Predicting death events from heart failures

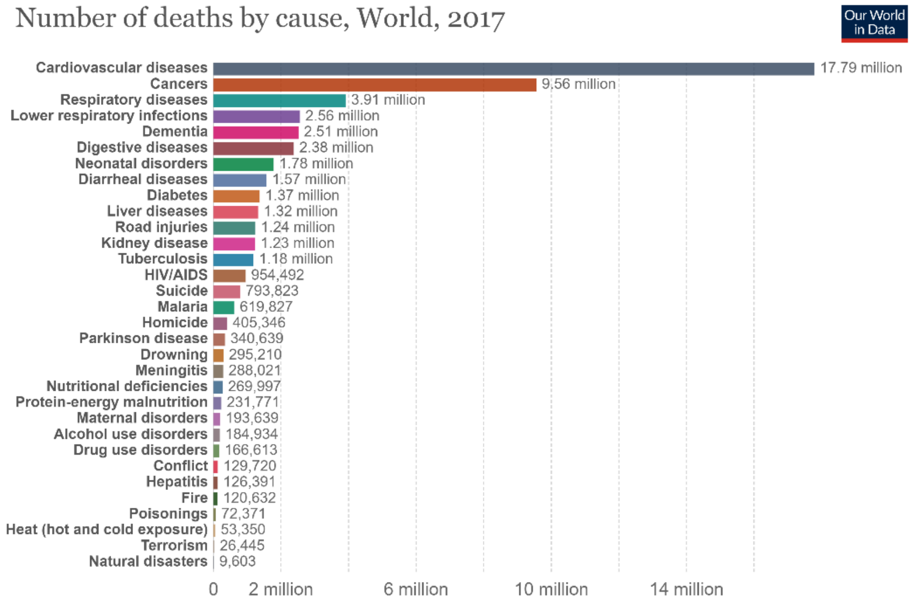
Brandon Yilmaz Castillo

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.



Source: IHME, Global Burden of Disease

OurWorldInData.org/causes-of-death • CC BY

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.

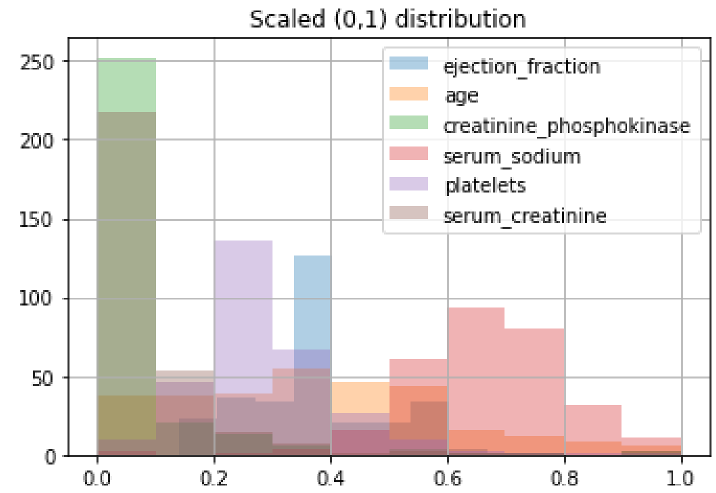
The main problems when data cleansing techniques are applied is the different measurement of the features. This can be a problem when Classification Algorithms are applied. Dummy variables are already ‘scaled’ in a sense, but continuous variables as platelets in the blood or serum creatinine are not.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

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

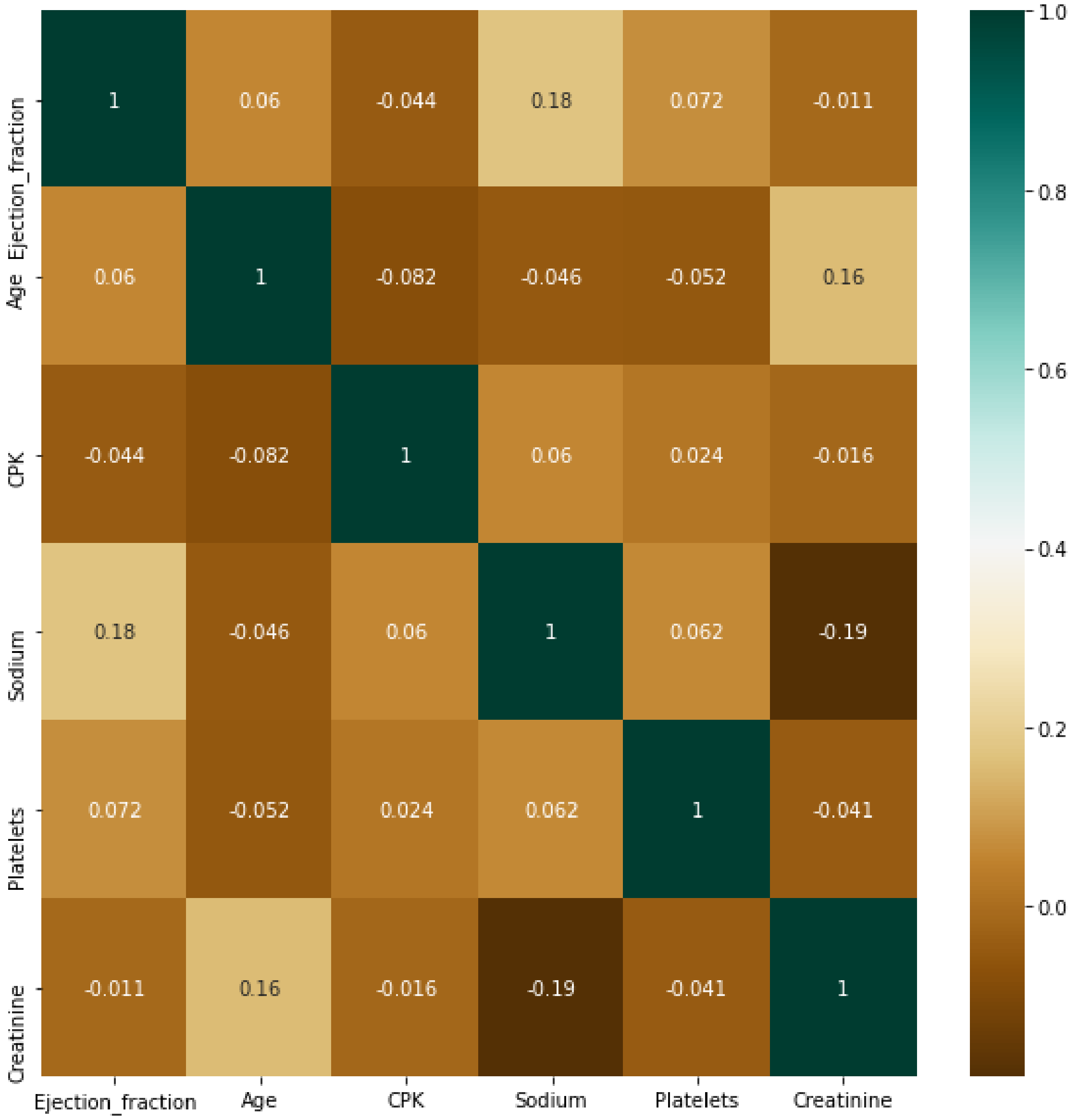
The Scaling formula applied was:

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

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 among the continuous variables, so there is no Multicollinearity between features.

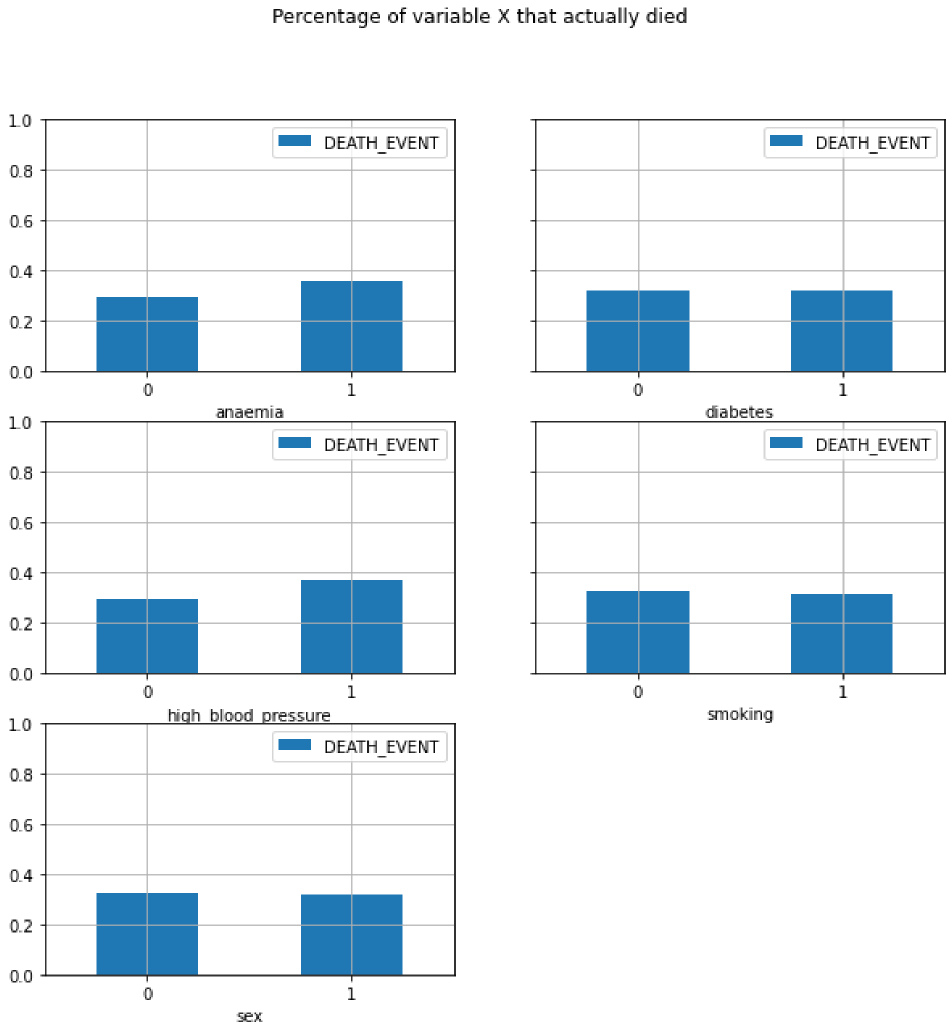


3.Exploratory Data Analysis

3.1Relationship between target variable and categorical variables

The categorical variables add up to 4 (excluding the target variable and gender). These categorical variables describe if the patient was a smoker, if he or she developed anaemia, the presence of high blood pressure (hypertension), and if the patient had diabetes. As we can state below, the **plot bar shows the percentage (mean) of death events for each value of the dummy.** Hence, it's a good way to visualize if there is a huge impact of the target variable conditioning by each dummy-binary value.

The first exploratory stage showed no big differences. One may argue that smoking increases heart failure, but doesn't mean that it causes eventually major risk to die after getting a heart failure. Same applies to diabetes. Nevertheless, the variable Anaemia and high blood pressure (Hypertension) could be good features.



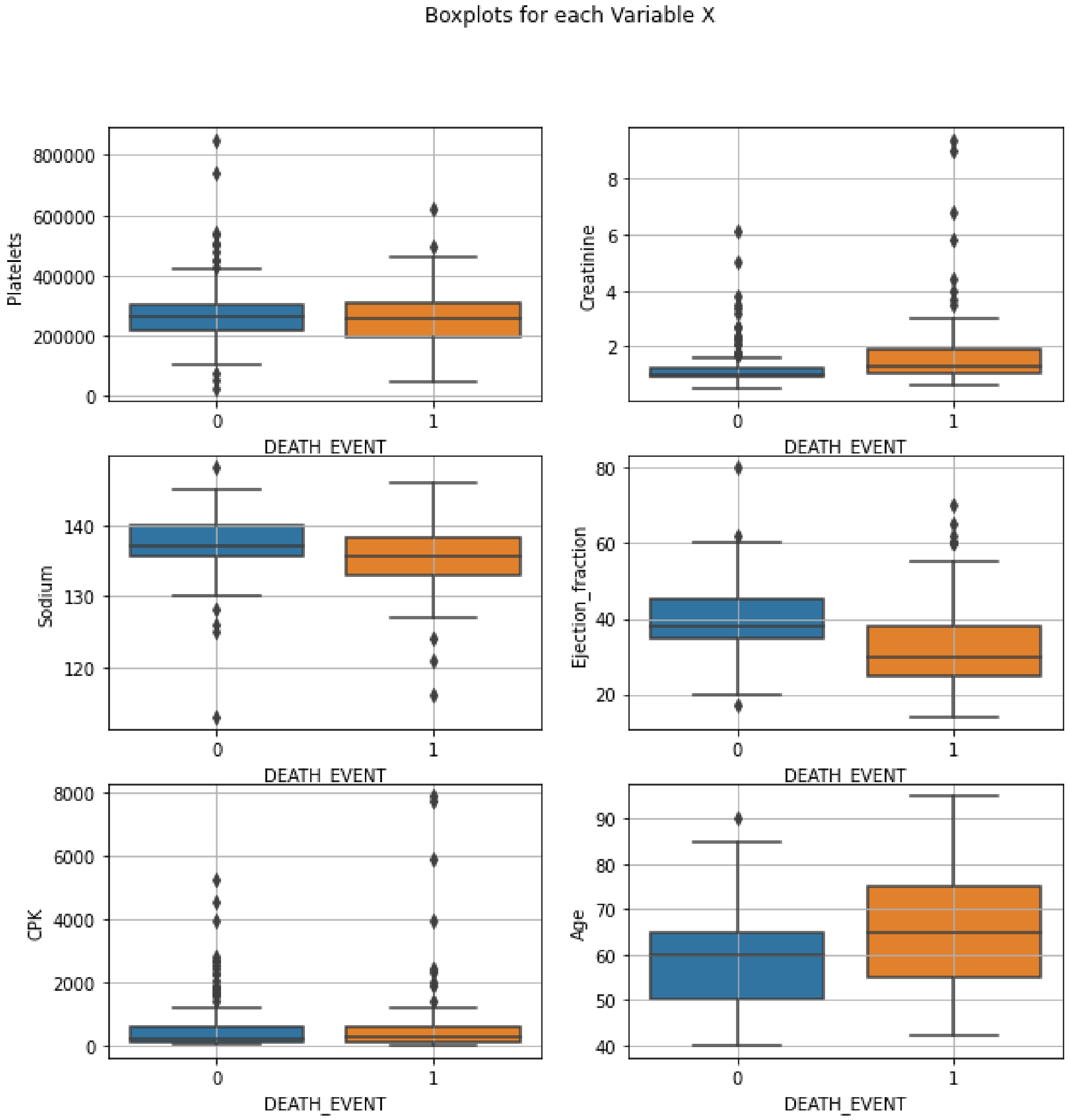
Additionally, ANOVA tests were carried out in order to analyse if the differences between the means were significant.

Variable	P-Value
Anaemia	0.25
Diabetes	0.97
Blood pressure (Hypertension)	0.17
Smoking	0.82
Sex	0.94

As we see, no one is statistically significant at 10%. But maybe this is caused by the small dataset, which is important in ANOVA tests. As we said, Anaemia and Blood pressure (lowest p-values) are kept as possible features.

3.2Relationship between target variable and continuous variables

The continuous variables are more related to technical issues. In this sense, they provide more information about the status of the patient’s heart and physical condition. Boxplots were drawn in order to analyse if there are significant differences between the variable ‘DEATH_EVENT’ and the continuous features.



Variables as 'CPK' and 'Platelets' show same distribution with patients who died and patients who not. On the other hand, variables as 'Age', 'Ejection fraction', 'Creatinine' and 'Sodium' show more significant differences in their distributions.

3.3Proposed Machine Learning models

As the target variable is a dummy (binary), a Classification Algorithm is the most suitable choice. The Algorithms proposed are: Logistic Regression, K-nearest neighbour, and Supported Vector Machine (SVM).

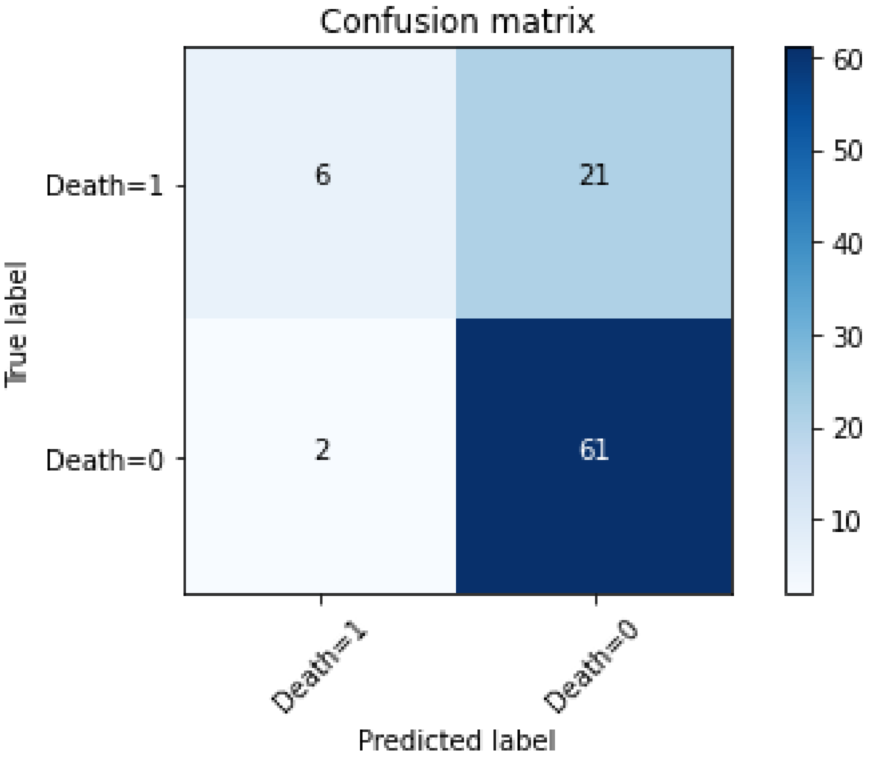
4.Modeling and Results

Before entering the sub-sections related to each model that was used for the analysis, the features that were finally used as predictors were: Ejection fraction, Creatinine, Age and Blood pressure (Hypertension). Just one categorical variable was used (the one with least p-value) in order to avoid insignificant variables.

As common in Machine Learning model evaluation, a split was made between the dataset in order to train and test the model with separate datapoints. The percentage of test-set is 30 %.

4.1Logistic Regression

After fitting the model with the train set, a confusion Matrix of the test set was built and a classification report. The model seems to predict pretty well the category Death=0, but pretty bad the category Death=1. This fact is visible in the recall metric and has consequences over the F1-score metric. Finally, the Jaccard-index shows a 0.74 and the LogLoss a 0.54 metric; the former looks standard, the latter shows very bad capability in predicting.



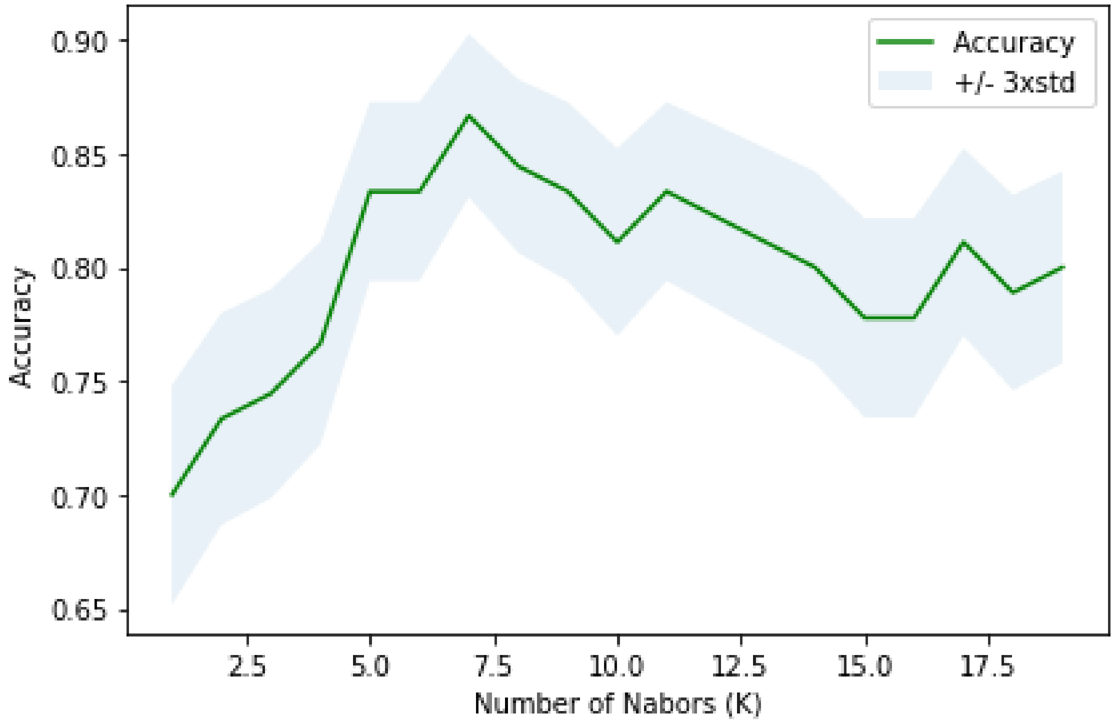
Variable	Precision	Recall	F1-score	Support
0	0.74	0.97	0.84	63
1	0.75	0.22	0.43	27
Accuracy			0.74	90
Macro avg	0.75	0.60	0.59	90
Weighted avg	0.75	0.74	0.69	90

In order to avoid overfitting a cross validation method was applied and consequently calculated the mean of the resulting scores:

LogLoss model above	Jaccard-index model above	Mean of Cross Validation (cv=10)
0.53	0.74	0.72

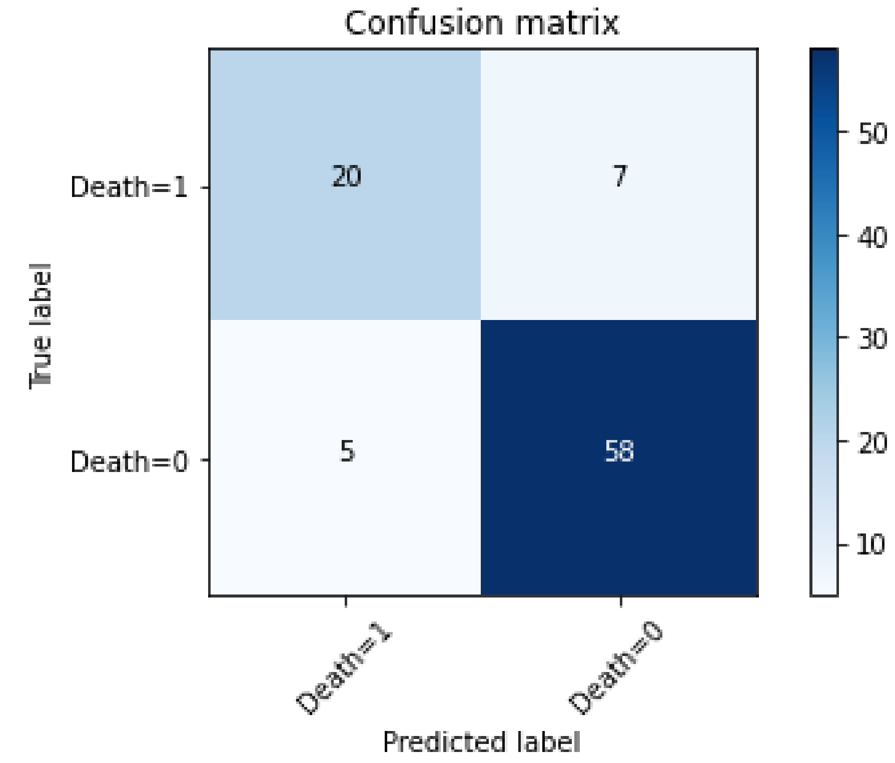
4.2K-nearest neighbor

For the K-Nearest neighbour algorithm the k parameter was estimated developing several models with different K's and choosing the one with the highest Accuracy. Obviously the index was calculated with the test-set.



The best K parameter was K=7.

The Confusion matrix of this Machine Learning model showed much better results than the results showed by the Logistic regression model.



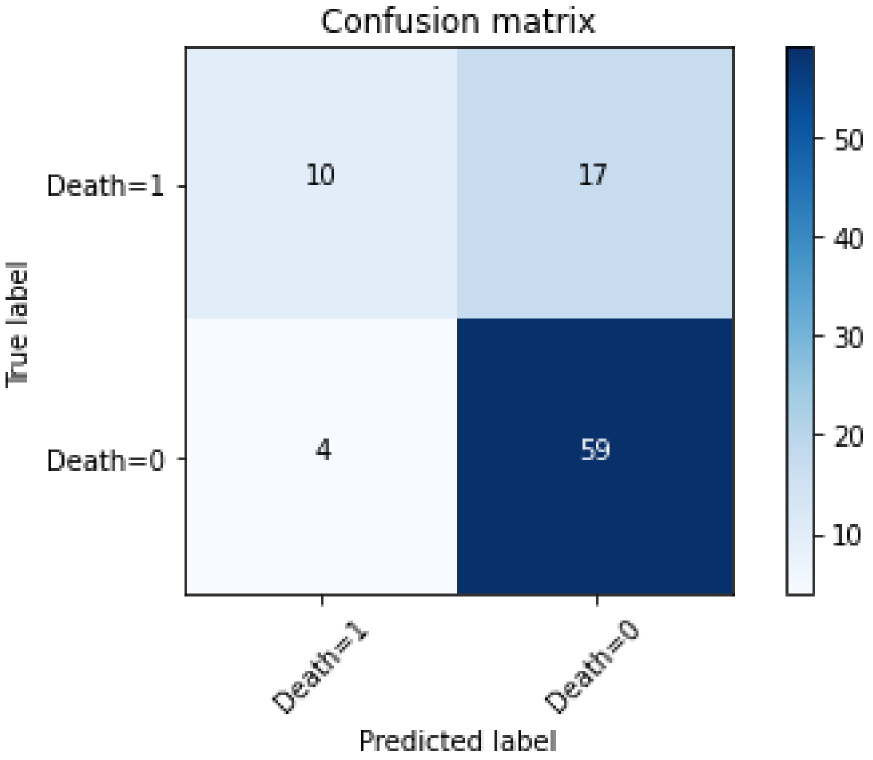
Variable	Precision	Recall	F1-score	Support
0	0.89	0.92	0.91	63
1	0.8	0.74	0.77	27
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Accuracy			0.87	90
Macro avg	0.85	0.83	0.84	90
Weighted avg	0.86	0.87	0.87	90

In order to avoid overfitting a cross validation method was applied and consequently calculated the mean of the resulting scores:

Jaccard-index model above	Mean of Cross Validation (cv=10)
0.87	0.77

4.3Support Vector Machine (SVM)

The Support Vector Machine was constructed with a RBF kernel. The model shows bad results at predicting values with label equal to 1, but extraordinary good results at predicting values with label equal to 0. The SVM model doesn't fit very well with small size datasets and surely this was one of the factors that made this model useless for this purpose.



Variable	Precision	Recall	F1-score	Support
0	0.78	0.94	0.83	63
1	0.71	0.37	0.49	27
Accuracy			0.77	90
Macro avg	0.75	0.65	0.67	90
Weighted avg	0.76	0.77	0.74	90

In order to avoid overfitting a cross validation method was applied and consequently calculated the mean of the resulting scores:

Jaccard-index model above	Mean of Cross Validation (cv=10)
0.77	0.73

5.Discussion

The models displayed above (Logistic Regression,KNN,SVM) are significantly different among them. One of the key indicators that shows us this variability is the Confussion Matrix: the Logistic Regression Model and the SVM model predict very poorly the label=1 (DEATH_EVENT); in contrast KNN model predicts pretty well this label. Of course there seems to be a trade-off between the two predictions (label=0,label=1), for example, SVM model has a recall on label=0 equal to 1, but 0.04 on label=1. This disparity isn’t optimum.

To sum up, the best model which is perfectly equilibrated is the KNN model. The KNN model shows pretty good metrics and balanced if we compare it to the SVM model case.

6.Conclusion

There are severe improvements that could be made in order to achieve better results on this topic. One of these improvements is getting a better Database, as we have seen having small-size datasets makes some Algorithms useless (see the case of the SVM model). Therefore, it’s highly recommendable to perform a similar analysis but with a much larger dataset.

Additionally, the KNN model has demonstrated to be very useful with a small dataset in which the variables are clearly well defined and add signal (not noise). Distance algorithms need variables that are significantly different in order to achieve good metrics.