Term End Milestone-2 (Project -1)

Predicting the Mortality based on Clinical Data of patients with Cardiovascular Disease

DSC, Bellevue University

Supraja Rapuru

DSC680-T301 Applied Data Science (2225-1)

Professor Catie Williams

06/26/2022

Topic: Predicting the Mortality based on Clinical Data of patients with Cardiovascular

Heart failure occurs when the heart is not able to pump enough blood to the body. HF are only a subgroup of all the cardiovascular diseases that comprehend also coronary heart diseases (heart attacks), cerebrovascular diseases (strokes) and other pathologies that altogether kill every year approximately 17 million people around the world.

Machine learning applied to medical records can be useful to predict the survival of a patient, highlighting patterns and even ranking the features to understand which are risk factors, possibly undetectable by doctors.

Business Problem :

Cardiovascular Disease often used interchangeably with “heart disease”, generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves, or rhythm, also are considered forms of heart disease.

The purpose of this project is to predict the effects of different parameters recorded in the data to predict mortality of the patient. By predicting so the physicians can determine high risk patients and can take better care of them thus helping them survive.

Background :

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.

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, hyperglycaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Data Explanation:

This project uses the UCI Machine Learning Repository's heart failure clinical records dataset

Heart failure clinical records Data Set contains the medical records of 299 patients who had heart failure. The dataset contains 11 clinical features (some of them are binary, others are numerical), the follow-up period and the label DEATH_EVENT that indicates whether or not the patient has died.

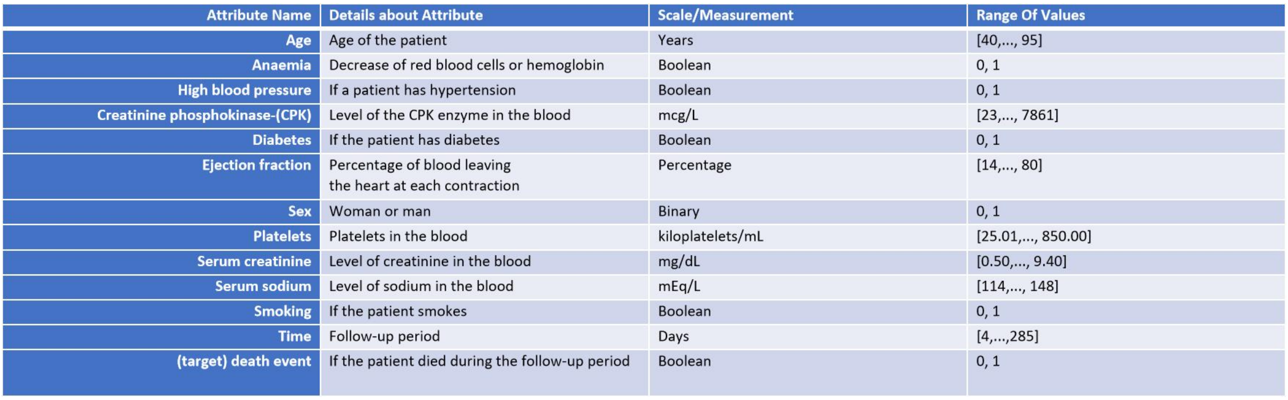
We can find some features strictly related to medical aspects like levels of enzymes, sodium, creatinine and platelets in the blood and others that are more common like age, sex or smoking.

We have 13 parameters in the dataset

We will be analyzing the effect of below parameters

- Age of Patient
- CPK Levels
- Ejection Fraction
- Platelets
- Serum Creatinine
- Serum Sodium

The table below lists out the details of all attributes in the dataset being used for the analysis



Attribute Name	Details about Attribute	Scale/Measurement	Range Of Values
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-(CPK)	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14,..., 80]
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

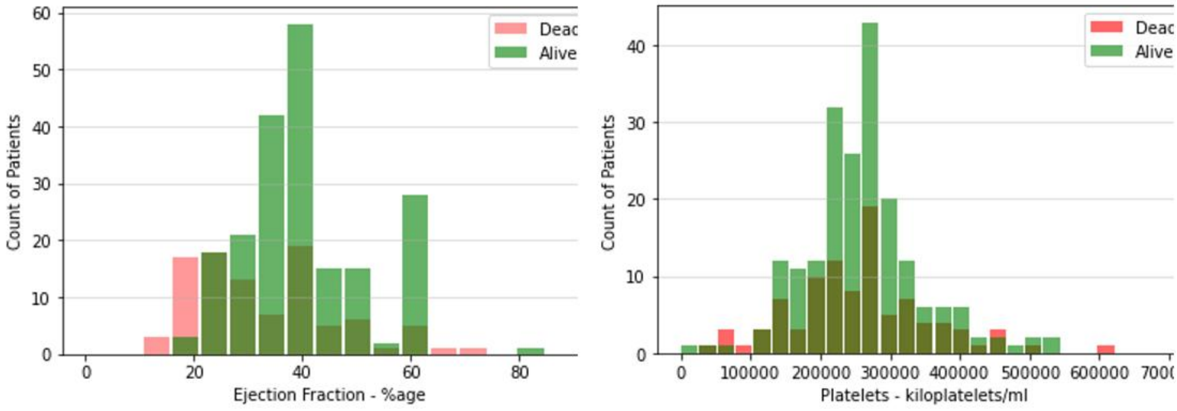
Methods:

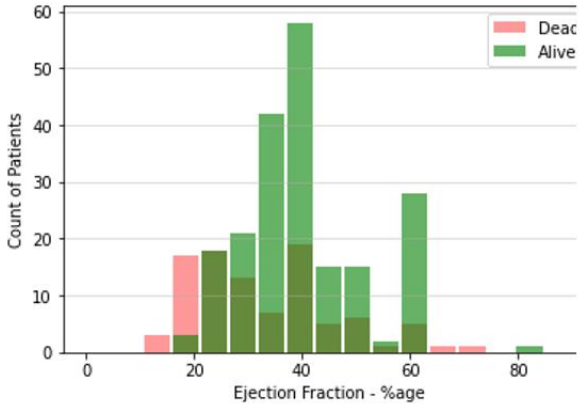
For this project, below series of classification models used to perform classification of the mortality.

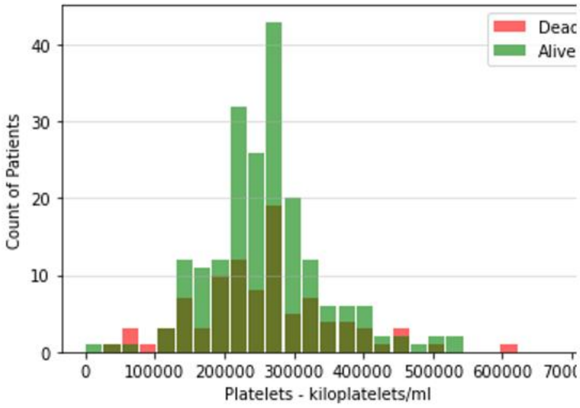
- Decision tree
- Random forest
- Linear regression
- Logistic regression
- Support vector machine (linear, poly, rbf)
- K nearest neighbors
- Naive bayes

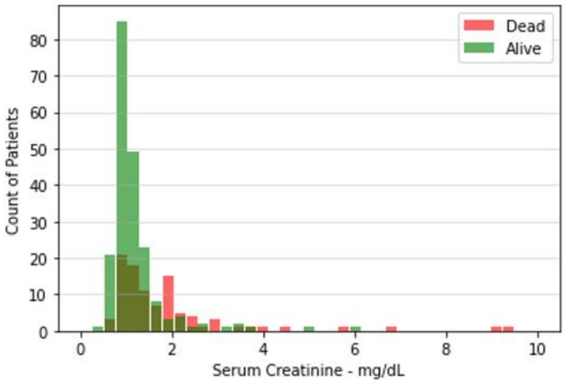
Analysis :

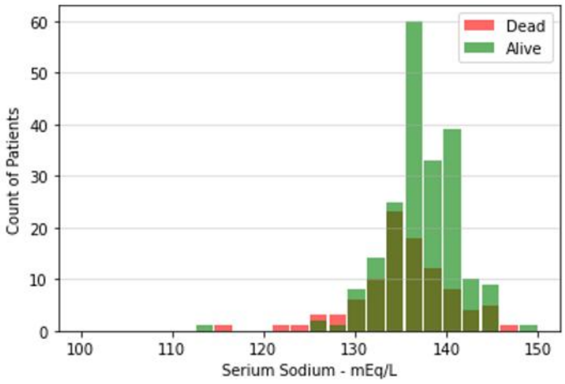
Histogram Plots for the variables in question







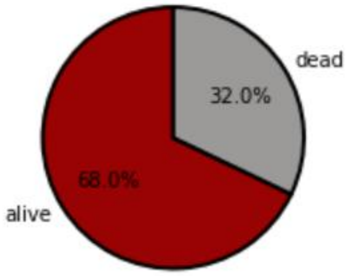




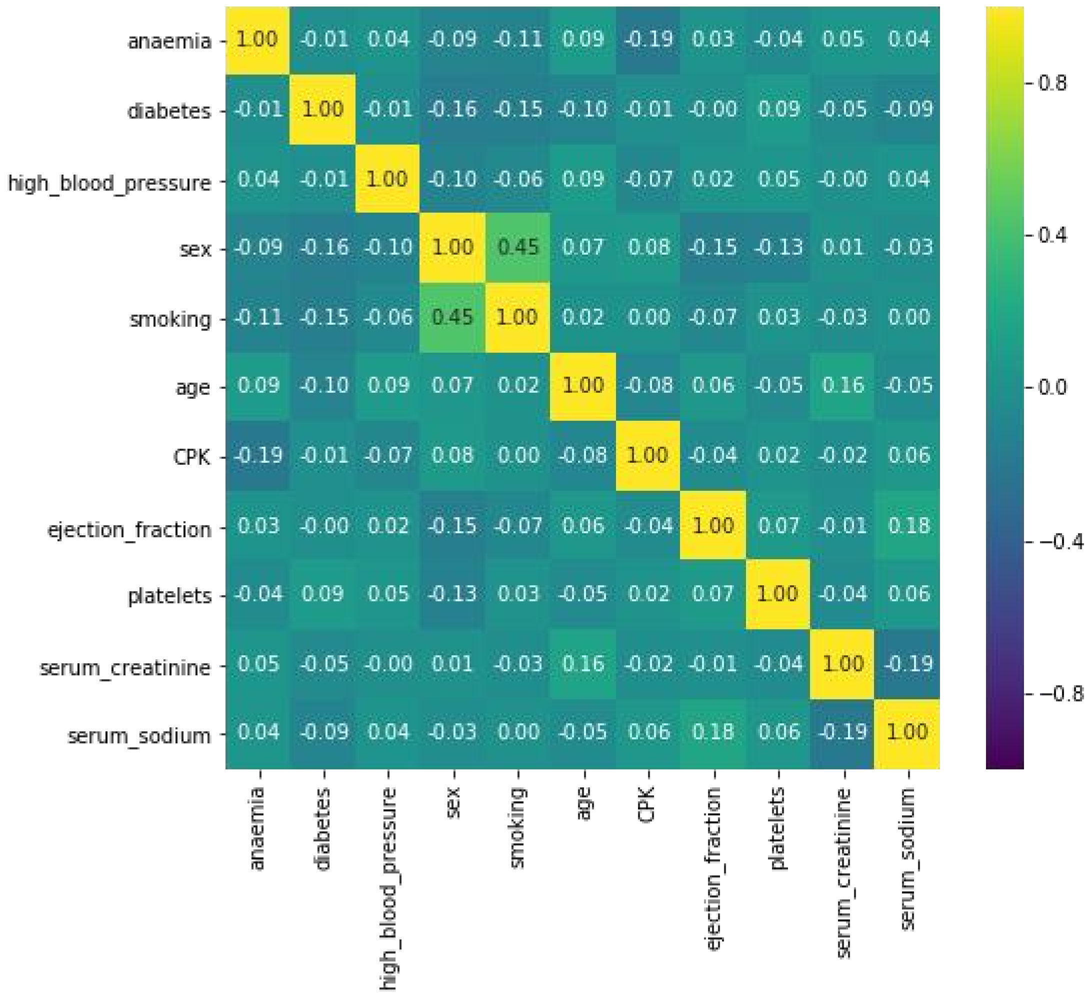
From the histograms, we may be able to see something but its not clearly visible

1. Age effects the mortality. As age increases the risk becomes higher
2. With lower Ejection fraction the risk increases
3. With higher serum creatinine levels risk increases
4. With lower serum sodium levels risk increases

One thing to take into account is the possible class imbalance.



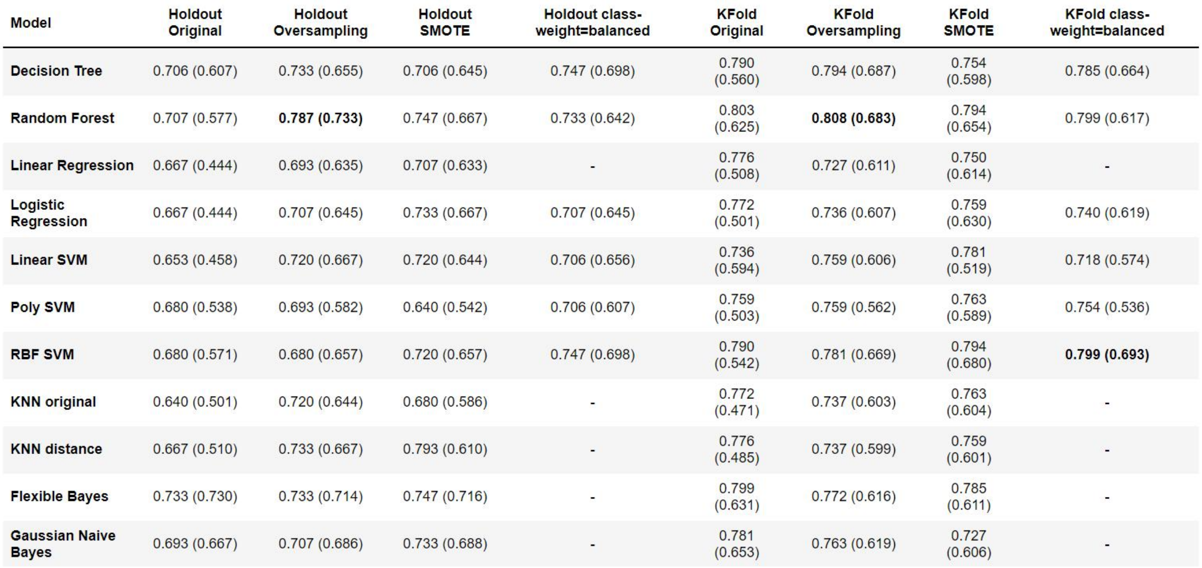
As we can see, even if not so strong, there is a class imbalance. This can leads to biased results that can be noticed by measures such as recall, precision or f1.To handle class inbalance it's possible to re-balance the dataset with different techniques. I used SMOTE for this project and my implementation of SMOTE is based on the original paper by N. V. Chawla et al. [2].



From the heatmap we can see that in general, features are quite uncorrelated with the exception of sex and smoking that seems to be slightly positively correlated.

Conclusion:

Here we can see the results obtained with different models and different rebalancing techniques for the Hearth Disease dataset. Both accuracy and f1 score (inside parenthesis) are showed.



Model	Holdout Original	Holdout Oversampling	Holdout SMOTE	Holdout class-weight=balanced	KFold Original	KFold Oversampling	KFold SMOTE	KFold class-weight=balanced
Decision Tree	0.706 (0.607)	0.733 (0.655)	0.706 (0.645)	0.747 (0.698)	0.790 (0.560)	0.794 (0.687)	0.754 (0.598)	0.785 (0.664)
Random Forest	0.707 (0.577)	**0.787 (0.733)**	0.747 (0.667)	0.733 (0.642)	0.803 (0.625)	**0.808 (0.683)**	0.794 (0.654)	0.799 (0.617)
Linear Regression	0.667 (0.444)	0.693 (0.635)	0.707 (0.633)	-	0.776 (0.508)	0.727 (0.611)	0.750 (0.614)	-
Logistic Regression	0.667 (0.444)	0.707 (0.645)	0.733 (0.667)	0.707 (0.645)	0.772 (0.501)	0.736 (0.607)	0.759 (0.630)	0.740 (0.619)
Linear SVM	0.653 (0.458)	0.720 (0.667)	0.720 (0.644)	0.706 (0.656)	0.736 (0.594)	0.759 (0.606)	0.781 (0.519)	0.718 (0.574)
Poly SVM	0.680 (0.538)	0.693 (0.582)	0.640 (0.542)	0.706 (0.607)	0.759 (0.503)	0.759 (0.562)	0.763 (0.589)	0.754 (0.536)
RBF SVM	0.680 (0.571)	0.680 (0.657)	0.720 (0.657)	0.747 (0.698)	0.790 (0.542)	0.781 (0.669)	0.794 (0.680)	**0.799 (0.693)**
KNN original	0.640 (0.501)	0.720 (0.644)	0.680 (0.586)	-	0.772 (0.471)	0.737 (0.603)	0.763 (0.604)	-
KNN distance	0.667 (0.510)	0.733 (0.667)	0.793 (0.610)	-	0.776 (0.485)	0.737 (0.599)	0.759 (0.601)	-
Flexible Bayes	0.733 (0.730)	0.733 (0.714)	0.747 (0.716)	-	0.799 (0.631)	0.772 (0.616)	0.785 (0.611)	-
Gaussian Naive Bayes	0.693 (0.667)	0.707 (0.686)	0.733 (0.688)	-	0.781 (0.653)	0.763 (0.619)	0.727 (0.606)	-

We can clearly see how using some rebalancing techniques the f1 score increase substantially. In some cases SMOTE performs better with respect to random oversampling, and the opposite in others. Furthermore, where is possible to apply it, also the use of the class-weight parameter increases the performances, sometimes outperforming the other techniques.

Then we noticed how using Gaussian Naive Bayes, even without respecting the hypothesis, leads to good results and also with a Bayes Classifier with KDE, the results are in line.

Best overall model seems to be the random forest trained on the over sampled dataset, that delivers the best results in terms of accuracy and f1 score.

Also RBF-SVM with class-weights=balanced provides some good results on KFold.

For the models that allow it, it's possible to evaluate the ROC curve to select a threshold according to the main goal (minimize false positives or maximize true positives) but the results in the table are obtained by fixing the threshold at 0.5.

The overall results seem in line with the ones obtained in the reference paper [1] but it's needed to keep in mind that the metrics are highly influenced by the small dimension of the dataset (75 samples in holdout validation set).

Assumption:

There were some concerns regarding the sample size of the data. Since the data was from a single location, there might be some other factors in play which can be due to habits of people to that specific region or part of the world and may not be reflecting in the data. If we get more geographically separated data, we might come up with better model. This model may be more effective in the region where data comes from but our assumption that it will work for all might be wrong.

Limitations:

While analyzing separately, I did notice time (in days) between follow up visits was making the risk lower. It is anyways evident that if you go on regular follow ups, you will be able to know the problem before its too late and you may have the opportunity to act to decrease the risk.

Challenges:

The biggest challenge in dealing with the small dataset, imbalanced and unavailability of time duration between visits. Dealing with this information while still finding meaningful patterns was a challenge.

Future Uses:

This information can be used as supporting material by medical professionals and suggest preventive course of action in time for patients. Also can be extended into smart wearable devices so that patients can be alerted/reminded the actions to be taken.

Recommendations:

To ensure risk factors are properly identified in the collected values, I recommend obtaining clinical information about the patients at periodic intervals and calibrate model

Implementation Plan:

As we already discussed in the methodology section about some of the implementation details.

So, the language used in this project is Python programming. We're running python code in anaconda navigator's Jupyter notebook. Jupyter notebook is much faster than Python IDE tools like PyCharm or Visual studio for implementing ML algorithms. The advantage of Jupyter notebook is that while writing code, it's really helpful for Data visualization and plotting some graphs like histogram and heatmap of correlated matrices.

Ethical Considerations:

- This Data contains processed physical images information related to multiple verities of rice and does not contains any PII-related information.
- Datasets and information on data were extracted from the public websites → UCL machine learning repositories.
- This data research is not going to harm any privacy.

References :

[1] D. Chicco, G. Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone", 2020

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, 2002

[3] G. H. John, P. Langley. Estimating Continuous Distributions in Bayesian Classifiers, 1995

10 Questions from the Audience

What are the significant risk factors you identified which may lead to HF from this case study?

How is your model accuracy defined and is it reliable enough?

What is the timeline for implementation?

How are you plan to govern and calibrate the model to ensure the model is working as intended?

Are there any additional checks or due diligence expected by consumers of these model in case of false negatives?

Should we be concerned that the model could potentially adversely impact certain demographic groups?

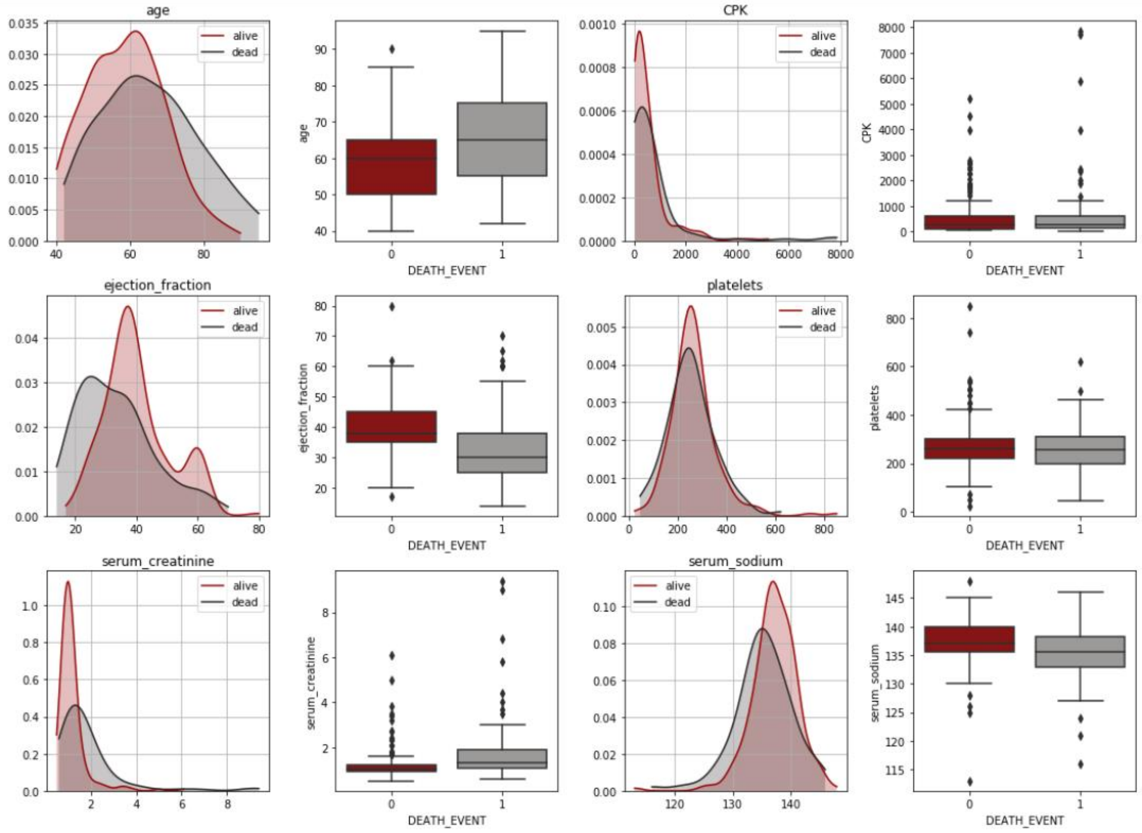
What do you suggest to make the model better?

What is your plan to get buy in from medical professionals to use this model?

Are there similar solutions in the market?

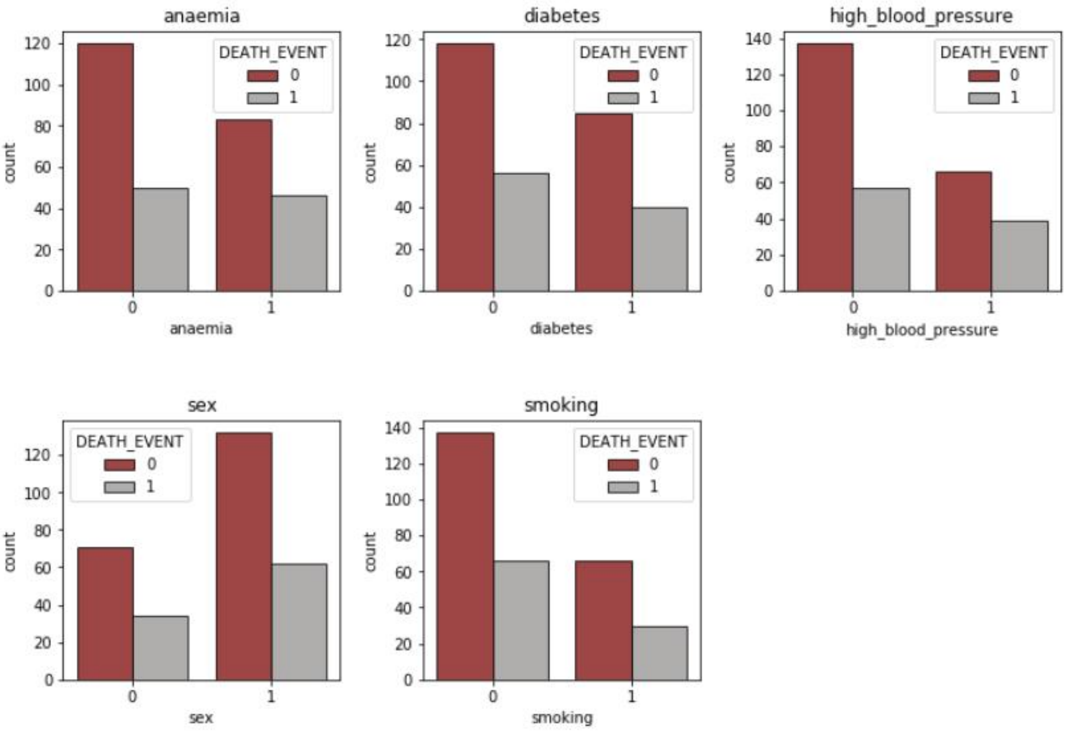
What is your plan to explain AI/ML to patients and gain confidence in using this model?

Appendix :

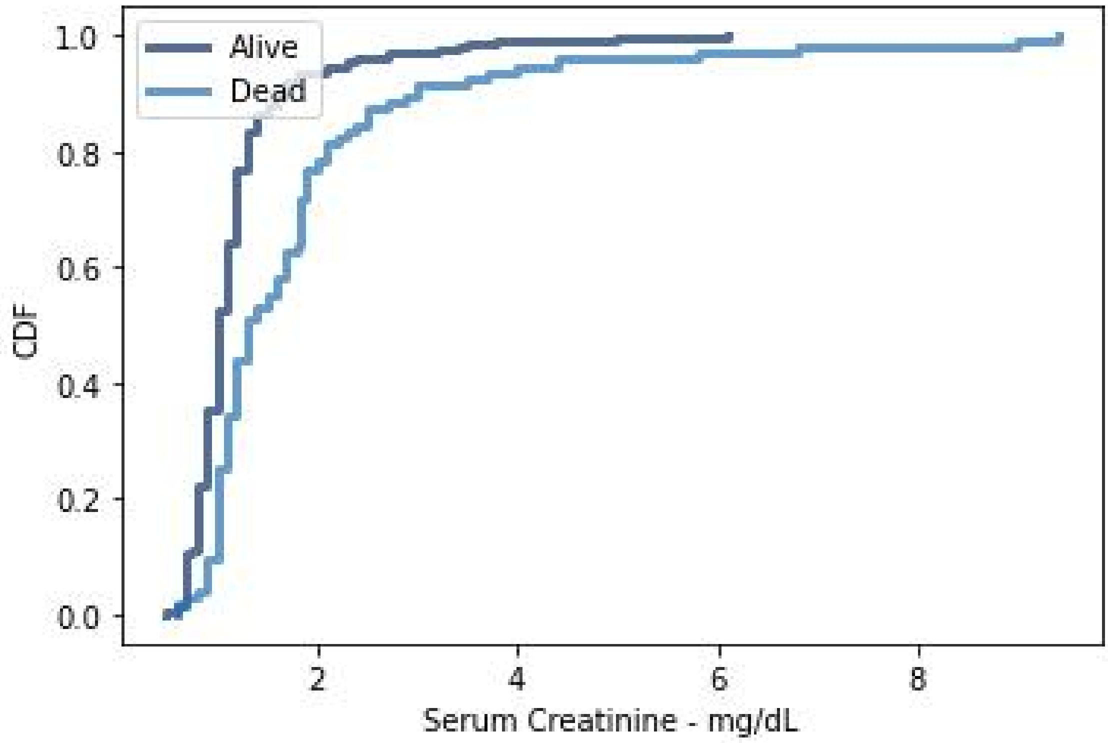


As we can see already from the unnormalized distribution plot of the features, the most informative ones seem to be **ejection_fraction** and **serum_creatinine**. This is confirmed by the fact that in the original paper by Chicco and Jurman [1] the analysis has been conducted taking into account only these two features.

Here are reported the categorical features



CDF – Serum Creatinine



From the CDF we see that for both Alive and Dead Patients, the max of the population lies between 1 & 2. As per the healthline.com which says:

“In general, however, normal creatinine levels range from 0.9 to 1.3 mg/dL in men and 0.6 to 1.1 mg/dL in women who are 18 to 60 years old. Normal levels are roughly the same for people over 60. High serum creatinine levels in the blood indicate that the kidneys aren't functioning properly.”

So our data seems to reflect the same. Also if we look at the CDF for Alive vs Dead Patients, the latter population is shifted more towards the higher numbers thus supporting our study