Predict Survival of patients with heart failure

Ravi Mugesh[UXDJ7L]

Eötvös Loránd University, H-1053 Budapest, Egyetem tér 1-3, Hungary
uxdj7l@inf.elte.hu

Abstract. Cardiovascular diseases 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. Machine learning can predict patients survival from their data and can individuate the most important features among those included in their medical records. In this report, we analyze a dataset of 299 patients. We apply Logistic Regression to predict the patients survival. This discovery has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not. Indeed, medical doctors aiming at understanding if a patient will survive after heart failure may focus mainly on serum creatinine and ejection fraction.

Keywords: Heart failure · Cardiovascular heart diseases · Serum creatinine · Ejection fraction · Machine learning · Data mining.

1 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 and this dataset contains 12 features that can 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.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. Scientists can take advantage of machine learning not only for clinical prediction, but also for feature ranking.

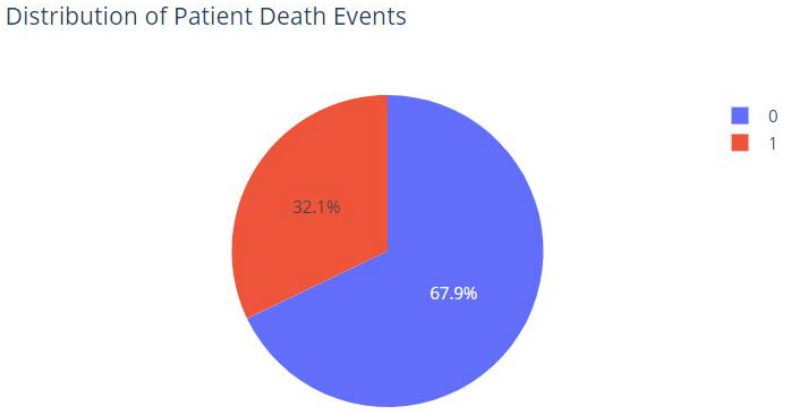
2 Methods and Experiment

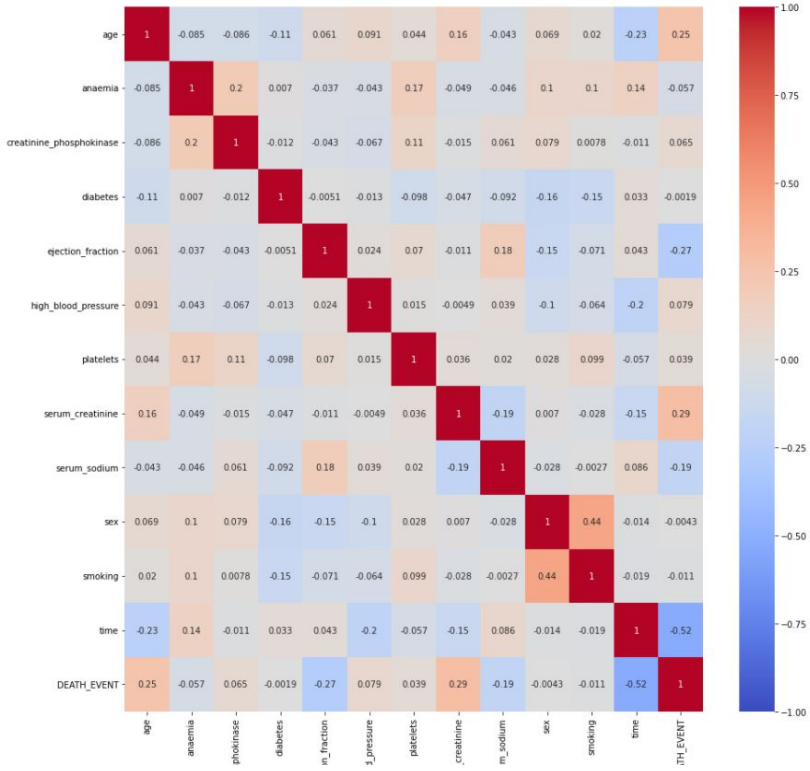
2.1 Dataset

We analyzed a dataset containing the medical records of 299 heart failure patients. The patients consisted of 194 men and 105 women, and their ages range between 40 and 95 years old. The dataset contains 13 features. Some features are binary such as anaemia, high blood pressure, diabetes, sex, and smoking. Regarding the features, the creatinine phosphokinase 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 follow-up period.

Table 1. Explanations, measurement units, and range of each feature of the dataset

Feature	Information	Measurement	Interval
Age	Age of the person	Years	[40,...,95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
Creatinine Phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[0,..,7861]
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection Fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14,...,80]
High blood pressure	If the patient has hypertension	Boolean	0, 1
Platelets	Platelets in the blood	kiloplatelets /mL	[0,..,232461]
Serum Creatinine	Level of serum creatinine in the blood	mg/dL	[0.50,..,9.40]
Serum Sodium	Level of serum sodium in the blood	mEg/L	[113,..,148]
Sex	Woman or man	Years	0, 1
Smoking	If the patient smokes or not	Boolean	0, 1
Time	Follow-up period	Days	[4,..,285]
Death Event	If the patient deceased during the follow-up period	Boolean	0, 1


Distribution of Patient Death Events
0
1
32.1%
67.9%


age 1 -0.085 -0.086 -0.11 0.061 0.091 0.044 0.16 -0.043 0.069 0.02 -0.23 0.25
anaemia -0.085 1 0.2 0.007 -0.037 -0.043 0.17 -0.049 -0.046 0.1 0.1 0.14 -0.057
creatinine_phosphokinase -0.086 0.2 1 -0.012 -0.043 -0.067 0.11 -0.015 0.061 0.079 0.0078 -0.011 0.065
diabetes -0.11 0.007 -0.012 1 -0.0051 -0.013 -0.098 -0.047 -0.092 -0.16 -0.15 0.033 -0.0019
ejection_fraction 0.061 -0.037 -0.043 -0.0051 1 0.024 0.07 -0.011 0.18 -0.15 -0.071 0.043 -0.27
high_blood_pressure 0.091 -0.043 -0.067 -0.013 0.024 1 0.015 -0.0049 0.039 -0.1 -0.064 -0.2 0.079
platelets 0.044 0.17 0.11 -0.098 0.07 0.015 1 0.036 0.02 0.028 0.099 -0.057 0.039
serum_creatinine 0.16 -0.049 -0.015 -0.047 -0.011 -0.0049 0.036 1 -0.19 0.007 -0.028 -0.15 0.29
serum_sodium -0.043 -0.046 0.061 -0.092 0.18 0.039 0.02 -0.19 1 -0.028 -0.0027 0.086 -0.19
sex 0.069 0.1 0.079 -0.16 -0.15 -0.1 0.028 0.007 -0.028 1 0.44 -0.014 -0.0043
smoking 0.02 0.1 0.0078 -0.15 -0.071 -0.064 0.099 -0.028 -0.0027 0.44 1 -0.019 -0.011
time -0.23 0.14 -0.011 0.033 0.043 -0.2 -0.057 -0.15 0.086 -0.014 -0.019 1 -0.52
DEATH_EVENT 0.25 -0.057 0.065 -0.0019 -0.27 0.079 0.039 0.29 -0.19 -0.0043 -0.011 -0.52 1
age
anaemia
phokinase
diabetes
in_fraction
s_pressure
platelets
_creatinine
m_sodium
sex
smoking
time
TH_EVENT
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

2.2 Preprocessing

The dataset does not have much null values. Those null values have been replaced by their respective mean values. Preprocessing techniques play an important role when passing the data for classification or prediction purposes.

Normalization is used to put heterogeneous data into a smaller ranges such as 0 to 1, -1 to +1. Normally done when there is need to convert raw data into another format to make the processing efficient. We have applied it to bring different features like, credit amount, duration of credit, etc in to same a range. There are different normalization techniques such as, Min-Max, Decimal Scaling and Z=score normalization. We have used Min-Max normalization for the data set.

Table 2. Normalization (Min-Max)

Ejection Fraction	Normalized EF	Serum Creatinine	Normalized SC
20	0.090909	1.1	0.067416
38	0.363636	1.2	0.078652
45	0.469697	1.3	0.089888
60	0.696970	1.9	0.157303

Min-Max In this, for every feature minimum value get transformed into 0 and maximum values get transformed into 1 and all other values get transformed into value between 0 and 1.
Formula:

$$V' = (V - min(A))/(max(A) - min(A)) \tag{1}$$

2.3 Classification Models

Algorithms which takes some input and provides labels/categories in output are known as classification models.

Logistic Regression Supervised method used to predict the probability of input value to get the required prediction for classification problem. Independent variables are analysed to determine the binary outcome with the results falling into one of two categories, P(Y=0—X),P(Y=1—X). Logistic Regression is similar to Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'Logistic function' instead of a linear function. Hypothesis of logistic regression is to limit the cost function between 0 and 1 $0 \leq h\Theta(x) \leq 1$. Sigmoid simply tries to convert the independent variable X into a expression probability

that ranges between 0 and 1, with respect to dependent variable Y. Equation (2) represents formula for sigmoid function.

$$h\Theta(Z) = \frac{1}{(1 + e^{-Z})} \tag{2}$$

2.4 Clustering

Clustering is the process of dividing the entire data into groups(also known as clusters) based on the patterns in the data. It is a unsupervised learning method, as no target is provided, algorithm takes the independent data and tries to group them.

K-Mean Clustering minimizes the distance between points amoung clusters with centroids. K-means is a centroid based algorithm, where main objective is to calculate the Euclidean distance to assign a point to the cluster.

$$d = \sqrt{x^2 + y^2} \tag{3}$$

K-means has simple steps to create clusters. Start by randomly choosing K value i.e. number of clusters to be formed. Next, select number of centroids equal to k. Calculate the distance between each points and centroids, and assign points to the nearest centroid. Compute mean for each clusters, and reclcuate distances between points and new centroids. Repeat the process until the stopping creteria is met i.e. centroids of new formed clusters do not change or points remain in the same clusters or specified number of iterations are reached.

To decide value of k, one way is to try different values of k and plot a graph k versus calculated variance among the cluster of different k values. Then find a visible elbow in the graph, the point where elbow is formed is considered as optimal k value. The Elbow method is used to find the elbow in the elbow plot. The elbow is found when the dataset becomes flat or linear after applying the cluster analysis algorithm. The elbow plot shows the elbow at the point where the number of clusters starts increasing. By using Elbow Technique, we can identify the optimal k value. For example, from figure 1 k=2 has the curve/bend, so is the optimal number number of clusters to be formed.

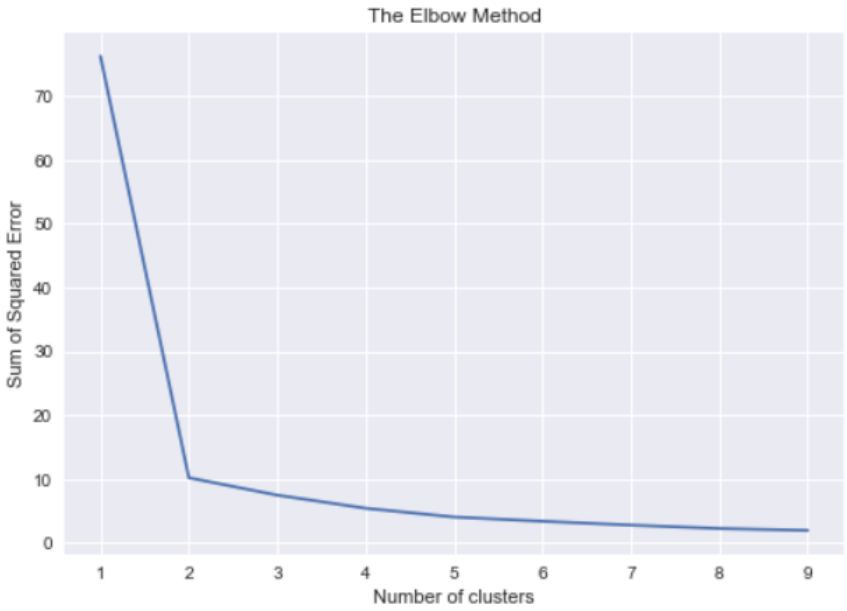


Fig. 1. Elbow Technique

3 Results

3.1 Evaluating Logistic Regression

Fig. 2. Success Rate and Confusion Matrix

3.2 Clusters

We depicted a scatterplot with the serum creatinine values on the x axis and the ejection fraction values on the y axis, and we colored every patient point based on survival status. This plot shows a clear distinction between alive patients and dead patients.

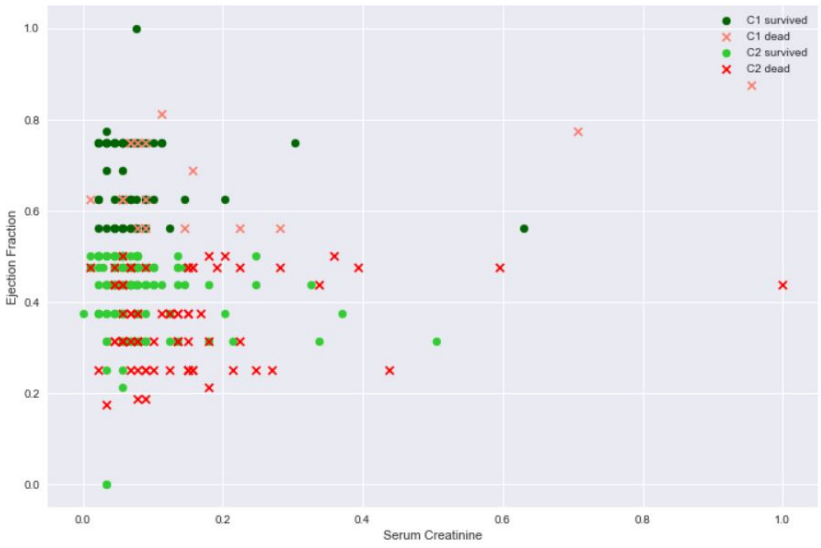


Fig. 3. Serum Creatinine versus Ejection Fraction

4 Conclusion

In our work, the fact that our traditional analysis selected ejection fraction and serum creatinine as the two most relevant features confirmed the relevance of the feature ranking executed with machine learning. Moreover, our approach showed that machine learning can be used on health records of patients with cardiovascular hearth diseases.

References

1. https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389
2. Guide to K-means