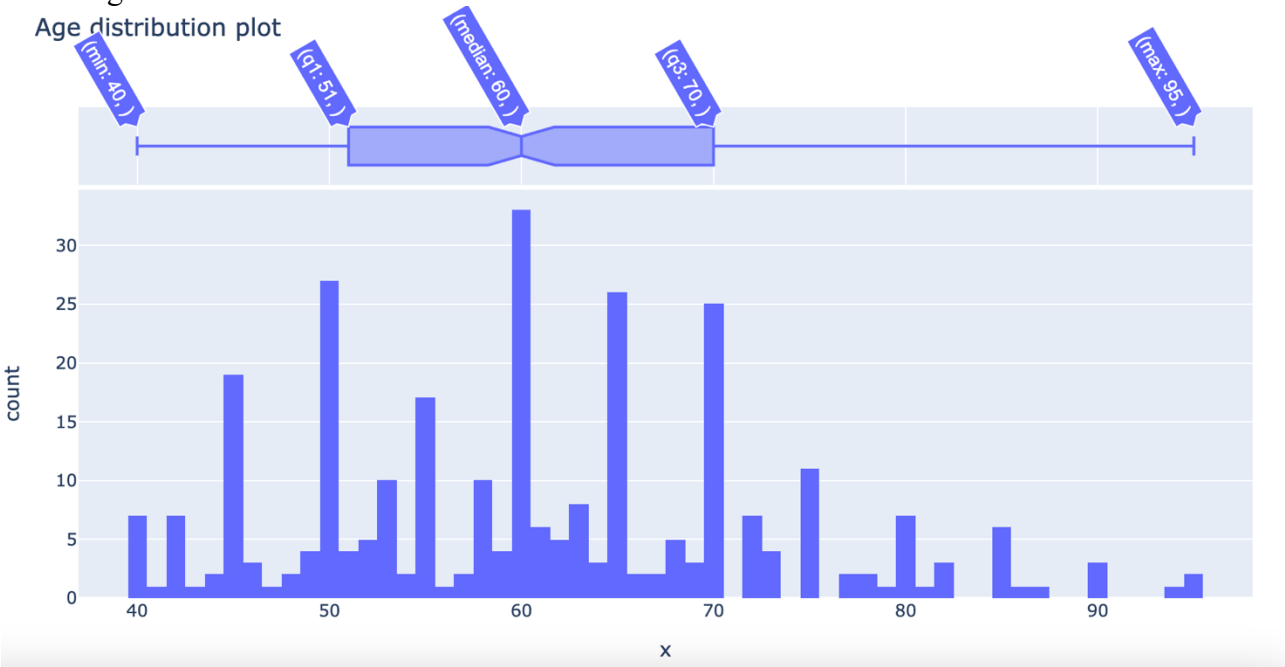


Figure 3 - Age Distribution Plot

Age wise 40 to 70 the spread is high; higher than 80 age people are very low.

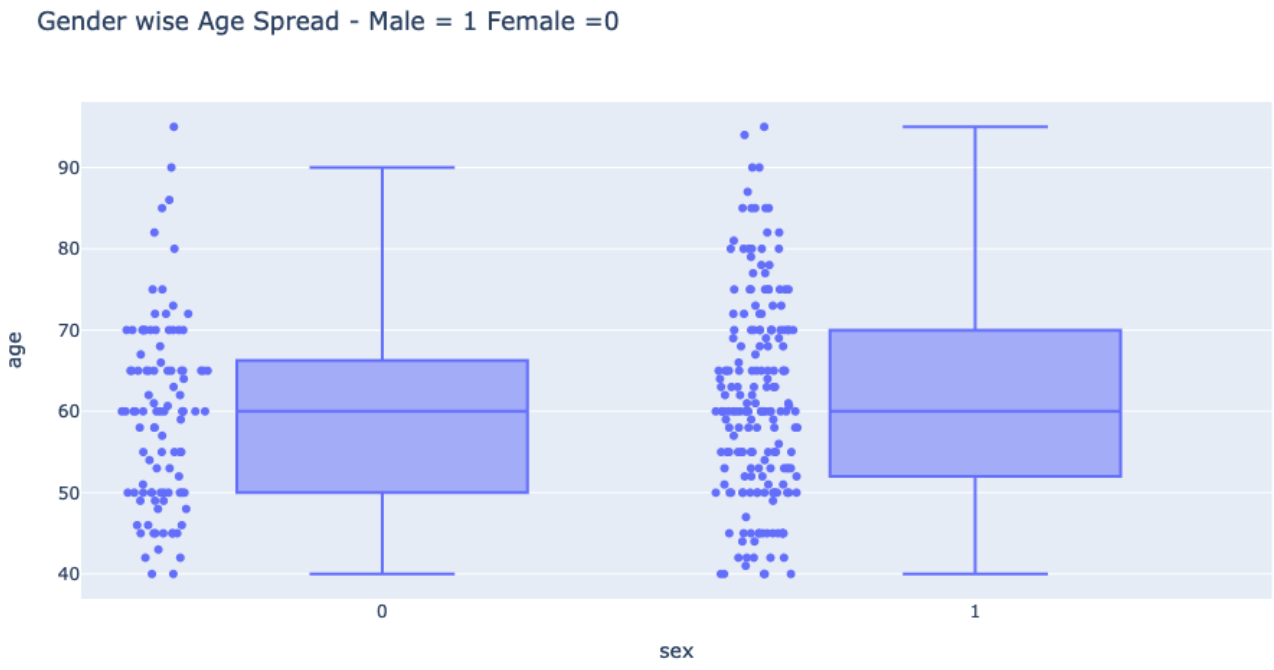


Figure 4 - Gender wise Age spread

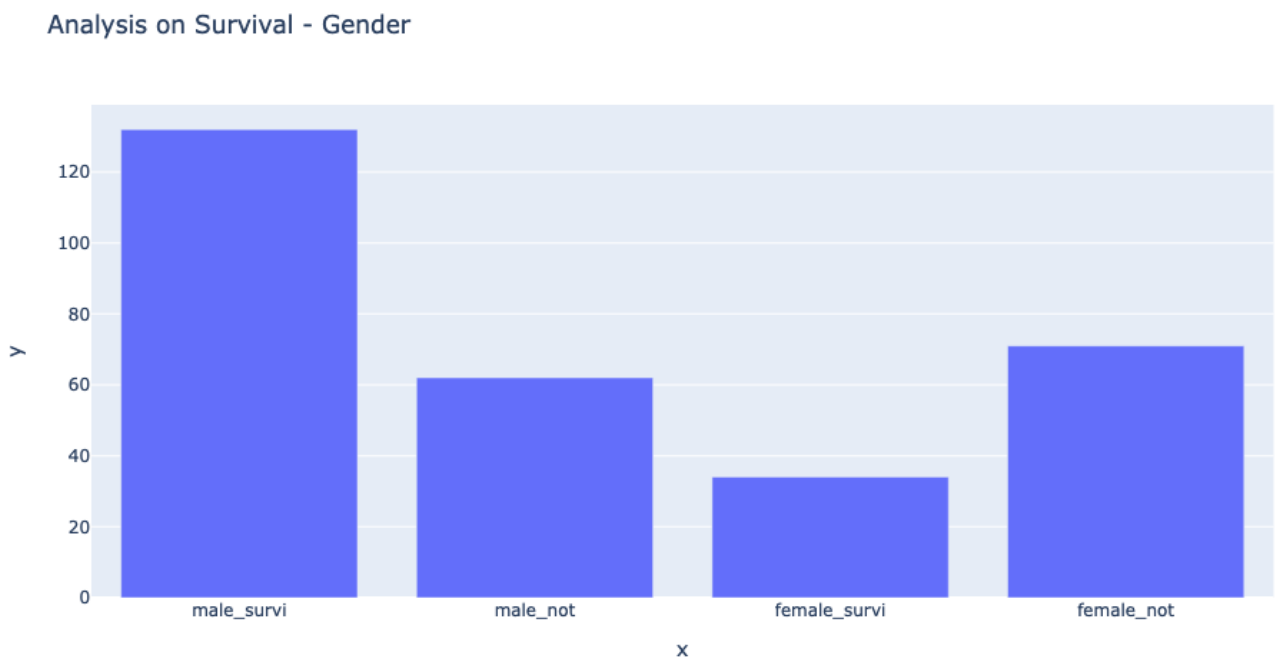


Figure 5 - Analysis on Survival – Gender

In percentage: male survived are 44.1%, female survived 23.8%, male not survived 20.7% and female not survived 11.4%

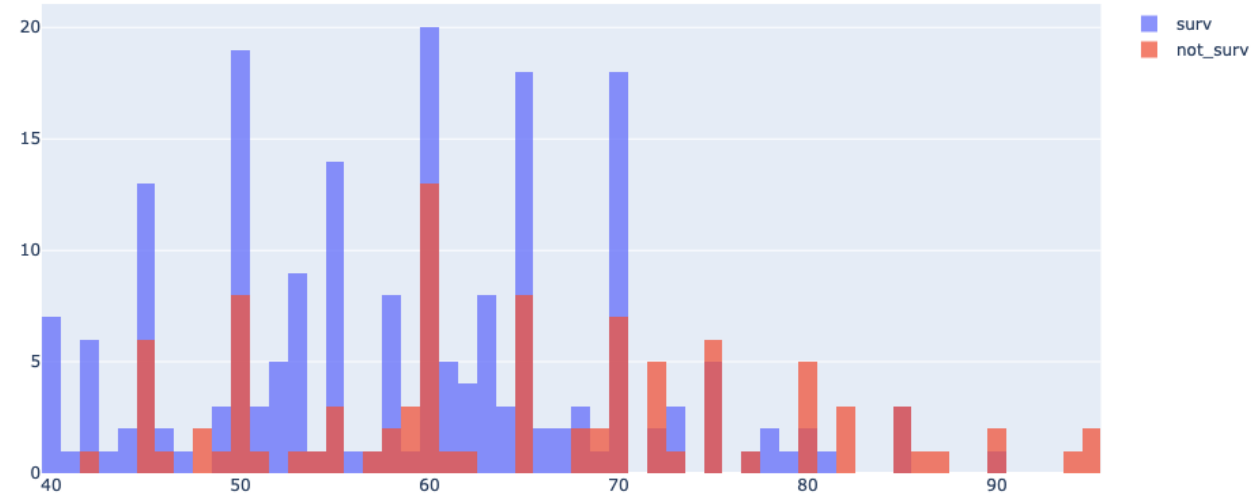


Figure 6 - Analysis in Age on Survival status

We can see that Survival is high on 40 to 70 and that Not Survival spread through all the ages.
Note that the blue and red bars are overlapped.

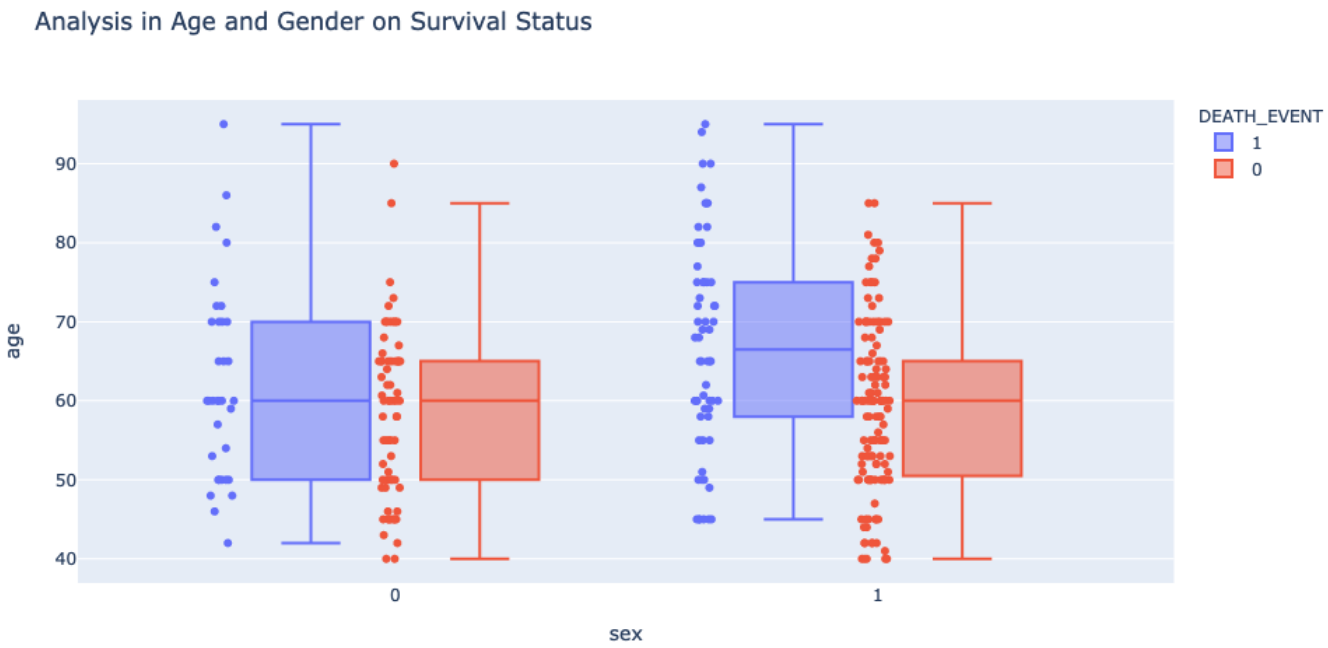


Figure 7 - Analysis in Age and Gender on Survival Status

Survival spread is high on 40 to 70. In particular for male is high between 50 and 60, and for female between 60 and 70.

Figure 8 - Analysis in Age and Smoking on Survival Status

Survival is high for not smoking people on 60 to 70, instead for smoking people on 50 to 60. In general Survival level is lower for smoking people than not smoking person.

Now, let's look at Diabetes, Anaemia and High Blood Pressure.

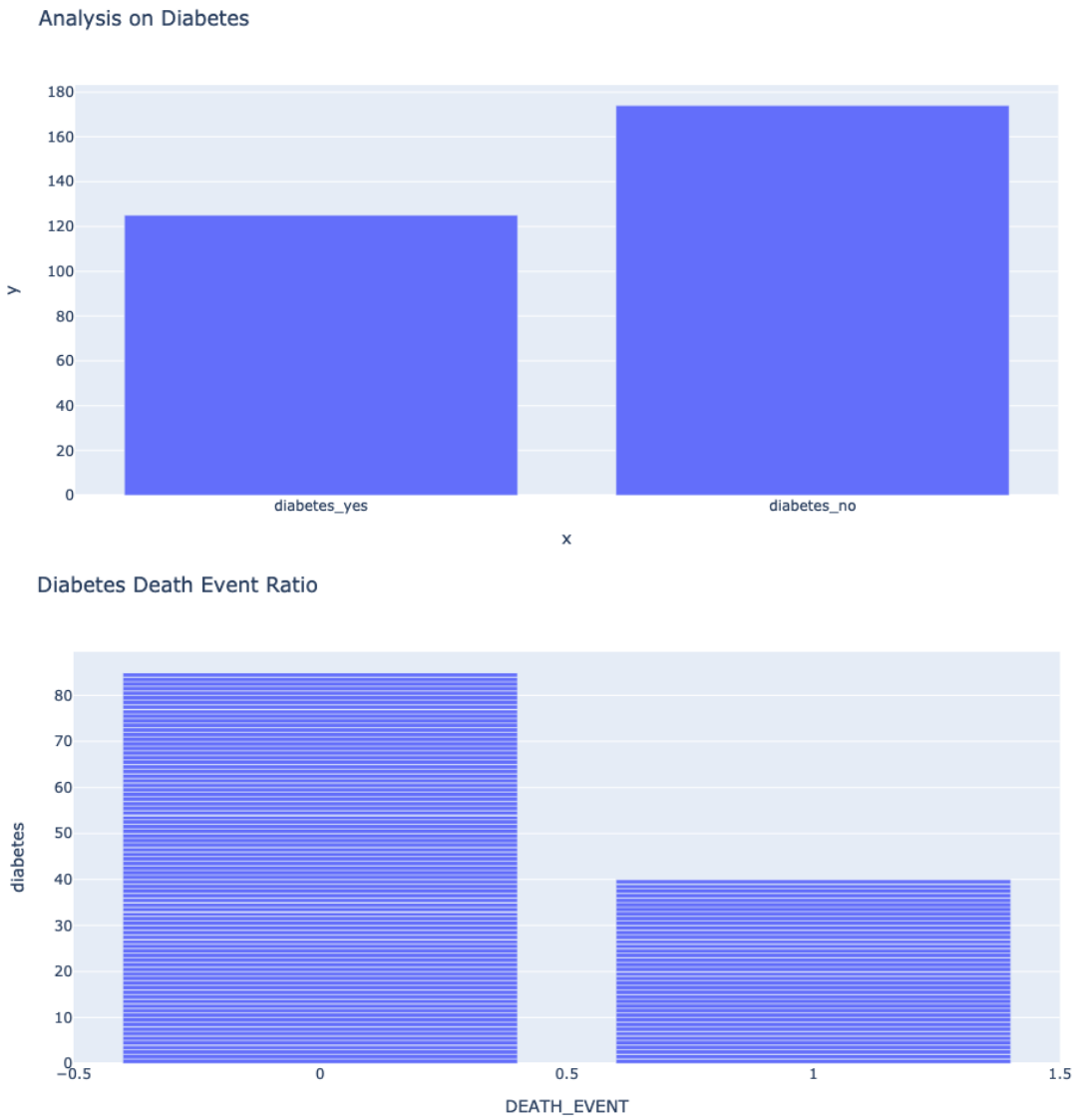


Table 2 – Hists of Diabetes

From these charts we can see that there are 174 people (58.2%) without diabetes and 125 (41.8%) with diabetes. Of the 174 Non diabetes person, 118 survived and 56 have not survived. From 125 with diabetes, 85 persons survived and 40 have not survived.

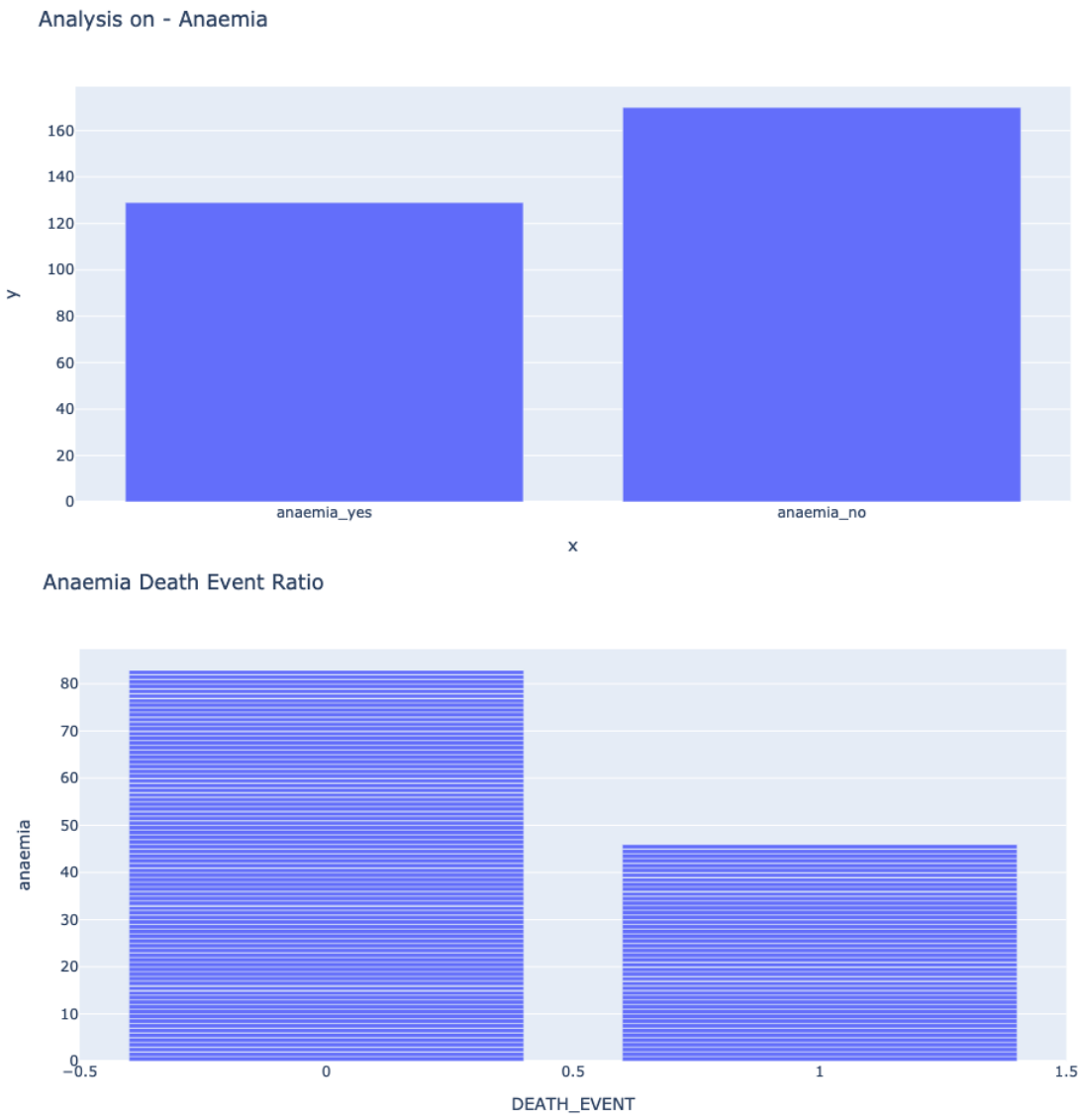


Table 3 – Hists of Anaemia

From these charts we can conclude that in our dataset 170 people (56.9%) are Non anaemic and 129 (43.1%) are anaemic. From Non anaemic 120 are survived and 50 not survived. From anaemic 83 survived and 46 not survived.

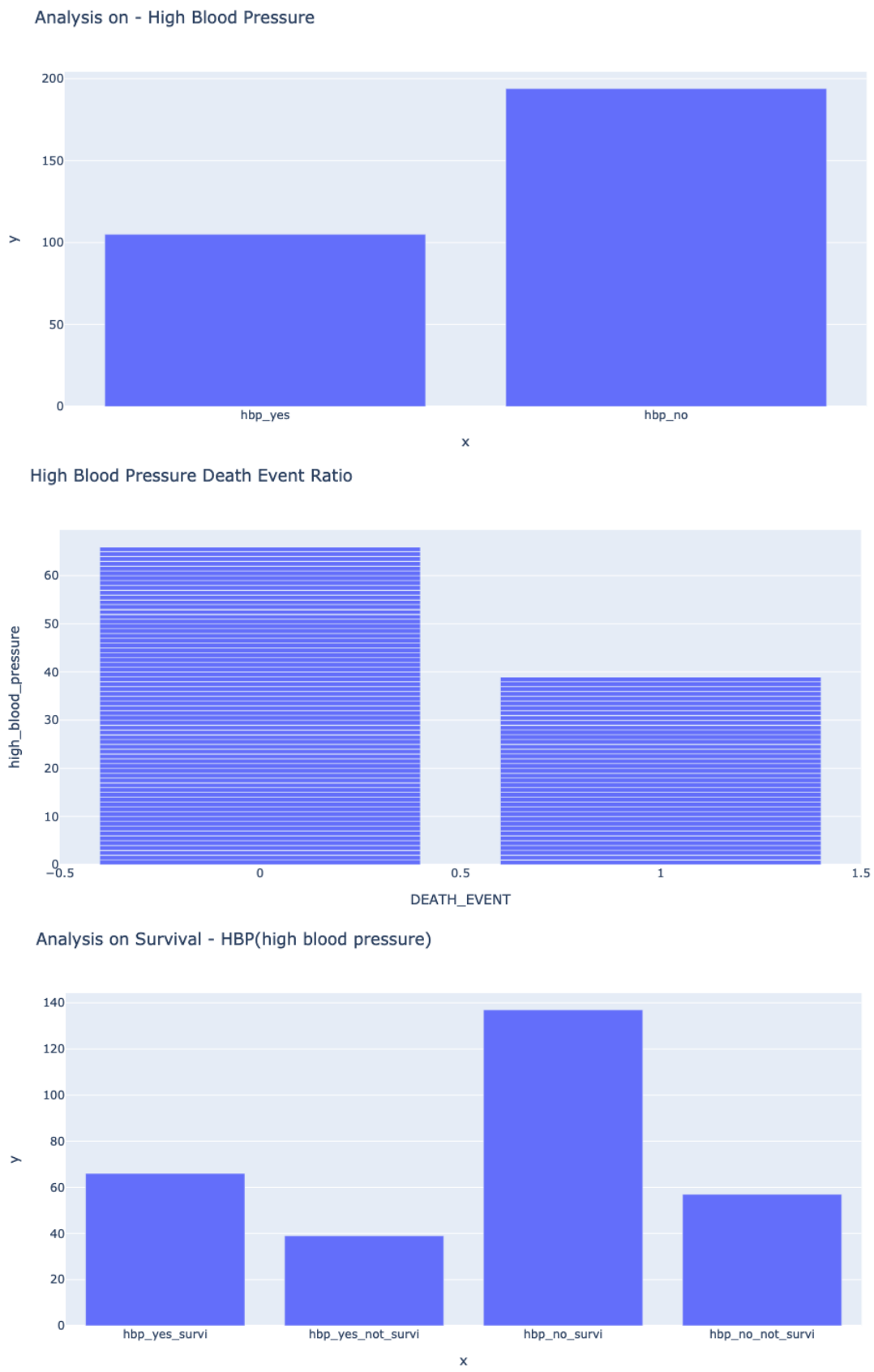
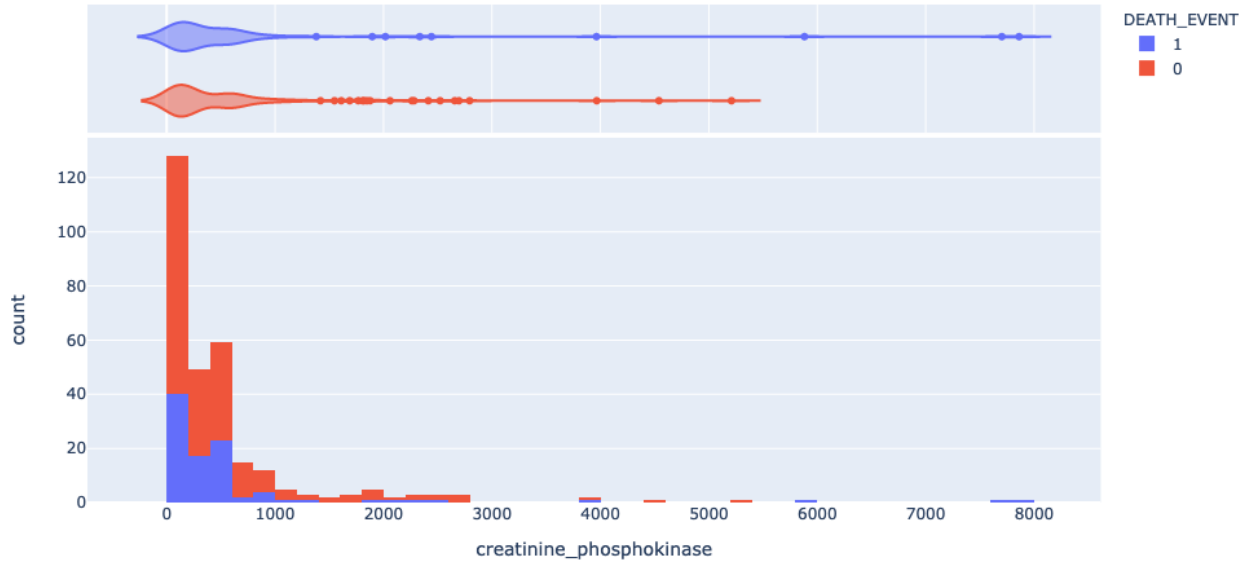
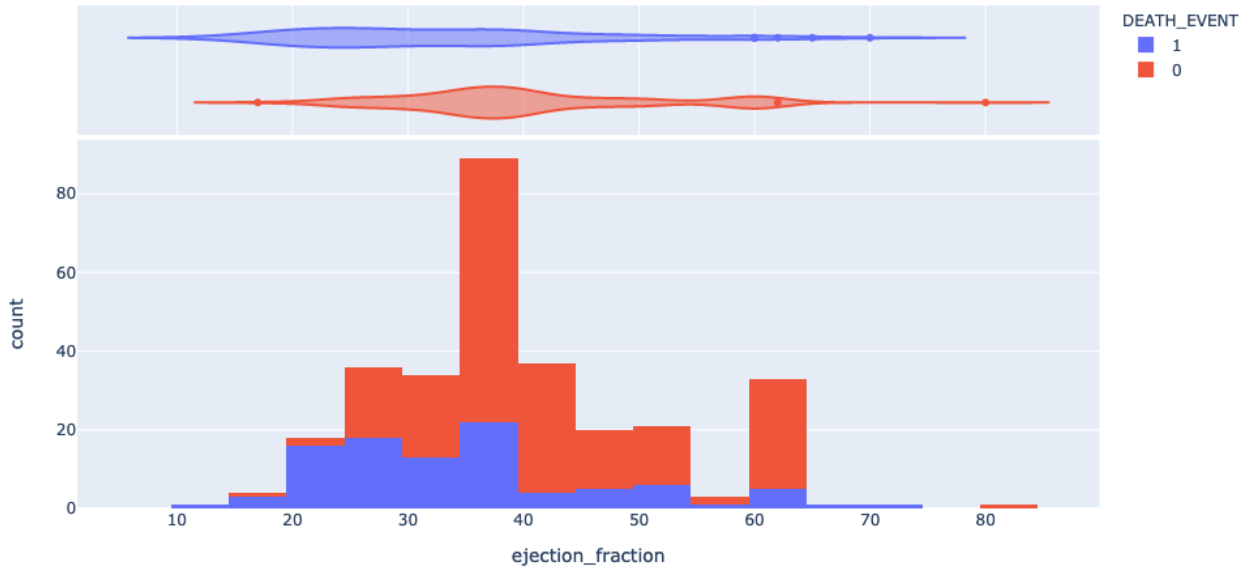
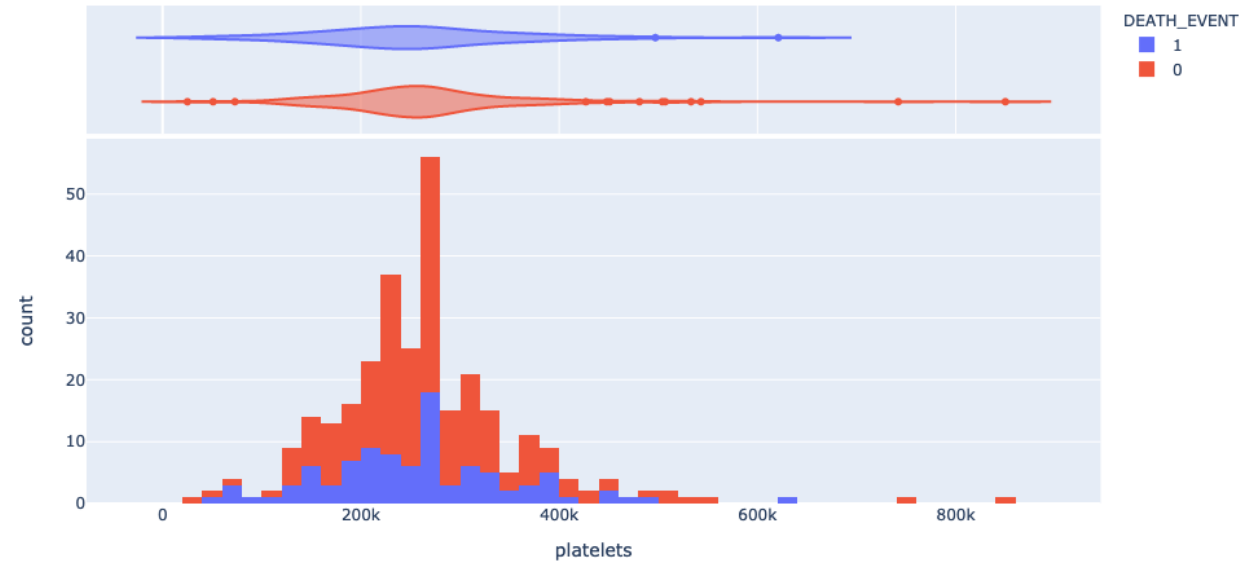


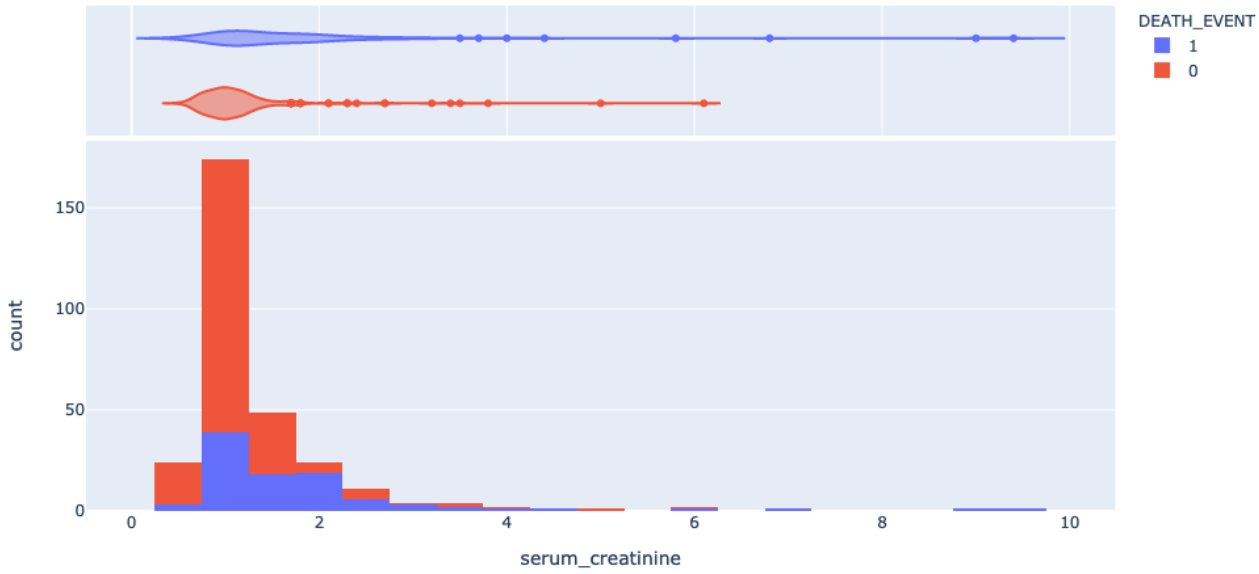
Table 3 – Hists of High Blood Pressure

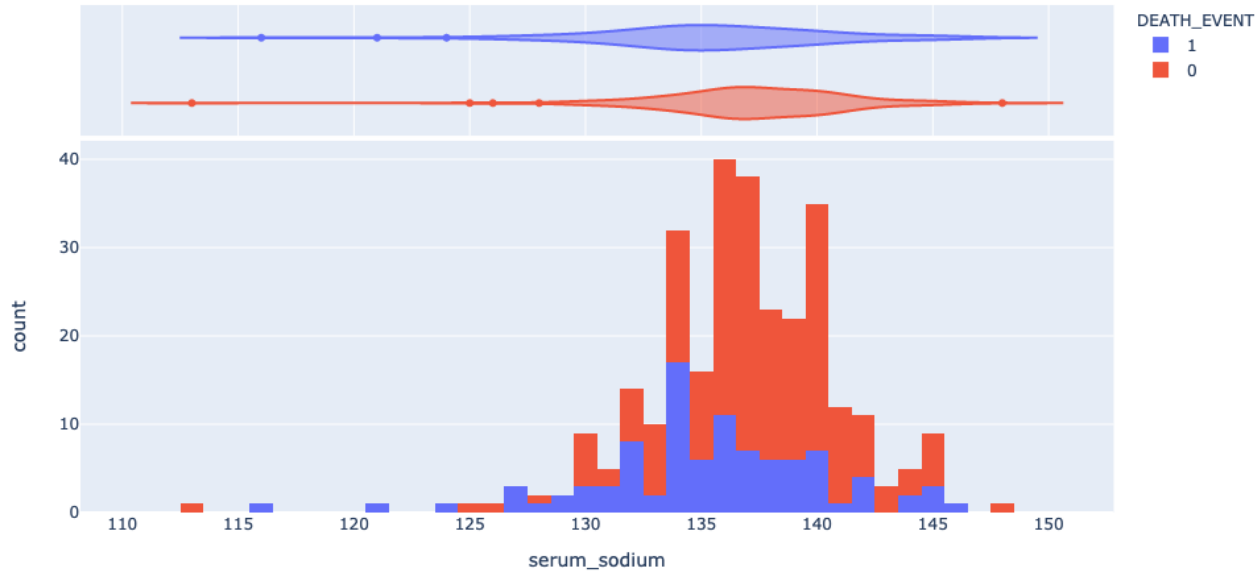
From these graphs we can say that there are 194 people (64.9%) that are not HBP and 105 (35.1%) that are HBP. From 194 not HBP, 137 survived and 57 no. From 105 HBP, 66 survived and 39 no.
Finishing, insights from the last variables.


DEATH_EVENT
1
0
count
120
100
80
60
40
20
0
0
1000
2000
3000
4000
5000
6000
7000
8000
creatinine_phosphokinase

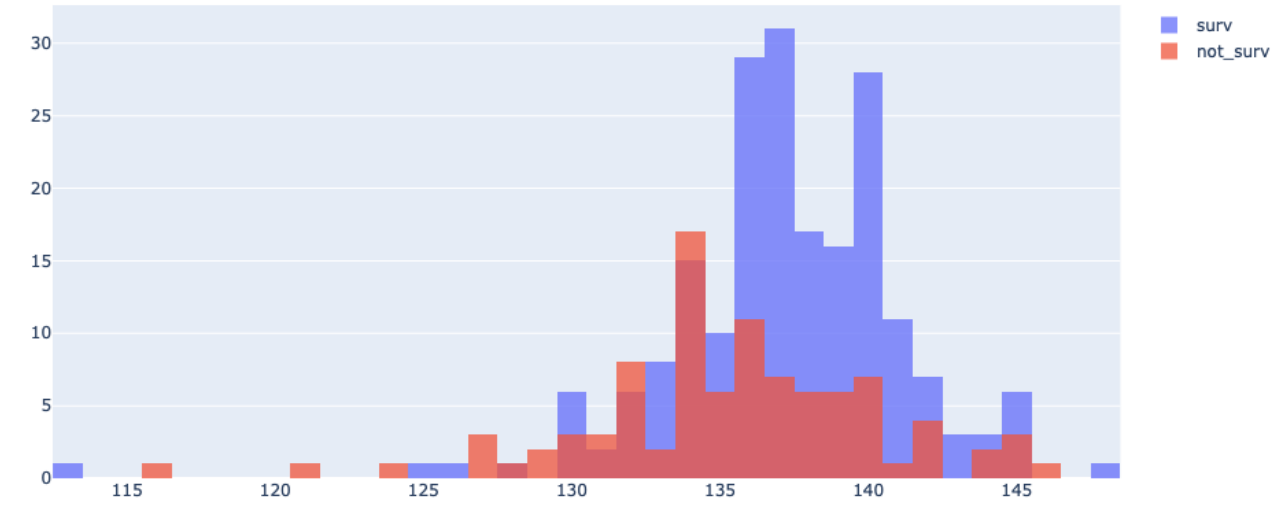

DEATH_EVENT
1
0
count
80
60
40
20
0
10
20
30
40
50
60
70
80
ejection_fraction


DEATH_EVENT
1
0
count
50
40
30
20
10
0
0
200k
400k
600k
800k
platelets





Analysis in Serum Sodium on Survival Status



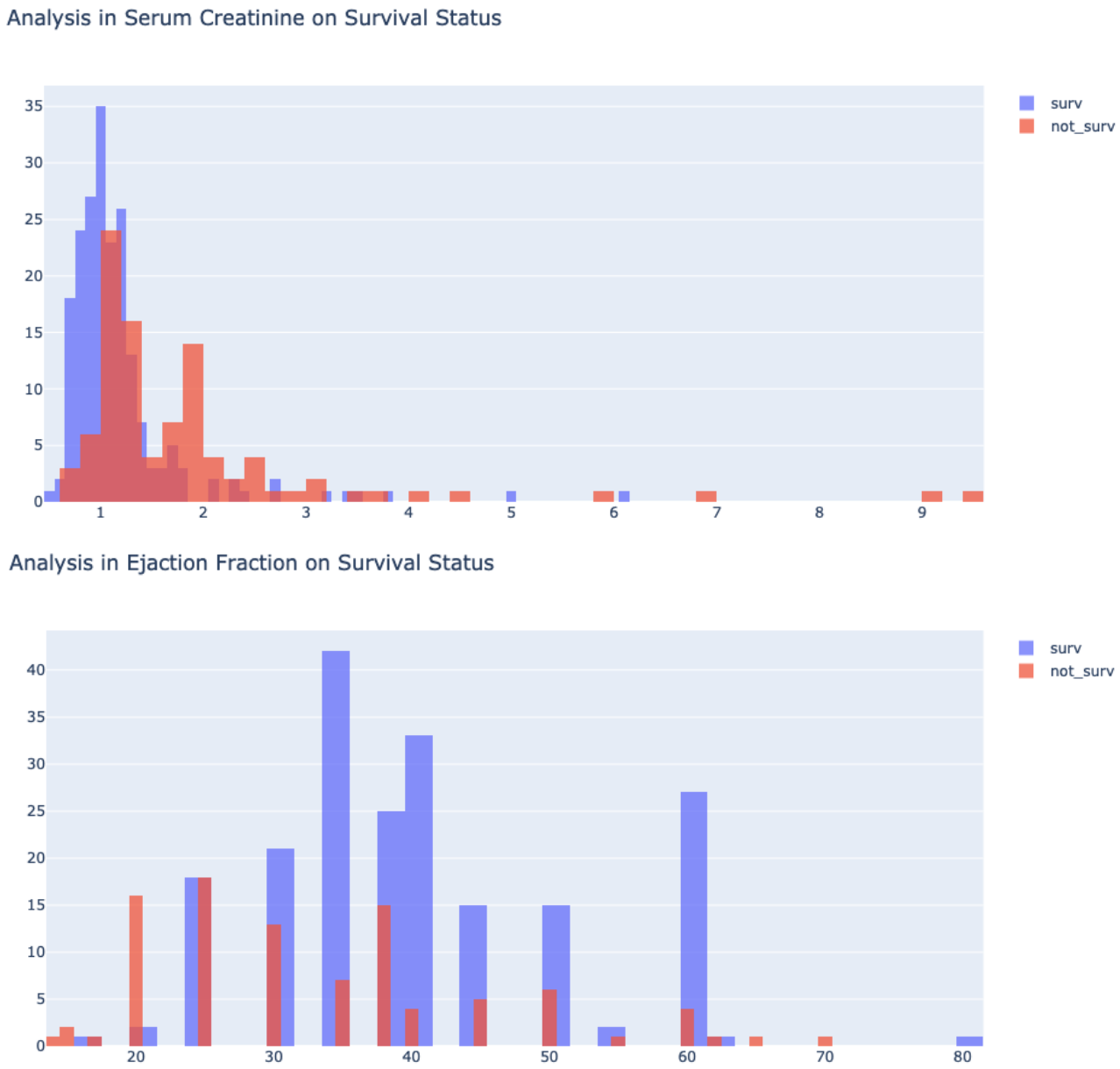


Table 5 – Hists of other factors

All these histograms show the relationship between the Survival level and the variable (creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium).

DATA MODELING

First of all, I investigated the most important features of the cardiovascular heart disease patient's dataset. I plotted the heatmaps showing the correlation between variables, using 3 different methods: Pearson, Spearman and Kendall correlation.

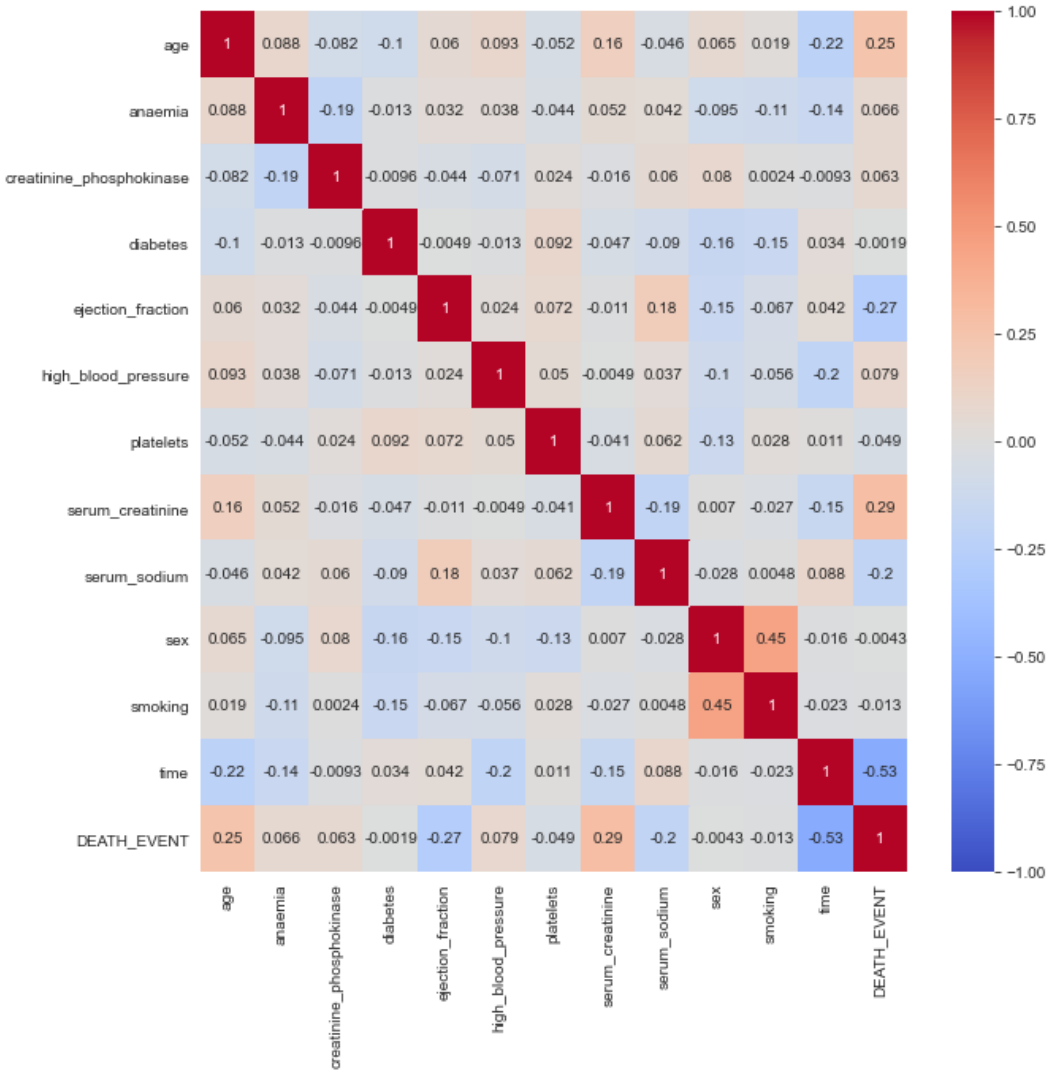


Figure 9 - Pearson correlation

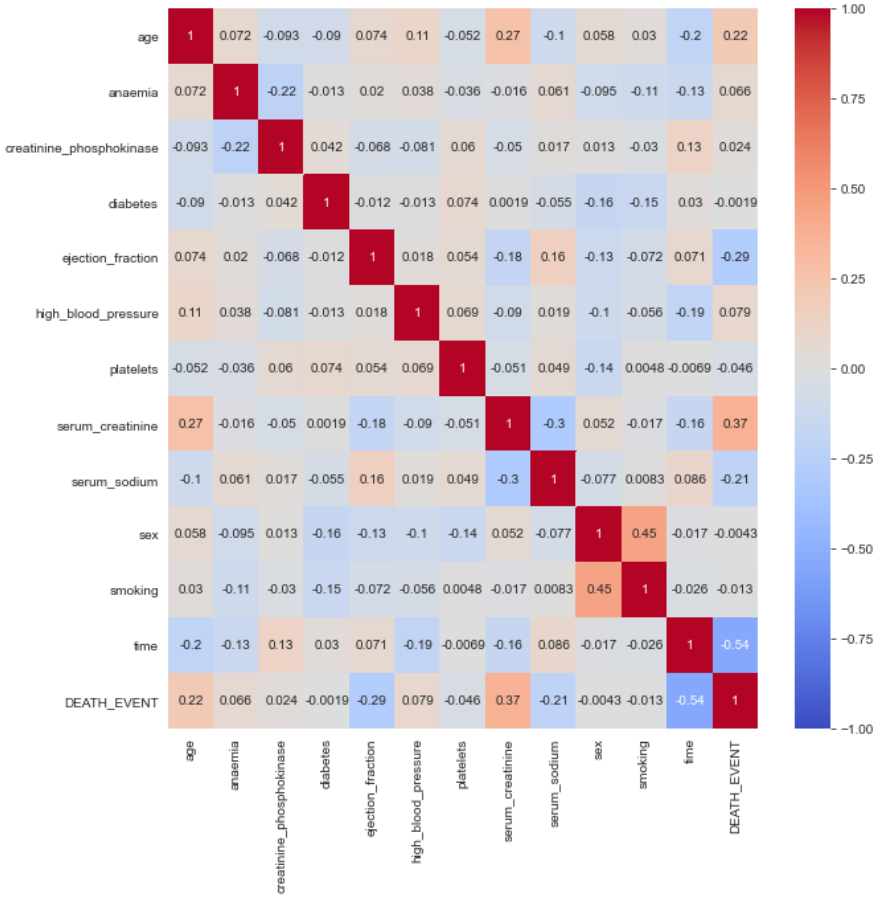


Figure 10 - Spearman correlation

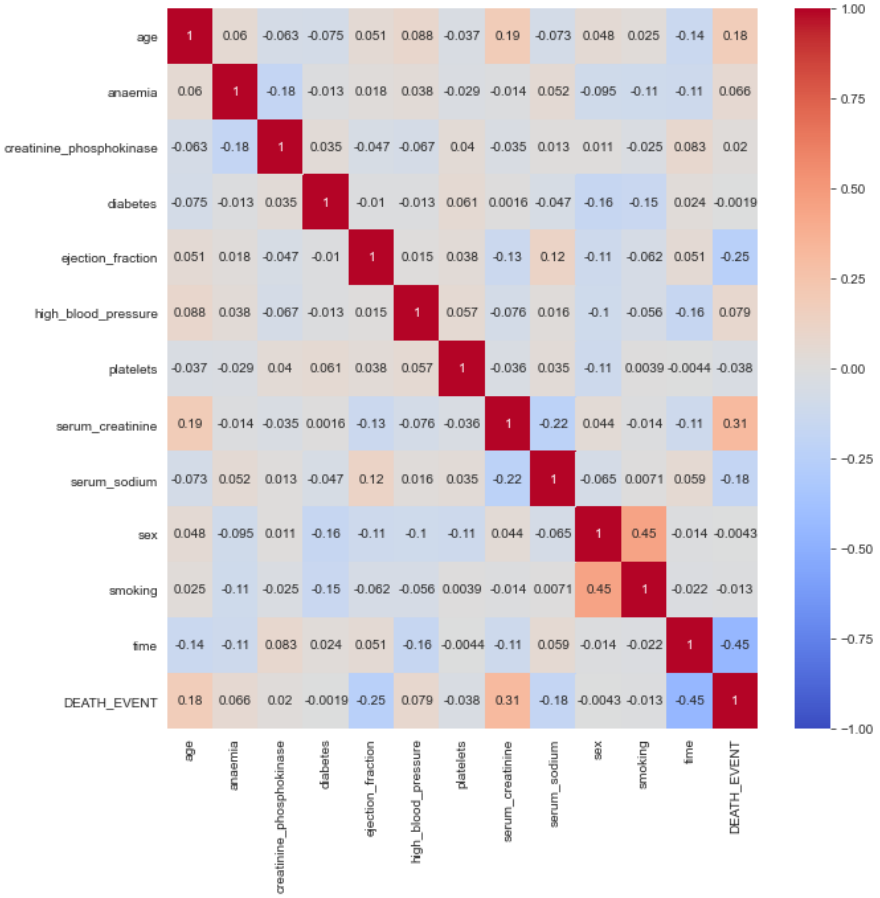


Figure 11 - Kendall Tau correlation

Checking the last row (or column because the matrices are symmetric) of the 3 matrices I will select only the features that are highly correlate with DEATH_EVENT.
I will choose only correlation greater than or equal to 0.25 (in absolute value).

From Pearson correlation: age (0.25), ejection_fraction (-0.27), serum_creatinine (0.29) and time (-0.53).

From Spearman correlation: ejection_fraction (-0.29), serum_creatinine (0.37) and time (-0.54).

From Kendall Tau correlation: ejection_fraction (-0.25), serum_creatinine (0.31) and time (-0.45).

Since the age feature has a correlation greater than 0.25 only in the first case, I will not consider it.

So, at this point the remaining features are ejection_fraction, serum_creatinine and time.
First of all I will look at ejection_fraction and serum_creatinine, and then I will talk about time.

To verify further the predictive power of serum creatinine and ejection fraction, I depicted a scatterplot with the serum creatinine on the x axis and the ejection fraction on the y axis, and I colored every patient-point base on survival status (survived or dead).

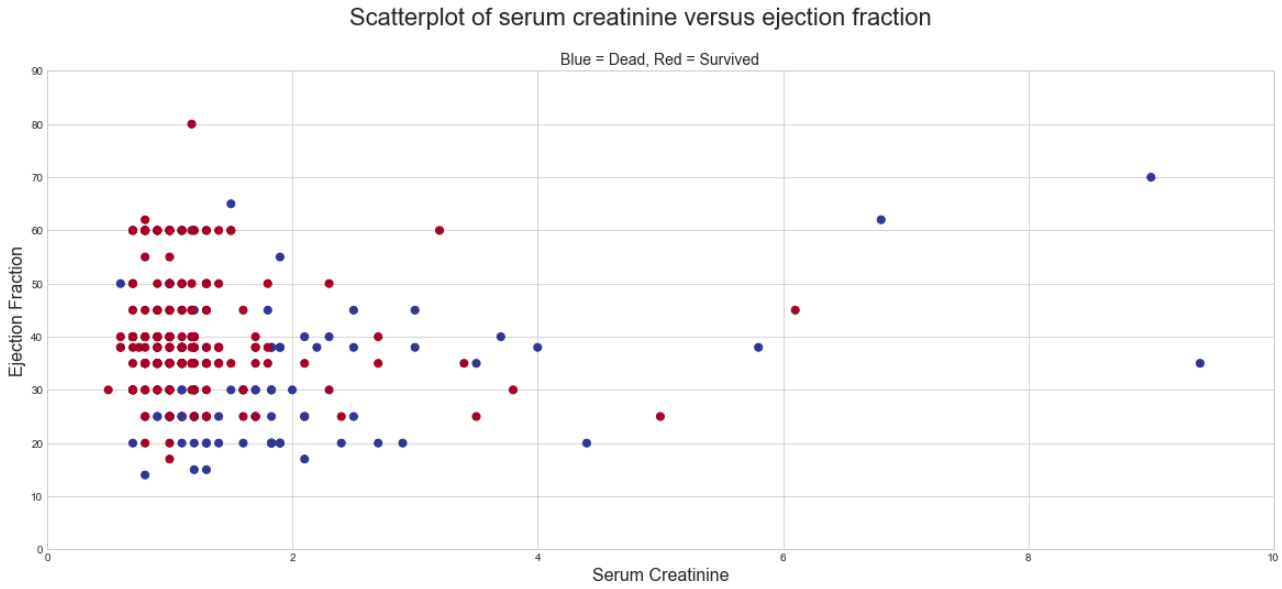


Figure 12 - Scatterplot of serum creatinine versus ejection fraction

This plot shows a clear distinction between alive patients and dead patients. It is possible to depict a straight line that separates the two variables.
Indeed, the majority of blue points (dead patients) is on the right; The death ratio increases with the level of serum creatinine, in particular for values greater than 1.5, and for low levels of ejection fraction.

Now, let's talk about time.
I will do 2 types of analysis. The first one in which I will exclude the variable time and the second one in which I will consider it.

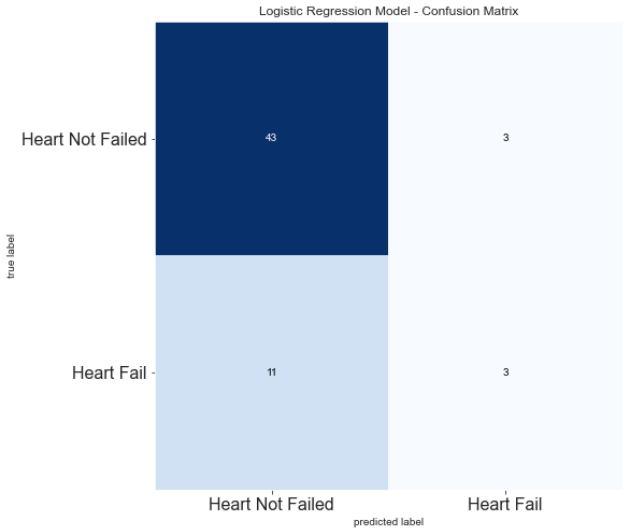
The time feature refers to the time from the start of the study after which the study was terminated. Either the people lost contact with the subject (presumably they were declared healthy and left) or they died due to heart failure. Time therefore has an artificially high predictive power and should not be used to predict, as it would not be known in time to make a prediction. Basically, if you would want to replicate the analysis on some real patient data, you would not have the information of time, because is the time that stops when a patient dies which you won't know before.
However, time can be an important factor in the survival of patients and should not be eliminated completely from this study. So, I will investigate the possible relationship between time and the survival of patients: is the time related to the chance of survival of the patient?

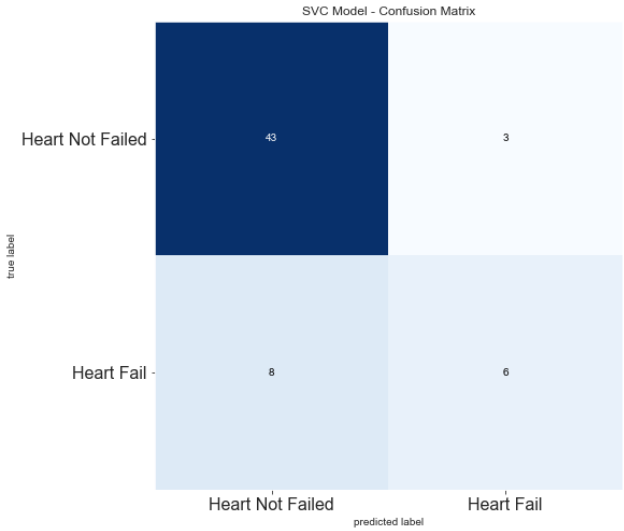
To predict patient survival, I employed eight different methods from different machine learning areas. The classifiers include one Logistic Regression, three tree-based methods (Random Forests, Decision Trees and Extra Trees) one Support Vector Machine (with Radial basis function kernel), one instance-based learning model (k-Nearest Neighbors) and two probabilistic classifiers (Gaussian and Bernoulli Naïve Bayes).

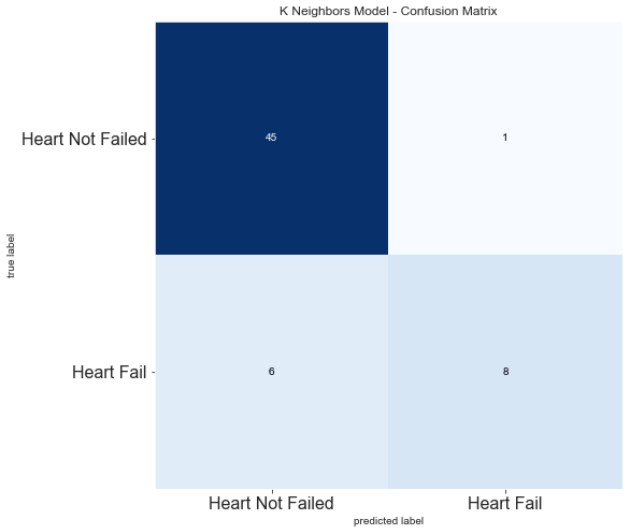
I split the dataset into 80% for training set (239 randomly selected patients) and 20% (the remaining 60 patients) for test set.

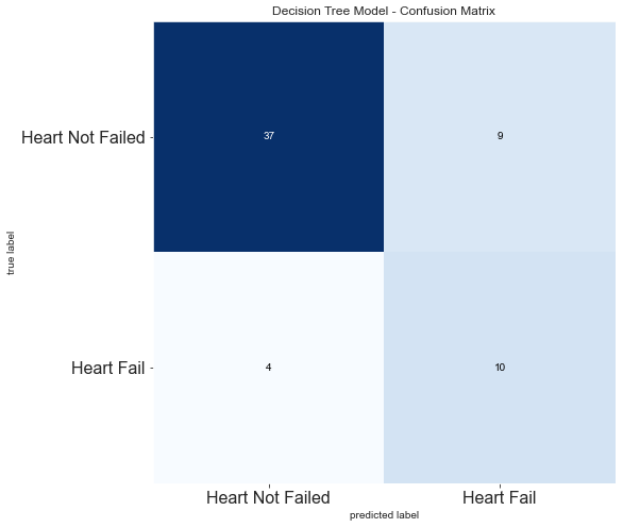
I measured the prediction results through confusion matrices.

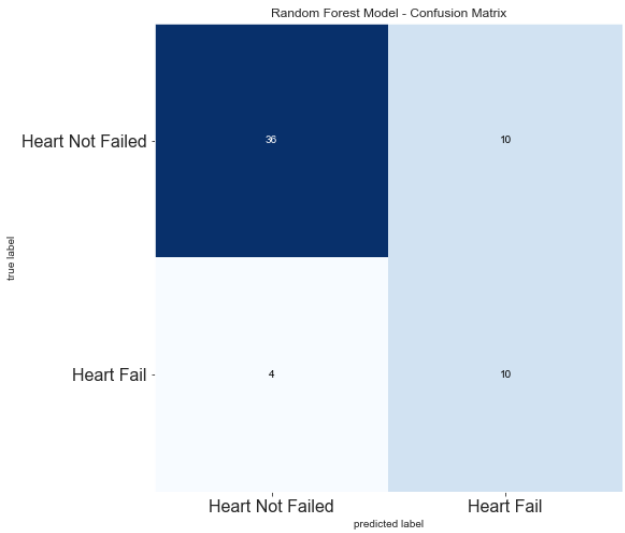
Now I will present the results obtained in the first case of the analysis, using only 2 features, ejection_fraction and serum_creatinine.

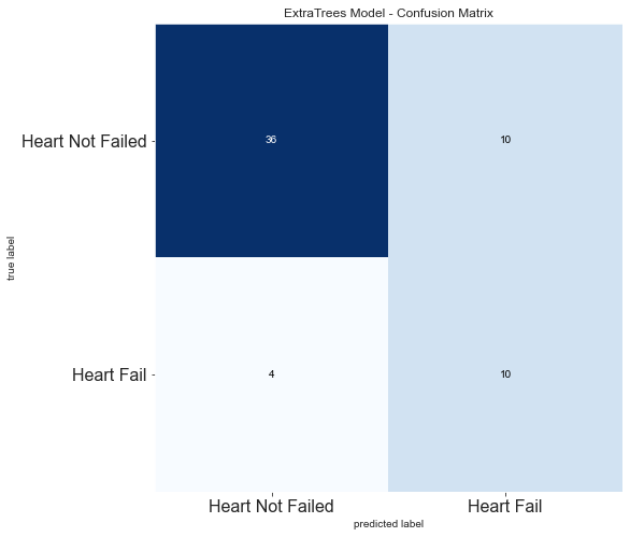












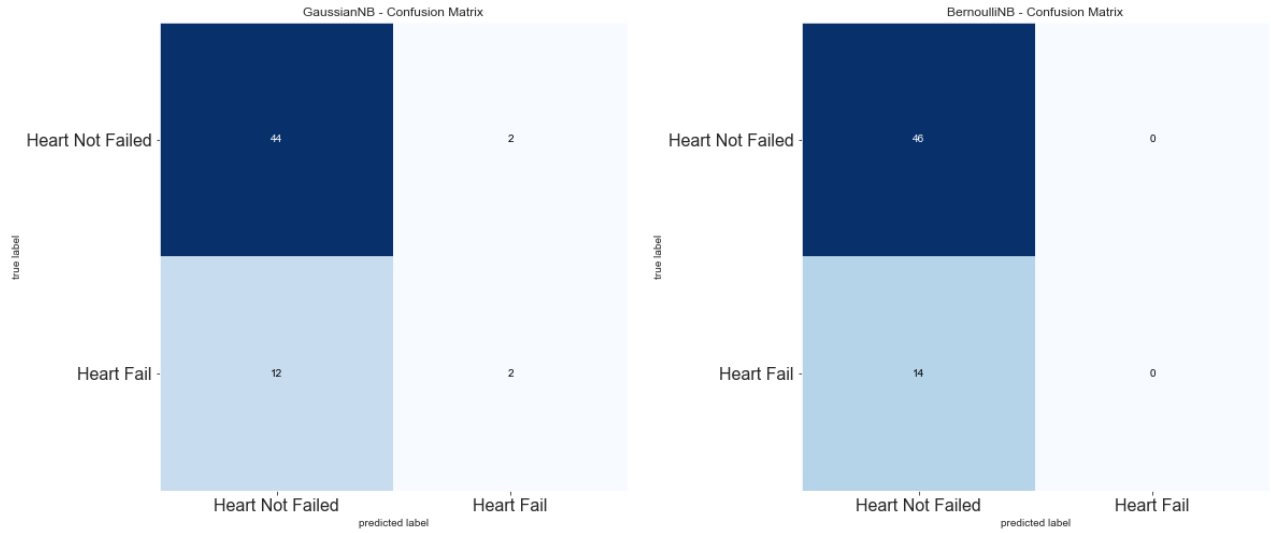


Table 6 – Confusion matrices of the first model

We know that the dataset is unbalanced because 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. For this reason, all the methods obtained better prediction scores on the true negative rate, rather than on the true positive rate. These results occur because the algorithm can see more negative elements during training, and therefore they are more trained to recognize deceased patient profiles during testing.

Summary of the results:

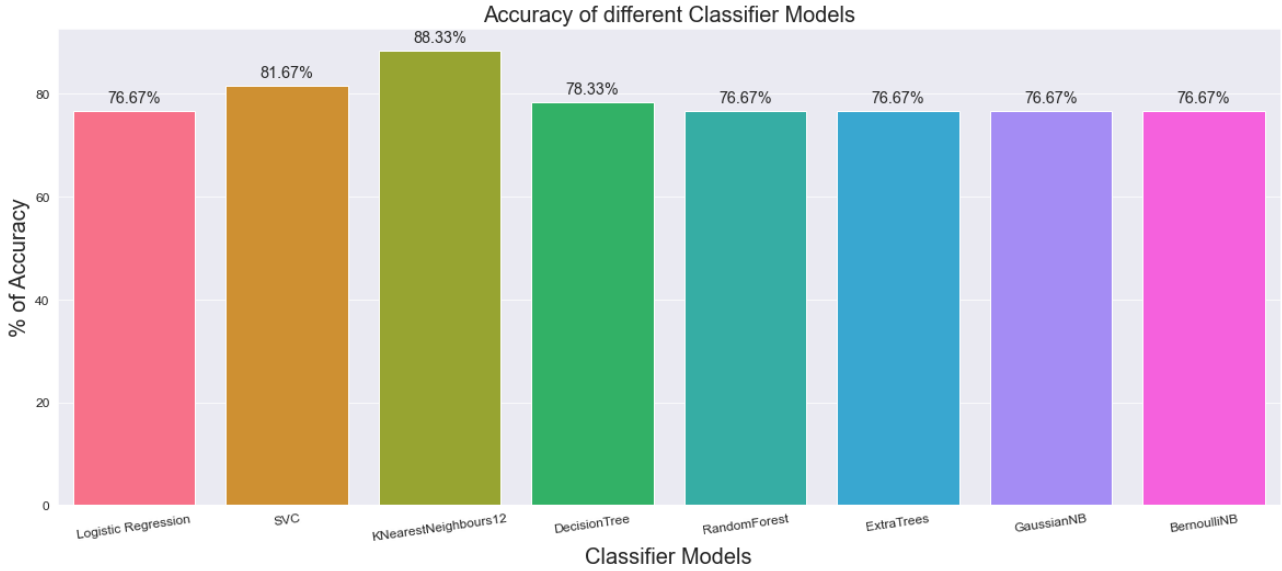


Figure 13 - Summary of the 1st case: only 2 features

These results are showing that k-Nearest Neighbors outperformed all the other methods, by obtaining the top accuracy, 88.33%.
The three tree-based methods, Random Forests, Decision Trees and ExtraTrees obtained the top results on true positives, predicting correctly the majority of deceased patients.

I will not present the confusion matrices of the other analysis; I will show only the summaries.

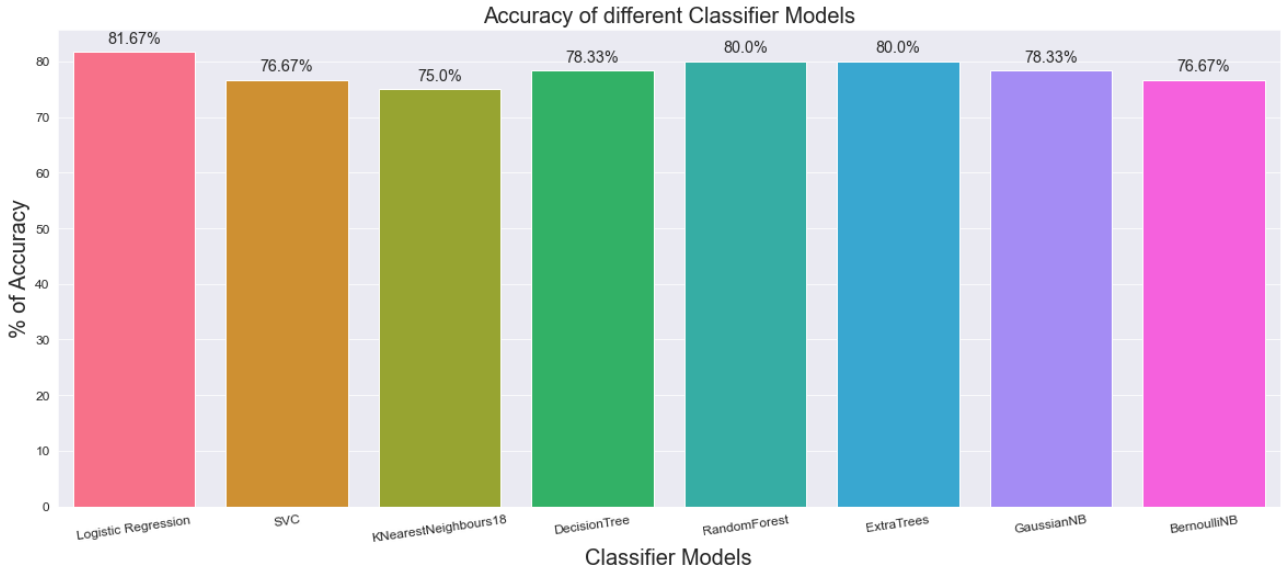


Figure 14 – Summary of the 2nd case: all the features (no time)

In the picture above it is possible to see that the best performance using all the features (no time) is achieved by the Logistic Regression with an accuracy of 81.67%. However, in the previous case, using only ejection_fraction and serum_creatinine, the performance of k-Nearest Neighbors was better, with 88.33% of accuracy. This means that the prediction made on these two features alone can be more accurate than the predictions made on the complete dataset. This aspect is particularly encouraging for the hospital settings: in case many laboratory tests results, and clinical features were missing from the electronic health record of a patient, doctors could still be able to predict patient survival by just analyzing the ejection_fraction and serum_creatinine values.

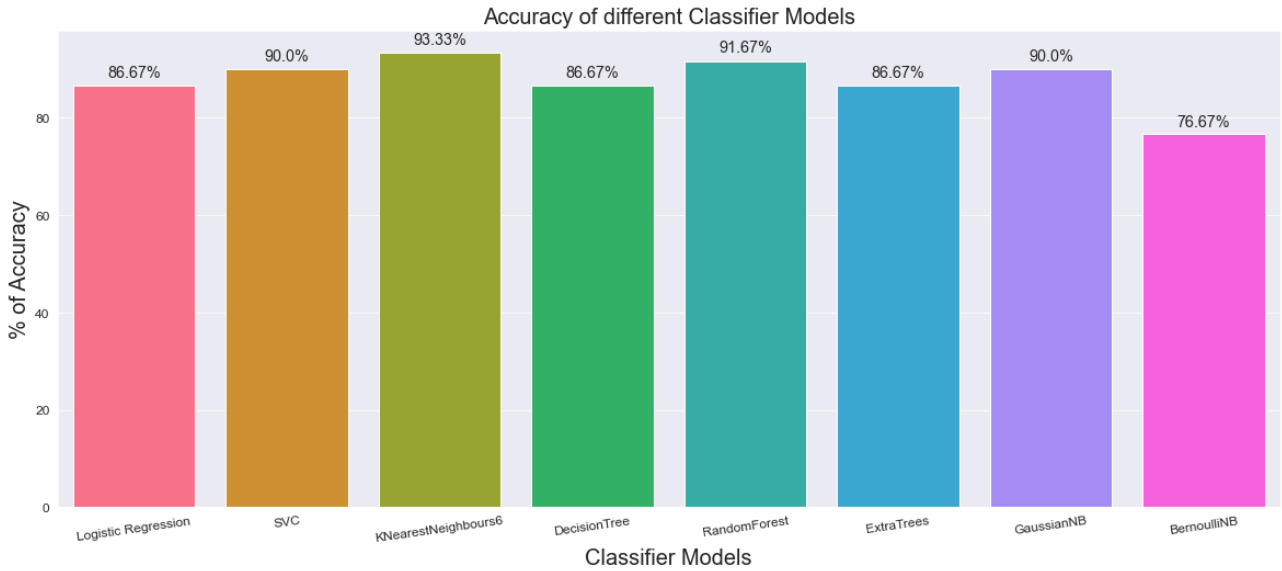


Figure 15 - Summary of the 3rd case: only 2 features and time

In this case of only 2 features and time the results were better of the previous cases that did not use time. K-Nearest Neighbors obtained an accuracy of 93.33%. The second-best accuracy was obtained from Random Forest that is the model with the highest true positives rate. Indeed, this model found 13 TP.

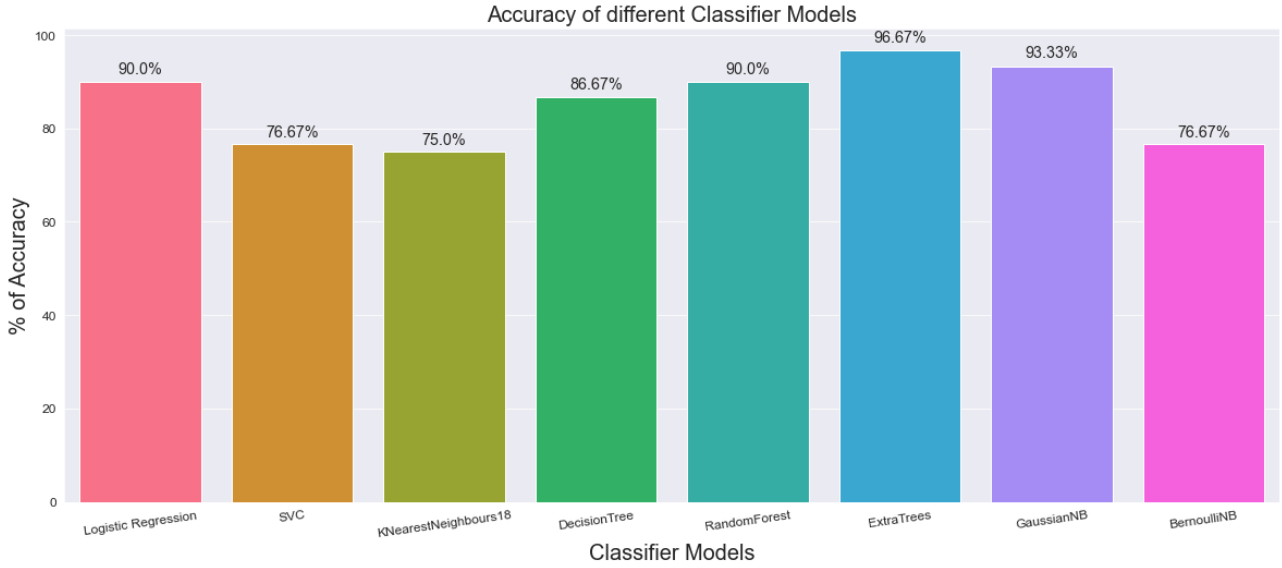


Figure 16 - Summary of the 4th case: all the features and time

Finally, in the last picture there are the results of the models using all the features including time. ExtraTrees outperformed all the other methods of this and also of the previous cases, with 96.67% accuracy.

However, as I said before, in a real situation, we cannot have any value for time, because we won't know before.

COMPARISON WITH A RELEVANT PAPER

A relevant paper about this dataset was written by Davide Chicco and Giuseppe Jurman, with title "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone".

In their work, for methods that needed hyper-parameter optimization (neural network, Support Vector Machine, and *k*-Nearest Neighbors), the authors split the dataset into 60% (179 randomly selected patients) for the training set, 20% (60 randomly selected patients) for the validation set, and 20% (the remaining 60 patients) for the test set. For the other methods (Random Forests, One Rule, Linear Regression, Naïve Bayes, and Decision Tree), instead, they split the dataset into 80% (239 randomly selected patients) for the training set, and 20% (the remaining 60 patients) for the test set.
They applied each method 100 times and reported the mean result score.
The authors also did different analysis, based on using only 2 features (ejection_fraction and serum_creatinine) or all the features, and also depending on the feature of time.

I am going to report the values of accuracy only for the classifiers we both used.

	My values	Paper
Random Forests	80.0%	74.0%
Decision Trees	78.33%	73.7%
GaussianNB	78.33%	69.6%
SVC	76.67%	69.0%
k-Nearest Neighbors	75.00%	62.4%

Table 7 – Comparison accuracies of models with all the features (no time)

It is clear that I obtained better results, in particular in the case of kNN with a difference of 12.6%. However, in the paper It is not reported how many neighbors in the kNN algorithm the authors used, so probably they did not use the best configuration. Instead in my analysis, I wrote only the accuracy in the best case, because I plotted the results of several trials, from 1 neighbor to 25 neighbors. For example, in the case of this model, the best accuracy was with 18 neighbors.

Moreover, as I said before, the authors used 60% for training set, 20% for validation set and 20% for test set. However, this dataset contains only 299 observations and I think that using 20% for validation test cause the training set to have only few observations.

Then, for the case in which the analysis is made only by using serum_creatinine and ejection_fraction, the authors used 2 different machine learning algorithms: Random Forests, Gradient boosting and SVM radial. Also, here they did 100 execution and reported the mean value.

	My values	Paper
Random Forests	76.67%	58.5%
SVC	81.67%	54.3%

Table 8 – Comparison accuracies of models with only EF and SC (no time)

Finally, including time:

	My values	Paper
Logistic Regression (EF, SC and time)	86.67%	83.8%
Logistic Regression (all features and time)	90.0%	83.3%

Table 9 – Comparison accuracies of models

CLUSTER ANALYSIS

We can ask ourselves if the structure of the original dataset is made in such a way that allow us to say that are two groups of patients, dead and survived. In other words, we can ask ourselves if the patients are grouped in 2 clusters, one cluster for dead patients and one cluster for survived patients.
I computer the kMeans Algorithm iteratively, changing the number of clusters. I started analyzing from 2 clusters to 25 clusters.
I used all the features. I wanted to see how the separateness of my clusters behave according to some metric, in this case I used the Silhouette Coefficient.
Before of using kMeans I scaled the data, since kMeans is a distance-based algorithm.
I plotted the Silhouette Coefficient for the different number of clusters.

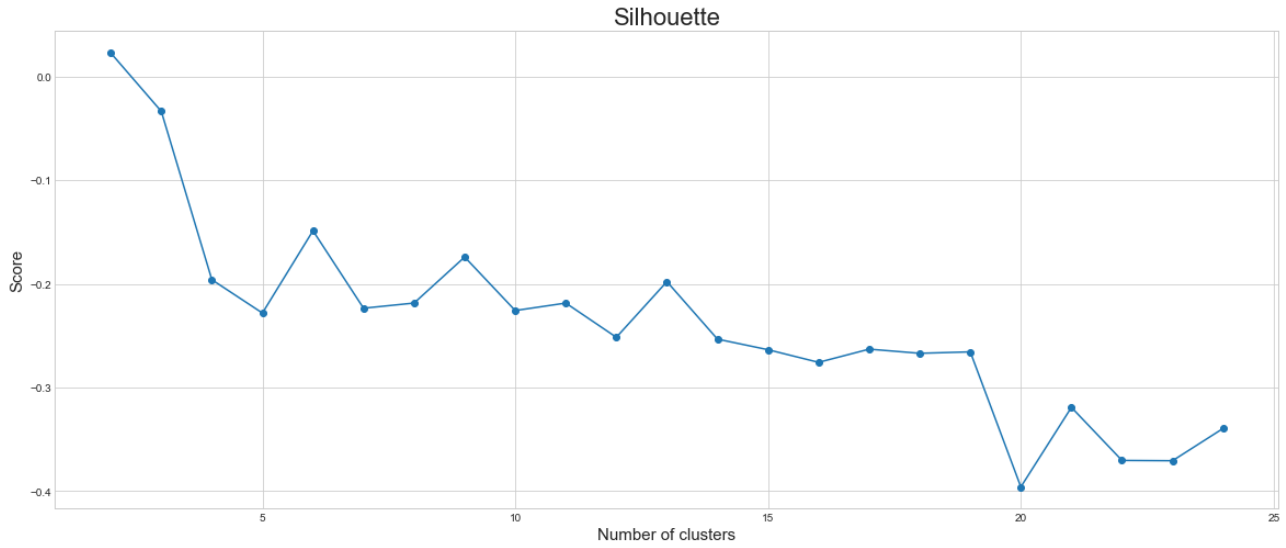


Figure 17 - Silhouette Coefficient of all features

As we can see, the best value is in correspondence of 2 clusters, with a Silhouette Coefficient a little bit greater than 0. From the Scikit-Learn Guide: "The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar."

So, this best value is not good.
I decided to try using only the features of ejection fraction and serum creatinine:

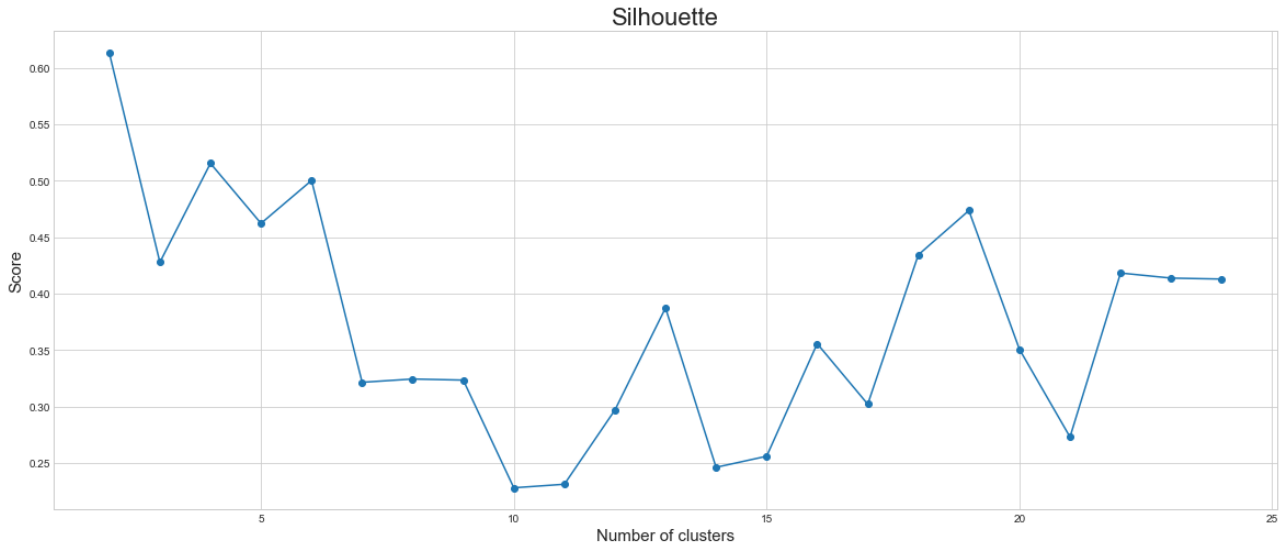


Figure 18 - Silhouette Coefficient with only 2 features

It is easy to note that the results in this case are better. The best case is again in correspondence of 2 clusters, with a good Silhouette Coefficient, a value greater than 0.60.

For purposes of visualization, I used a decomposition technique.

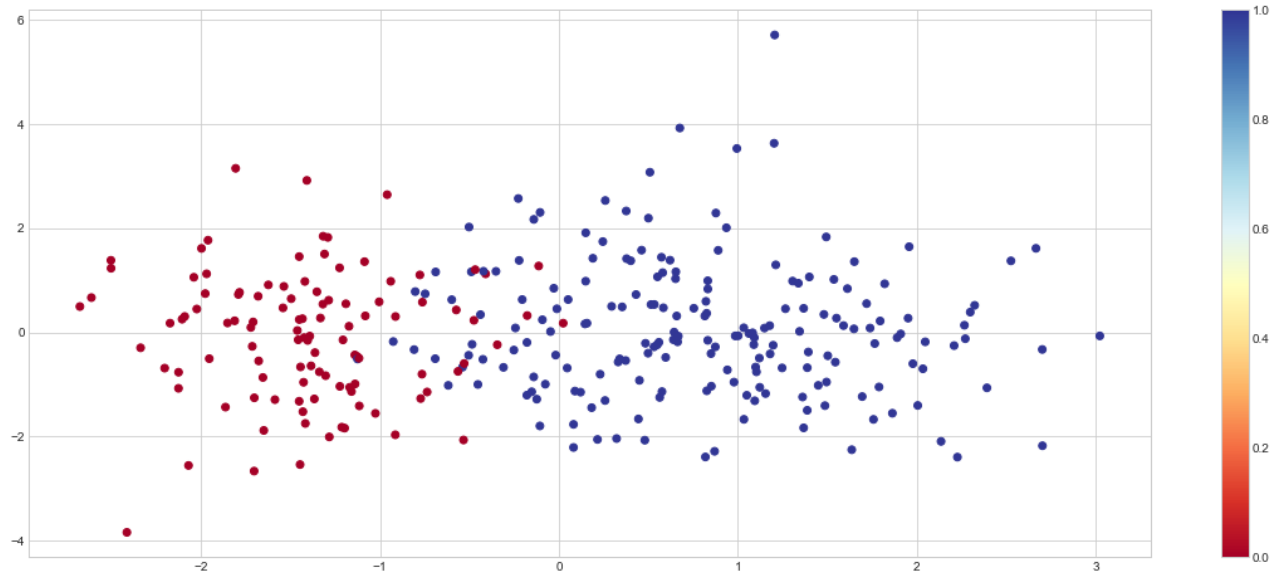


Figure 19 – First and second principal components (all features) with 2 clsuters

This plot shows first and second principal component (it is simply a projection of how all the features look in 2D space).
Colors are the clusters. For every color we can see in which cluster the algorithm (k-means) assigned each observation to.
This first plot was obtained using all the features. There are only 2 colors because I plotted the case with only 2 clusters. For example, I can plot also the case of 25 clusters to show the confusion with lot of clusters.

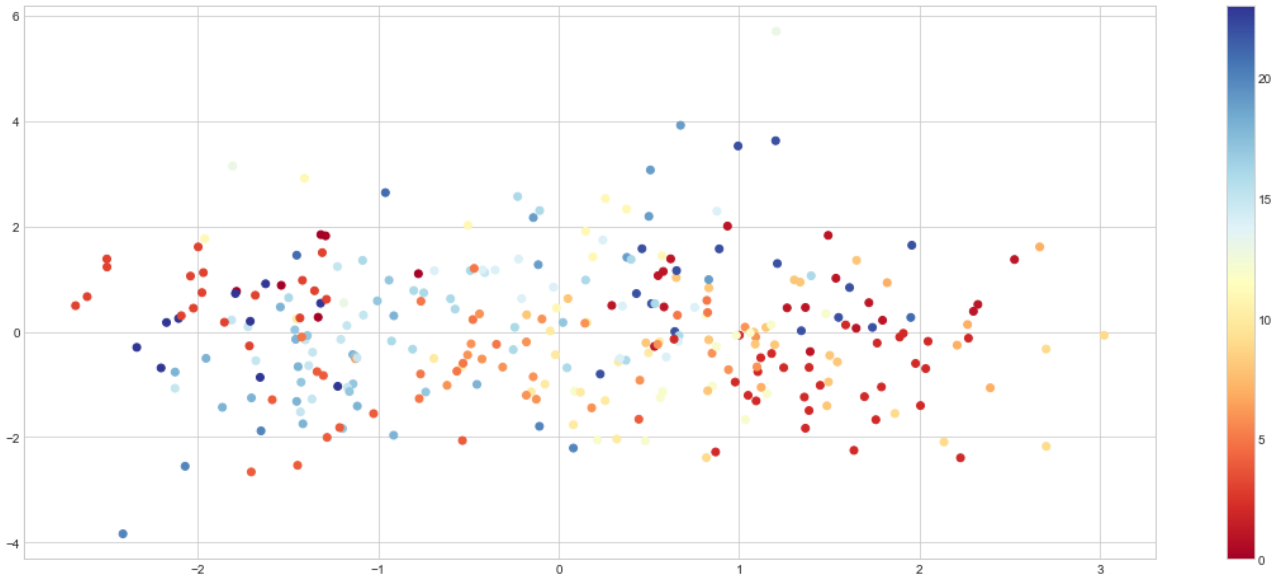


Figure 20 - First and second principal components (all features) with 25 clusters

And now I show the plot using only ejection fraction and serum creatinine, with only 2 clusters.

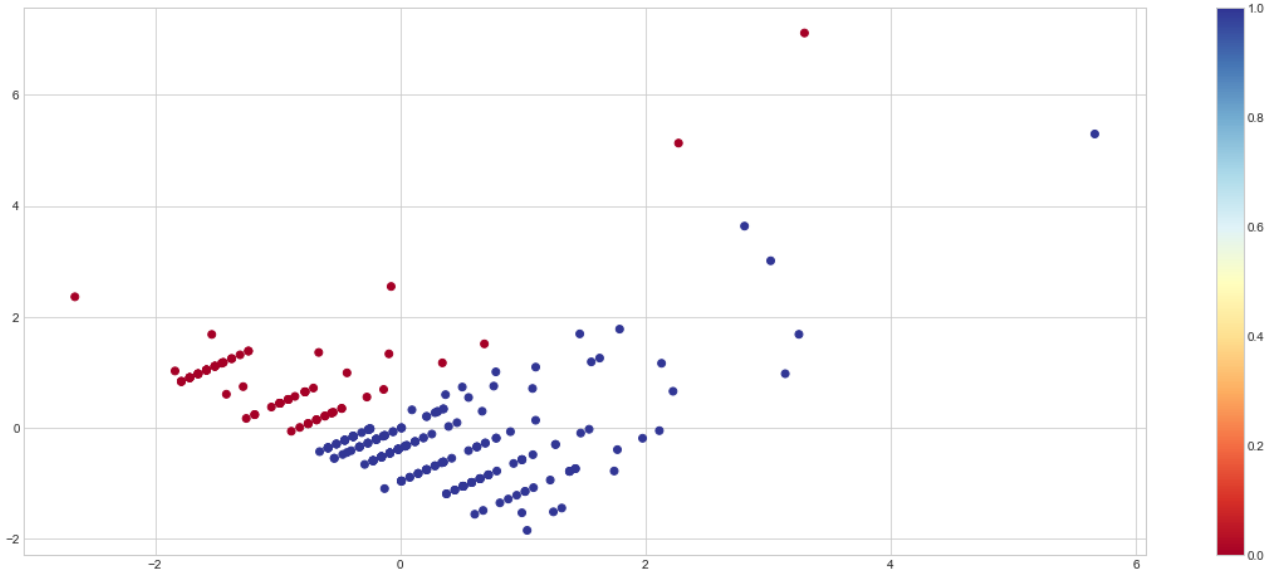


Figure 21 - First and second principal components (only 2 features) with 2 clusters

Last idea that I applied in my analysis: from the Silhouette Coefficient and PCA I saw the clustering tendency of my data and I watched graphically, in 2D, these clusters. However, until now, I cannot say if these 2 clusters separate dead patients from survived patients

To try to understand this I constructed a confusion matrix for comparing the data of the column DEATH_EVENT and the data predicted by the clustering algorithm and check if one of the clusters contains majority of death or not death cases.

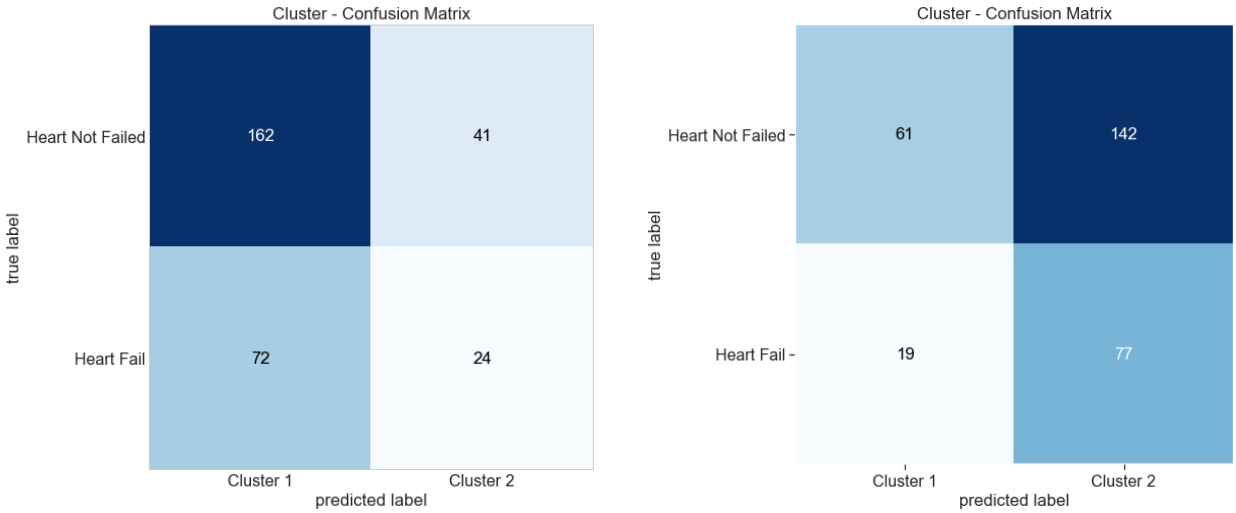


Table 10 – Left: CM of the algorithm using all the features. Right: CM of the algorithm using only 2 features

In the CM on the left, a lot of survived patients belongs to cluster 1 with 162 of 203 total, so the specificity, that measures how exact the assignment to the positive class is, in this case is 79.8%. These 162 patients are the majority, indeed in the same cluster there are other 72 patients that had a heart failure. So, in this cluster there are 234 observations and 162 (69.23%) are survived patients.
In both the matrices the precision, that measures how good the model is at assigning positive events to the positive class, (TP/(TP + FP)) is low; indeed, in the first matrix precision = 24/(24+41)=36.92% and in the second matrix it is 77/(77+124) = 38.31%.