Practical Data Science

Student ID: s3899733
Student Name: Kelvin Young
Student Email: s3899733@student.rmit.edu.au
Affiliations: RMIT University
Date of report: 20/05/2022
I certify that this is all my own original work. If I took any parts from elsewhere, then they were non essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honour code by typing “Yes”: *Yes.*

Table of Contents

Abstract/Executive Summary

The task is to determine whether a heart will or will not fail, based on twelve clinical features. We want to determine which features will be the most significant contributors in determining death. Two classification models will be used, and the end goal of this report is to give a recommended model in order to best achieve this task.

In this report we will preprocess and analyse a dataset of 299 patients, using both decision tree and k nearest neighbours classifiers to predict the survival of a patient. We will optimise both classifiers using feature selection and hypertuning, and derive a conclusive model recommendation based on results.

Our results show that only half of the dataset is useful in terms of determining the target variable: *serum creatinine, creatinine phosphokinase, age, time, serum sodium, and ejection fraction*. Further improvements made to this reduced feature count model result in more accurate predictions than the original dataset itself.

Introduction

For this report, the *Heart failure clinical records* dataset has been selected for classification modelling. The original dataset was provided by *Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza* from the *Government College University, Faisalabad, Pakistan.* The revised version of the dataset used in this report was provided by *Davide Chicco* from the *Krembil Research Institute, Toronto, Canada.*

A surface level explanation of the dataset is sourced from Chicco's paper; this table is also referenced by the UCI repository in lieu of data set information:

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or haemoglobin	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

Individual feature analysis

(the dot point number refers to the figure number in the jupyter notebook file):

1. The distribution of age appears to be right skewed, with a concentration of ages between [40, 80]. The average age of a recorded patient is 61 years old.
2. The distribution of anaemia is non-equal, there are 170 instances of 0 and 129 instances of 1. Therefore, of the 299 recorded patients, 56.86% had anaemia while the remaining 43.14% did not.
3. The distribution of the creatinine phosphokinase enzyme is massively right skewed; there is an intense concentration of patients with CPK values between [0, 1000]. Remarkably, 50% of the values can be found between just [23, 250]. The average CPK value was 581.84.
4. The distribution of diabetes is non-equal, there are 174 instances of 0 and 125 instances of 1. Therefore, of the 299 recorded patients, 58.19% had diabetes while the remaining 41.81% did not.
5. The distribution of ejection fractions appears to be slightly right skewed, with a large concentration of values between [35, 40]. There is also a smaller, isolated concentration of values between [55, 60]. The average value recorded was 38.08.
6. The distribution of high blood pressure is largely non-equal, there are 194 instances of 0 and 105 instances of 1. Therefore, of the 299 recorded patients, 64.88% were considered to have had high blood pressure, while the remaining 35.12% were not.
7. The distribution of platelets is right skewed, with a concentration of values between [200000, 400000]. The distribution features a large tail on the right. The average recorded value was 263358.03.
8. The distribution of serum creatinine is massively right skewed, with a very large concentration of values between [.5, 2]. Remarkably, 75% of the value can be found just between [.5, 1.4]. The average recorded value was 1.39.
9. The distribution of serum sodium is unexpectedly left skewed, with a concentration of values between [135, 140]. The average value was 136.63.
10. The distribution of biological sex is non-equal, there are 105 instances of 0 and 194 instances of 1. Chicco’s revision of the dataset defines 0 to be female, and 1 to be male; we will use the same convention. Of the 299 patients, 105 were male and 194 were female, resulting in a biological gender ratio of .54:1.

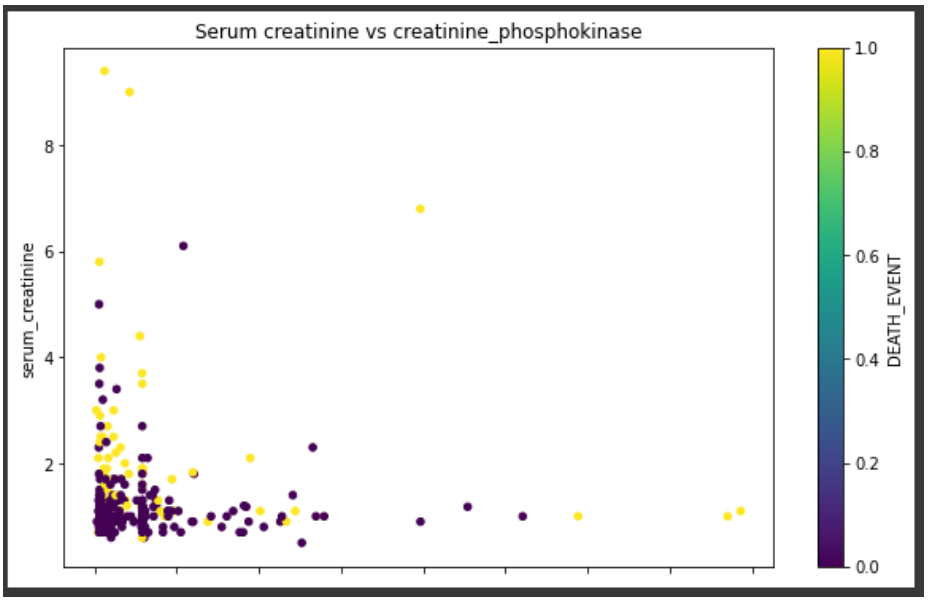
11. The distribution of smoking is non-equal, there are 203 instances of 0 and 96 instances of 1. Therefore, of the 299 patients, 67.89% did not smoke, while the other 32.11% did.
12. The distribution of follow-up time is tri-modal; there are three distinct peaks as seen in the figure. The average value was 130.26.
13. The distribution of the target variable death event is non-equal, there are 203 instances of 0 and 96 instances of 1. Therefore, of the 299 patients, 32.11% survived while the other 67.89% did not.
    a. Notably for our target variable; an unbalanced distribution could result in a skewed model, where the dominant class value could compose a statistically impactful majority of our training or testing set. Therefore, it is decided here in the process that we will use k folds cross validation in order to nullify this potential issue.
14. This figure shows the distribution of the target variable across the physical dataset itself; this is used to determine the necessity to stratify our training and testing split. Evidently, there is a large concentration of deaths in the first quadrant of the indices, therefore we will use stratified cross validation, for the same reasons above.

Feature pair analysis

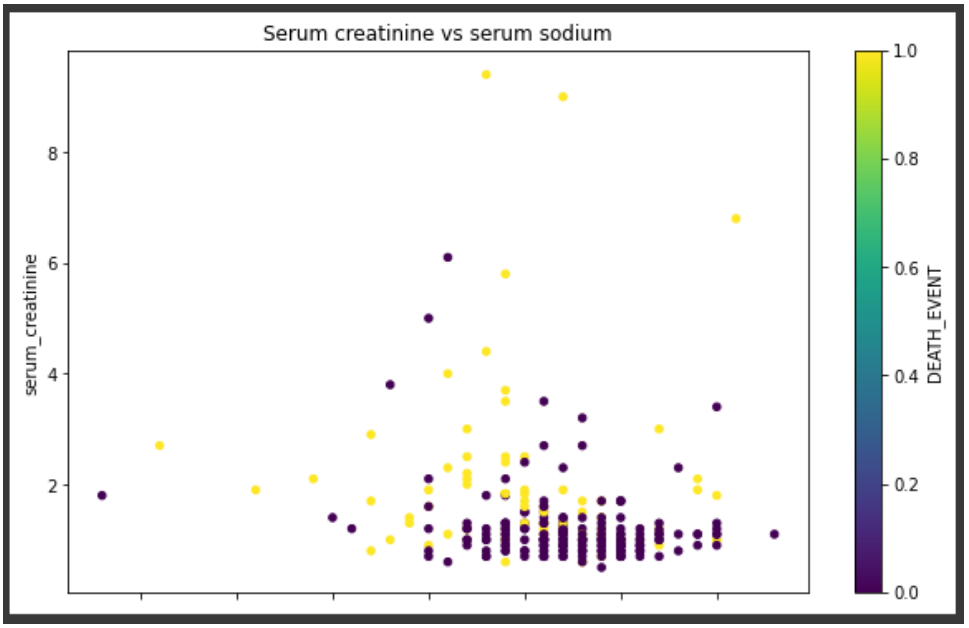
As there are a relatively large amount of features in this dataset, the feature pair analysis will be limited to the non-categorical kind.

15. A scatter matrix has been created in order to identify relationships amongst the numeric variables. All individual features are also visible here as kernel density estimations in the leading diagonal; these serve to verify our prior analysis of said features.
    a. Key feature pairs that stand out in figure 15 are those that exhibit polynomial relationships:
        i. Creatinine phosphokinase and serum creatinine
        ii. Serum sodium and serum creatinine
        iii. Serum creatinine and ejection fraction
    b. All key pairs will be graphed against each other, with a target label colormap applied in order to observe each pair's impact on the target feature.
16. There exists an inverse relationship between creatinine phosphokinase and serum creatinine, bar one rogue point. However, the distribution of the target value across this relationship appears to be random. This suggests that the physical impact on the

target value of both of these features are moderately separated, but non-trivial.

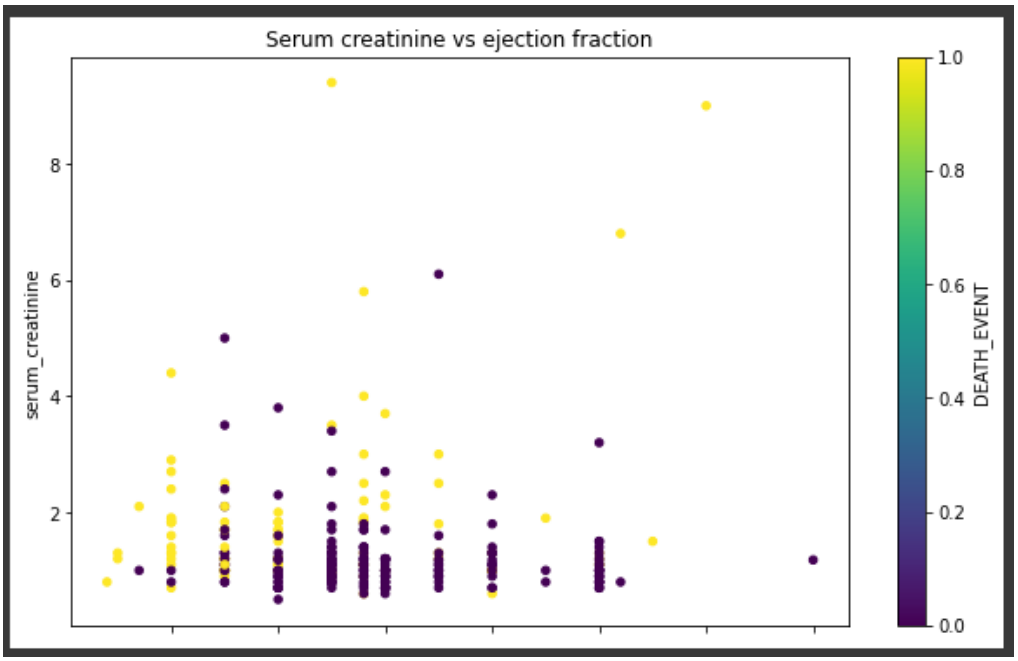


17. While the distribution of serum creatinine against serum sodium is mostly centred around a group of values on the bottom right of the graph, there are some very distinct curves and linear patterns of values that are disjoint from the main body. Also notable is the main body being mostly composed of target value 0; the distribution of target values in this graph are clearly defined and separate. This suggests that both serum creatinine and serum sodium may be impactful factors on the determination of death events.

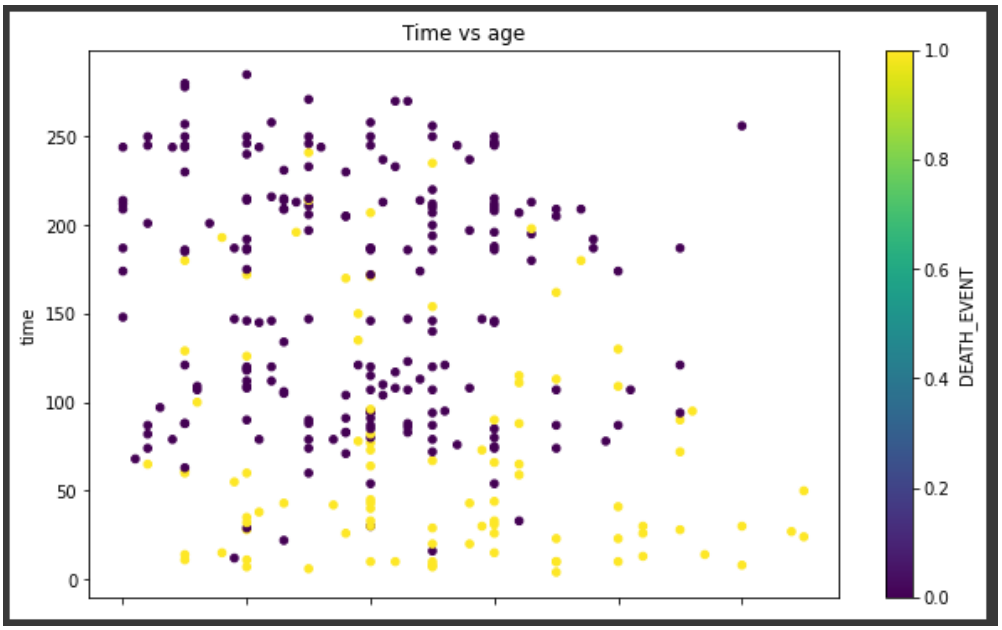


18. Almost identical behaviour to figure 17 is also exhibited here in the relationship between serum creatinine and ejection fraction. There is a very clear distinction between target values; further suggesting that ejection fraction is also an important factor in classifying the target variable (in addition to the previously hypothesised

serum creatinine).



19. While figure 19 does not exhibit any polynomial relationship, it is instead notably random. Whereas other graphs in the scatter matrix exhibit clustering around certain values, the time and age distribution is rather noisy. There is again a very clear distinction between target values here; suggesting again that now both time and age may be important factors in classifying the target variable.



Based on our feature pair analysis, we **hypothesise for** *serum creatinine, creatinine phosphokinase, age, time, serum sodium, and ejection fraction* to be key determining factors in the classification of death events. Therefore we expect these variables to be a part of the incoming feature selection, and dominant in the composition of the decision tree model.

Methodology

The dataset is first pre-processed to ensure it is fit for modelling. The jupyter notebook code segments are labelled by instruction, here in the report.

1. The dataset is checked for non-numeric data types (in the event that a number may be represented as a string of characters for example).
    a. As per the output of the cell, all thirteen of the columns are of numeric type. Notably, the datatype of the age column is a 64 bit float, rather than an integer as expected. Having age be represented by a float may result in some non-whole values.
    b. Two values of age are found to host decimal values, two instances of age 60.667. Therefore, in order to enforce consistency, the entire age column in type casted into integers.
2. The dataset is checked for numeric data values beyond constraints.
    a. The output of this cell is mostly in accordance with *Chicco and Jurman's* supplied constraints on the dataset, except for the platelets column. While the given kiloplatelet interval is [25.01, 850.00], the dataset itself yields an interval of [25100, 850000] instead.
    b. It is obvious that *Chicco and Jurman* have transposed the data; in response we will simply redefine our definition of the feature to instead be just platelets per mL rather than kiloplatelets.
3. The dataset is checked for null values.
    a. There are none, so there are no further steps.

One hot encoding was tested for its application and efficacy with this particular dataset, but we found it had very little impact on the results. Therefore the code has been left in for completion but will not be used to limit dataset manipulation.

After processing the data, now we can build our two models. The two classification models will be the decision tree, and k nearest neighbours.

The training and testing split will be 80:20, as per standard. The MinMaxScaler() from the sklearn.preprocessing library will be used in favour of the StandardScaler(), as to prevent modifying the distribution of the data itself. The dataset is also not visibly normally distributed either.

Results

Disclaimers: Results vary slightly when code is run multiple times. When referring to the pair of classification models collectively, they will be referenced in order of decision tree, followed by k nearest.

The base models constructed directly from the preprocessed data yield accuracy results of 0.70 and 0.55 respectively. Here the MinMaxScaler() has been used. Notably, the precision and recall are far worse for k nearest when determining a target value of 1 (i.e. the patient

dies). Interestingly the decision tree accuracy on the training set is a perfect 1. The k nearest recall of target value 1 is also astonishingly low at 0.08.

```
Decision tree results: [1.00, 0.70]
[[27  8]
 [10 15]]
              precision    recall  f1-score   support

           0       0.73      0.77      0.75        35
           1       0.65      0.60      0.63        25

    accuracy                           0.70        60
   macro avg       0.69      0.69      0.69        60
weighted avg       0.70      0.70      0.70        60
```

```
k nearest results: [0.77, 0.55]
[[31  4]
 [23  2]]
              precision    recall  f1-score   support

           0       0.57      0.89      0.70        35
           1       0.33      0.08      0.13        25

    accuracy                           0.55        60
   macro avg       0.45      0.48      0.41        60
weighted avg       0.47      0.55      0.46        60
```

Feature selection

After applying feature selection to both models their accuracies become 0.90 and 0.77 respectively. The recall and f1-score of the k nearest model sees remarkable improvements; the accuracy is much improved at 0.77.

```
Test set accuracy with features selected:
serum_sodium: 0.5666666666666667
anaemia: 0.5833333333333334
ejection_fraction: 0.6
time: 0.7166666666666667
creatinine_phosphokinase: 0.7666666666666667
Decision tree feature selection results: [1.00, 0.90]
[[38  5]
 [ 1 16]]
              precision    recall  f1-score   support

           0       0.97      0.88      0.93        43
           1       0.76      0.94      0.84        17

    accuracy                           0.90        60
   macro avg       0.87      0.91      0.88        60
weighted avg       0.91      0.90      0.90        60
```

```
Test set accuracy with features selected:
age: 0.6666666666666666
ejection_fraction: 0.7
high_blood_pressure: 0.7333333333333333
serum_creatinine: 0.7666666666666667
serum_sodium: 0.8
time: 0.8333333333333334
k nearest feature selection results: [0.87, 0.77]
[[30  6]
 [ 8 16]]
              precision    recall  f1-score   support

           0       0.79      0.83      0.81        36
           1       0.73      0.67      0.70        24

    accuracy                           0.77        60
   macro avg       0.76      0.75      0.75        60
weighted avg       0.76      0.77      0.76        60
```

As hypothesised, the text representation of the decision tree includes *time, serum creatinine, age, ejection fraction, creatinine phosphokinase, and serum sodium*. The selected features for the k nearest model also include *age, ejection fraction, serum creatinine, serum sodium, and time.* These observations support our hypothesis that the stated variables are relevant classification factors.

Note: The decision tree is represented using sklearn's tree.export_text() function in order to support the smaller screen size used by the author. As the text export is rather vertically large it is not shown here in the report, refer to figure 20 of the code instead).

Parameter hypertuning

Now we test parameter hypertuning using stratified k folds cross validation. Nominated parameters for the decision tree model include *max_depth, min_samples_leaf, and criterion,* while parameters for the k nearest model include *leaf_size, n_neighbours (equivalent to k), and p.* Hypertuning is applied **after** feature selection, therefore the number of features at this stage for both models are almost halved.

After applying both parameter hypertuning and stratified k folds cross validation to both models, their accuracies become 0.87, and 0.77 respectively. Notably, the precision for the k nearest model becomes 1.00, while the recall drops back down to 0.39.

```
Hypertuned decision tree parameters:
DecisionTreeClassifier(max_depth=2, min_samples_leaf=46)
Hypertuned training set accuracy: 0.84
Hypertuned test set accuracy: 0.87

Hypertuned k nearest parameters:
KNeighborsClassifier(leaf_size=8, n_neighbors=28, p=1)
Hypertuned training set accuracy: 0.85
Hypertuned test set accuracy: 0.77

Stratified kfolds cross validation hypertuned feature selection decision tree model:
[[40  2]
 [ 6 12]]
              precision    recall  f1-score   support

           0       0.87      0.95      0.91        42
           1       0.86      0.67      0.75        18

    accuracy                           0.87        60
   macro avg       0.86      0.81      0.83        60
weighted avg       0.87      0.87      0.86        60

Stratified kfolds cross validation hypertuned feature selection k nearest model:
[[37  0]
 [14  9]]
              precision    recall  f1-score   support

           0       0.73      1.00      0.84        37
           1       1.00      0.39      0.56        23

    accuracy                           0.77        60
   macro avg       0.86      0.70      0.70        60
weighted avg       0.83      0.77      0.73        60
```

The confusion matrices for both models see a reduction in false estimations despite both models suffering in accuracy and recall scores.

Discussion

The base decision tree and k nearest models first generated by the code see some interesting results. A notably low recall value for the k nearest model suggests that a large majority of true positive predictions made by the classifier on the test set were incorrect, further suggesting that the base k nearest model is massively overfit. The base decision tree model has no similarly outstanding results to warrant discussion.

The feature selected models see large improvements to both models; the k nearest model is no longer visibly overfit, and the decision tree model becomes extremely performant on the given training data.

The parameter hypertuned models see notably worse results; both model's recall value drops significantly, and likewise both models produce worse confusion matrices. Overall, the impact of hypertuning here is negligible, if not almost impactless.

Feature selection yielded most of the features we hypothesised earlier to be statistically significant: *serum creatinine, creatinine phosphokinase, age, time, serum sodium, and ejection fraction.* These results are in line with what others have observed on the same dataset. Further improvements made to this reduced feature count model result in more accurate predictions than the original dataset itself.

Chicco states that *"[we] identified age, serum creatinine (renal dysfunction), high blood pressure, ejection fraction and anaemia as top features".* This is similarly reflected in our report. Meanwhile Ahmed states that *"growing age, renal dysfunction (serum creatinine), [high blood pressure], higher level of anaemia and lower values of ejection fraction are the key factors"*. Our results are less similar but share many features.

By our own observations, we recommend the decision tree model, along with regular pipeline functions such as feature selection, k folds cross validation, and parameter hypertuning.

Conclusion

In conclusion, while the failure of the heart is not completely deterministic, our findings have shown correlations between *serum creatinine, creatinine phosphokinase, age, time, serum sodium, and ejection fraction* as key factors in the accurate prediction of their death.

Some improvements to be made to this investigation would be, regarding the k folds, to use k folds for only half the data set (i.e. use a validation set), and to only do feature selection on each individual k fold.

Currently, k folds are applied to the entire dataset, for both the training and test sets. As a result, the model is trained well for our particular distribution of the class variable, death_event, but is only trained well for this particular spread of target values. And because the test set bears the same target value ratio due to k folds, the results are biassed because of it. This is explored further in *The Elements of Statistical Learning* (Hastie et al.), where Hastie states that the sample "does not correctly mimic… a completely independent test set, since these predictors *have already seen* the left out samples". Hastie proceeds to explain how to avoid this issue, which is to apply feature selection per cross validation fold instead; in a sense the feature selection is blind to the cross validation within its own scope thus avoiding selection bias.

However, Hastie's proposed solution is computationally expensive, relative to the size of the dataset. While this is not a relevant issue to this particular report, a more sensible and efficient solution would be to split the dataset into two; with one half eligible for k folds application and the other ineligible. By having the test set be completely independent from the training set, in terms of data transformation, the classifier can no longer utilise its internal

bias to skew results. Another name to describe this would be a validation set; this result would instead better reflect the real world application of this particular model.

References

Heart failure clinical records Data Set
https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
Centre for Machine Learning and Intelligent Systems
Retrieved 20/05/2022

Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone
BMC Medical Informatics and Decision Making 20, 16 (2020)
Chicco D, et Jurman G. (et al)
https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5
Retrieved 20/05/2022

Survival analysis of heart failure patients: A case study
PLoS ONE 12(7), 0181001 (2017)
Ahmad T, Munir A, Bhatti SH, Aftab M, et Raza MA. (et al)
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001
Retrieved 20/05/2022

How NOT to perform feature selection!
Christos - Iraklis Tsatsoulis
https://www.nodalpoint.com/not-perform-feature-selection/
Retrieved 21/05/2022

The Elements of Statistical Learning, page 245
Published by Springer New York, NY
Hastie et al, 2009
Retrieved 21/05/2022