Heart Failure Analysis

Salman Salman

October 28, 2022

Abstract

Doctors use the term 'cardiovascular diseases (CVD)' to refer to a broad array of heart and circulation diseases (NHS, 2022). This includes both hereditary conditions and those that develop later in life, such as heart failure and stroke. Around 7.6 million people in UK are currently living with a CVD (BHF, 2022, p. 1). Heart disease early detection is a critical concern for people at risk of developing a CVD. The focus of this report will be the analysis of patients with heart failure using machine learning. The analysis includes an exploration of the data to reveal hidden patterns and find correlations, a supervised analysis (using several classification algorithms) to predict if a given patient is likely to die, and lastly, the patients will be clustered and analysed for similarities and patterns. The data set used in this report can be found here.

1 Background and Introduction

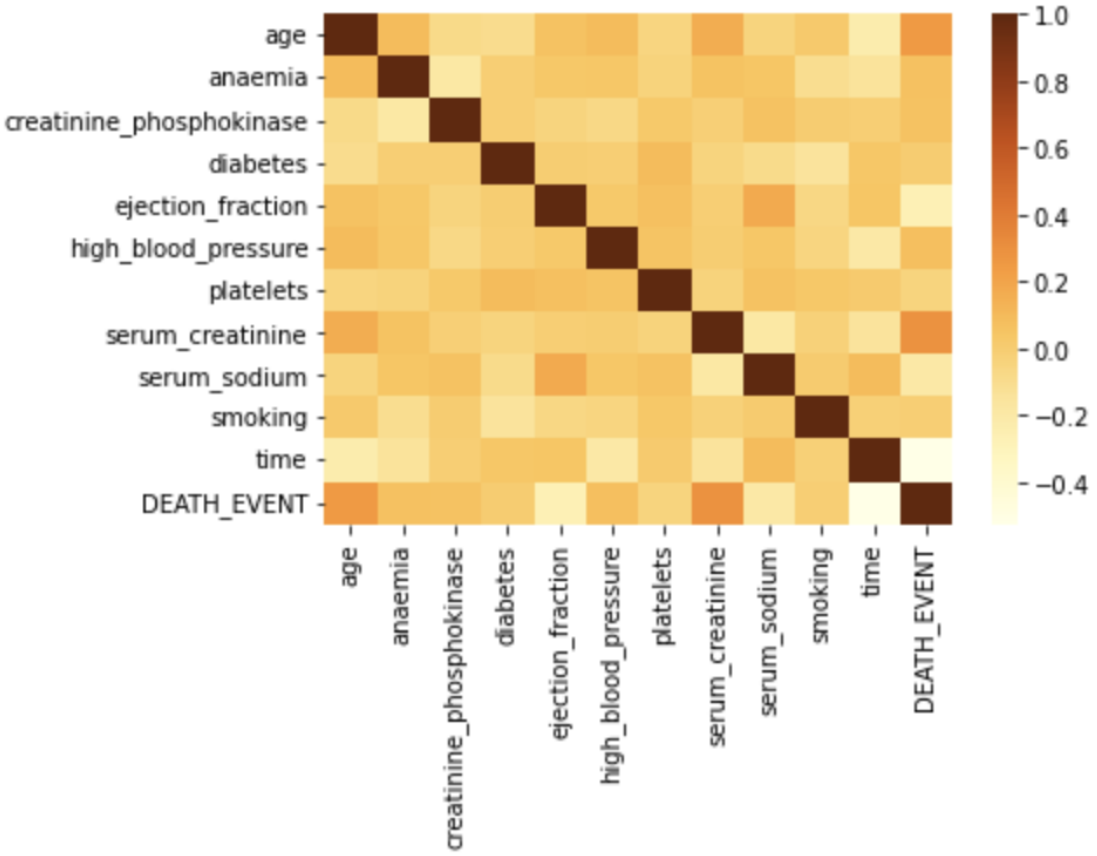
Heart failure occurs when the heart is no longer pumping blood properly around the body. According to the British Heart Foundation, there are around 200,000 new diagnoses of heart failure every year across the UK. People with heart failure were found to be 2-3 times more likely to have a stroke, which is the biggest killer in the UK, causing around 340,000 death year (BHF, 2022, p. 9). The most common risk factors are smoking, diabetes, and high blood pressure (NHS, 2022).

In this report these factors will be analysed in relation to each other and to other features in the data set. In machine learning, the data set is usually explored before applying any algorithms to it. The reasoning behind data exploration is to untangle the data and find relevant correlations that could be useful for our purposes. In our case, we are trying to understand heart failure, its causes and the patterns that occur in patients with that condition, including their medical results (sodium levels, e.g.). It is important to keep in mind that the results presented here are merely correlative and not conclusive. Machine learning algorithms can only uncover trends and patterns within the data but they cannot replace medical advice.

2 Exploring the Data

The data consists of 13 columns and 299 rows, with each row representing a patient. The columns are multivariate, meaning that some are binary while others are numerical. Of importance are the columns that contain risk factors, which could be utilised to predict a heart failure. These are diabetes, high blood pressure, and smoking. Other columns that serve as indicators of heart damage such as creatinine phosphokinase and ejection fraction will also be analysed and explored. Let us first look at each column in its medical context and in relation to other columns in the data set.

Figure 1: Heatmap for all columns

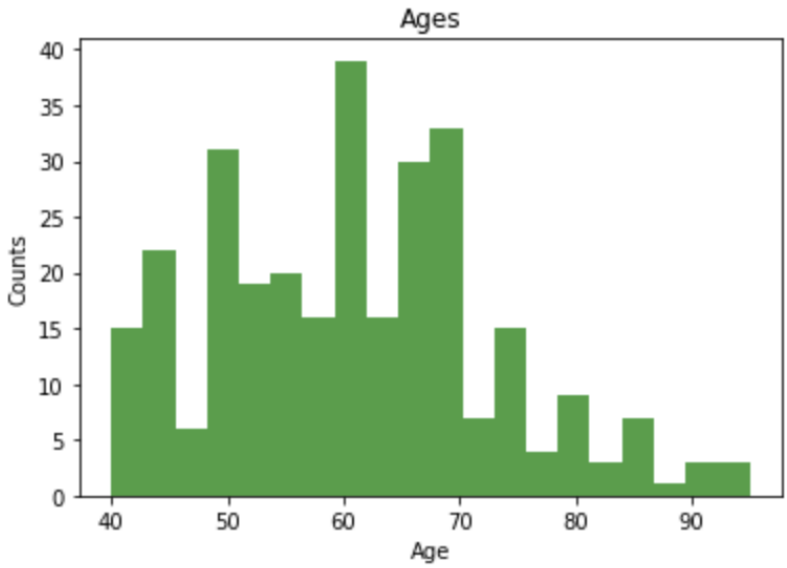


As we would expect, death positively correlate with age. More importantly, serum-creatinine correlates with death, which is what we would expect based on medical findings (Wannamethee et al., 1997). Lastly, we can spot a correlation between ejection fraction and serum-sodium. All of the columns in the heat map will be used in the analysis, except time. The time variable is irrelevant to the analysis in the report. The *time* feature is usually used for the Kaplan-Meier estimator, which is not covered here. With that in mind, let us now look at each column individually[1].

[1]The names of the columns were not altered. They were left in their original form. That's why the first letter is not capitalised

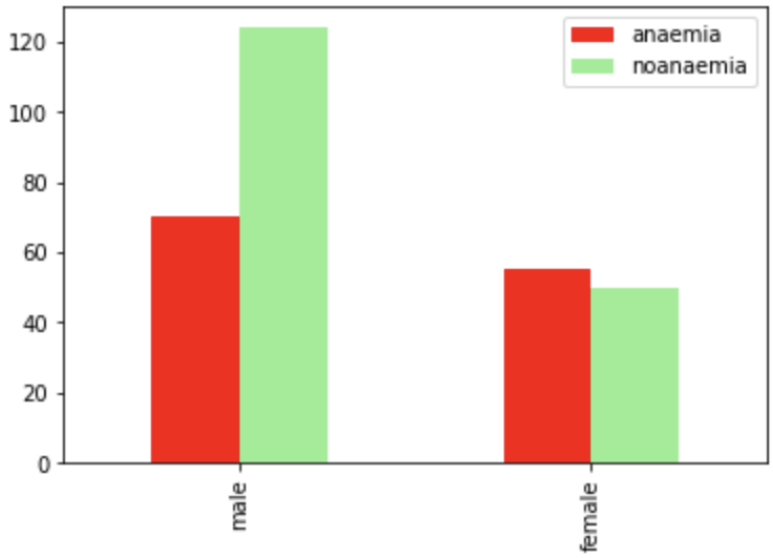
age: As is the case with any medical study, age is a crucial factor. People over the age of 65 are more likely to develop a CVD (NIH, 2022). This seems to jibe well with our data.

Figure 2: Histogram of Age



anaemia: Anaemia is a condition characterised by the lack of healthy red blood cells to carry adequate oxygen to the body's tissues (MayoClinic, 2022). What is of interest for this paper is how anaemia relates to heart failure. Each patient either has anaemia or he/she does not[2]. In the data set, men are less anaemic (only 36 % have anaemia as opposed to 52 % of the women).

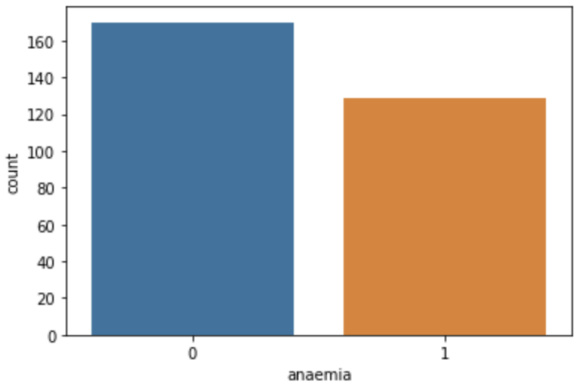
Figure 3: Anaemia in men and women



[2]the various types of anaemia are of no relevance to this analysis

Out of the 299 patients, 125 have anaemia.

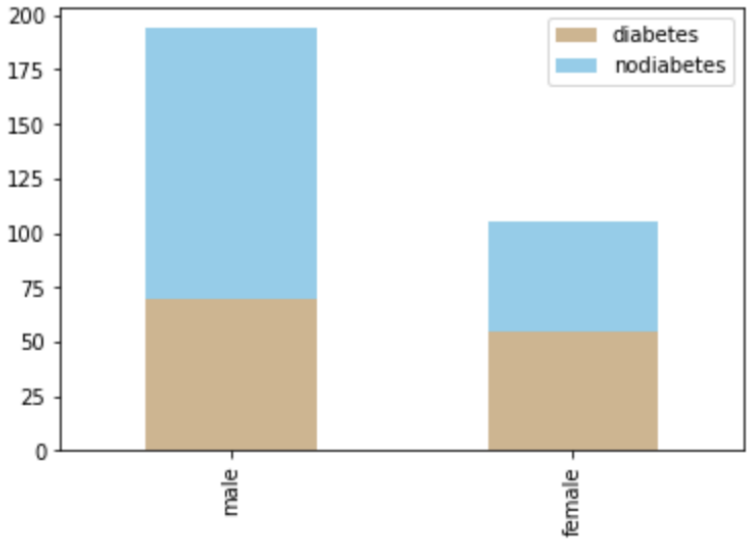
Figure 4: Patients with anaemia



creatinine phosphokinase: Creatine phosphokinase (CPK, or CK) is an enzyme found in the heart, brain, and skeletal muscles. When muscle tissue is damaged, CPK leaks into the blood. Therefore, high levels of CPK usually indicate some sort of stress or injury to the heart or other muscles. 74 % (222 out of 299) of the patients in the data set have CPK levels above 120, which is the upper bound for the normal range (Sinai, 1997). CPK is one of the numerical columns in the data set.

diabetes: People with diabetes are more likely to have heart failure (CDC, 2022). Diabetes comes in types, 1 and 2. However, the distinction is irrelevant to this report. Each person either has diabetes or he/she does not (0 for no and 1 for yes). In our data set men have less diabetes than women (36% compared to 52%).

Figure 5: men and women with diabetes



ejection-fraction:

Ejection fraction is a measurement of the percentage of blood leaving the heart each time it squeezes (contracts). Normal heart's ejection fraction is usually between 50 and 70 percent. Below 40 is low and could be evidence for heart failure. Ejection fraction is a numerical column and ranges from 0 to c.a. 90. From our data set we can infer that most patients in the data set have ejection fraction below 40 which makes sense since we are dealing with people who live with heart failure (Association, 2022a).

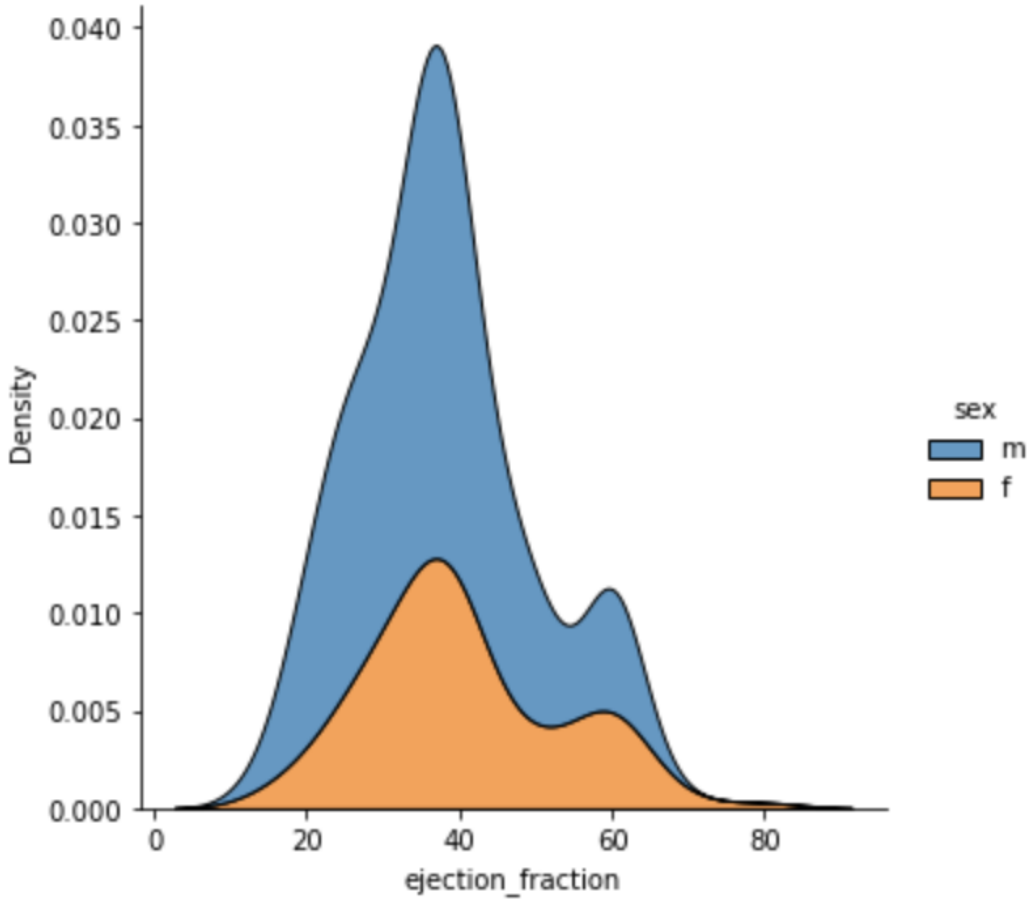


Figure 6: Density graph for ejection fraction

high-blood-pressure:

The thickening and/or stiffening of the heart's walls, as well as narrowing and constriction of blood vessels caused by high blood pressure, are the most common non-cardiac causes of heart failure (Association, 2022b). Each person in the data set either has high blood pressure or does not. Therefore, it is best visualised with a bar plot.

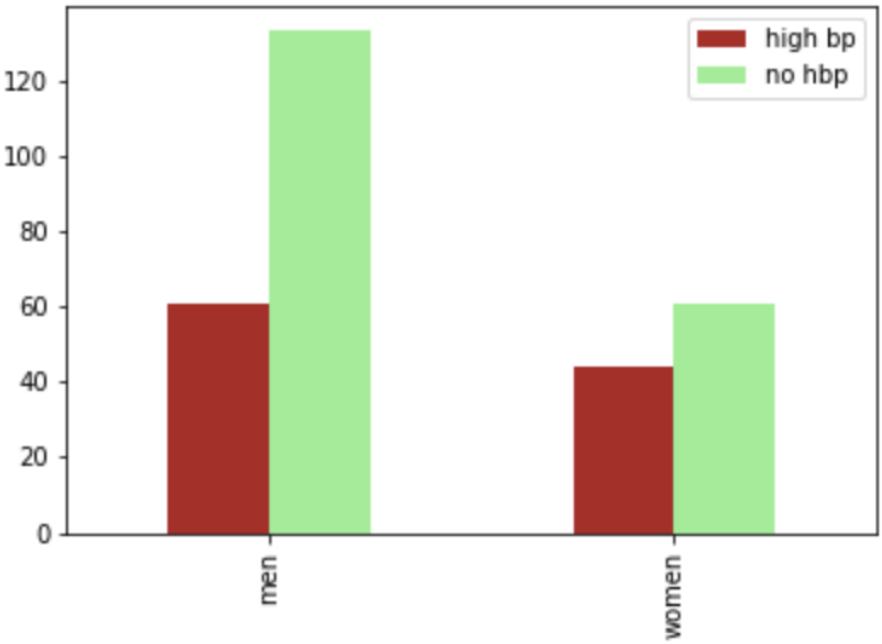


Figure 7: High Blood Pressure according to sex

From the above plot we can infer that in our data, men suffer more from high blood pressure than women.

platelets: Platelets, or thrombocytes, are small, colourless cell fragments in our blood that form clots and stop or prevent bleeding. Most people—in the data set— have 300,000 platelets. If we map put the dead in a graph of (sodium and platelets) we'll quickly realise that people with low sodium and low platelets seem more likely to die. Especially with people who have platelets under 250,000.

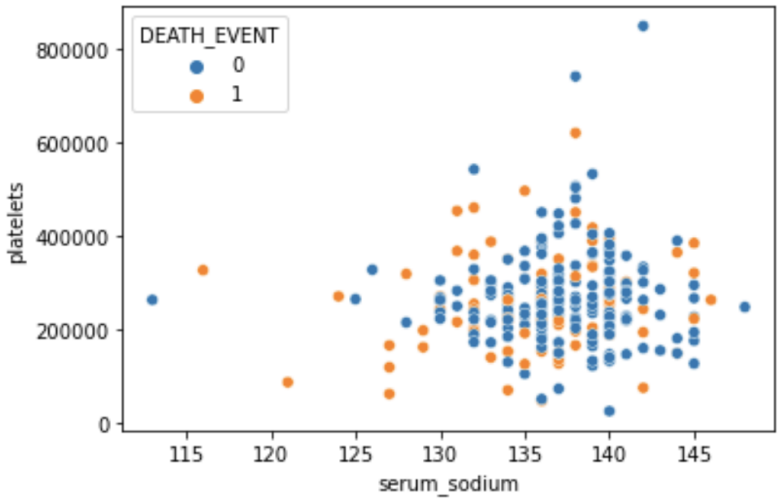


Figure 8: platelets

serum-creatinine: Creatinine is a waste product of the muscles. In a healthy body, the kidneys filter creatinine from the blood and excrete it through the urine. High levels of creatinine can indicate kidney issues. An increase in serum creatinine, also termed worsening renal function, commonly occurs in patients with heart failure and confirms the correlation above between high creatinine and death rates.

serum-sodium: Sodium is a substance cells need to work normally. Sodium helps make sure that your nerves and muscles can work as they should. Normal sodium levels are usually between 136 and 145 millimoles per litre (mmol/L). This is reflected in the data set as well. Plotting a density graph will show that the highest density falls within the normal range.

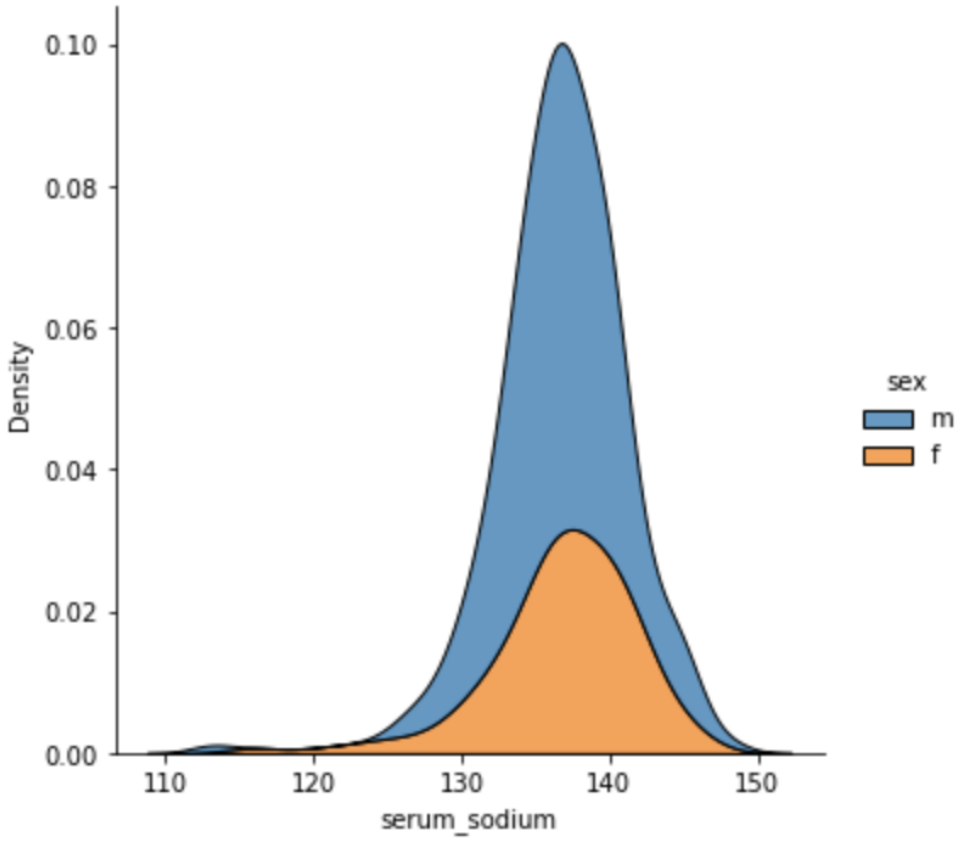


Figure 9: Sodium levels

smoking: At least 1 in 8 adults smoke cigarettes in the UK – that's between 6 and 8 million adults. It's estimated that at least 15,000 deaths in the UK each year from heart and circulatory diseases can be attributed to smoking (BHF, 2022).

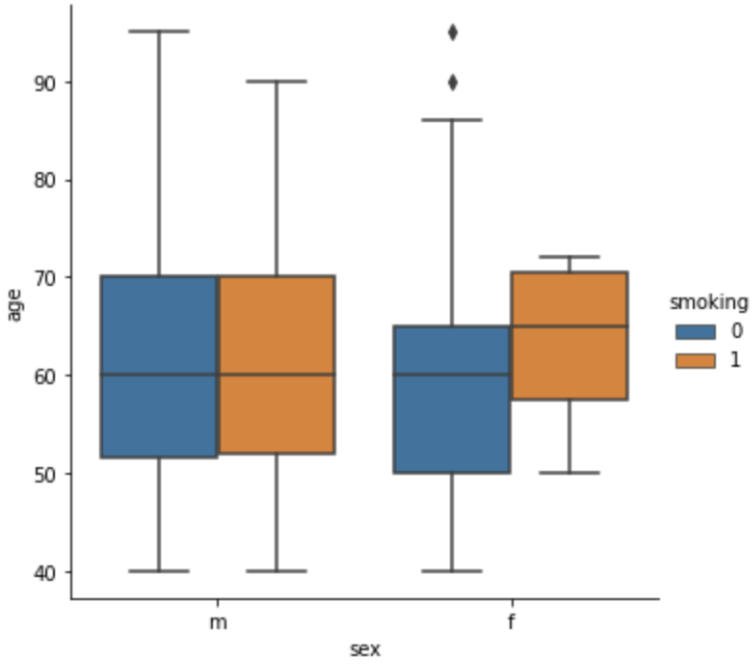


Figure 10: Smoking based on sex

DEATH-EVENT: This column tells us whether a patients had died or not. Most patients in our data set did not die.

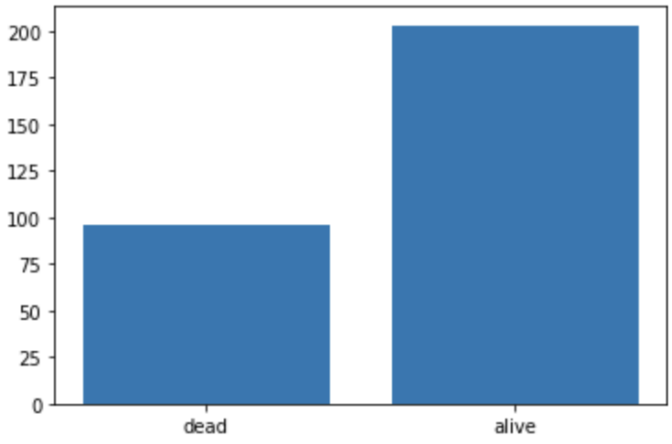


Figure 11: dead vs alive

2.1 Challenges

The data set presents us with several challenges. The first challenge we run into is the representation of *sex*. In the data set males and females are represented with either a 0 or 1, which is quite ambiguous. Consequently, it was changed to 'm' and 'f', though it had to be encoded back (i.e., one-hot encoding) so it could be fed into the algorithms. Another challenge was the *time* feature. Looking the discussion section on Kaggle, *time* was only used for the Kaplan-Meier estimator or dropped altogether (some even argued it was a false feature). Either way, it is not relevant to our analysis since we are not conducting a survival analysis. The last noteworthy challenge was the size of the data set. While the data did not contain any *null* values, it was relatively small. As a result, the models used in this report had to be heavily optimised in order to achieve an acceptable accuracy.

3 Supervised Analysis

In this section several supervised analysis algorithms will be presented to predict whether a person with heart failure will die or not. For simple classification, there are three algorithms to try out: K-nearest neighbours, logistic regression, and decision trees. We will use these algorithms to analyse the data set, and ultimately determine which one of the algorithms performs better and why.

3.1 Challenges

Classification problems present us with several challenges. Firstly, it is essential to empirically determine the test size split. If it is too low, the overall accuracy when predicting the test set will go down. A large test set won't leave us with enough instances to train with. The usual split is 70/30, 80/20 or 75/25. Based on the accuracy results, I chose 75/25 (70 for training, and 25 for testing). Another challenge we are faced with is the size of the data set. A data set of 299 instances is considered (generally) small. Therefore, the plain classification algorithms, i.e., without any optimisations, will yield a relatively low accuracy: c.a. 65%. This can, of course, be improved with data scaling, weight balancing, *etc.*

3.2 Logistic Regression

As described in the previous section, the plain algorithms such as logistic regression have a relatively low accuracy. To increase the accuracy, I scaled the data with *Standardscaler*, which raised the accuracy to c.a. 69%. Moreover, I turned on the *class_weight* variable, which balances the weights associated with each class. This raised the accuracy to 72%. In terms of precision, we realise that the model is more precise when predicting negative cases (i.e., people who didn't die) and this is what we would expect since we have more negative instances than positive ones (see Figure 11).

In this section, we will look at the actual results from the algorithm and visualise the confusion matrix. Firstly, let us take a look at the classification report.

	precision	recall	f1-score	support
alive	0.76	0.77	0.76	44
dead	0.67	0.65	0.66	31
accuracy			0.72	75
macro avg	0.71	0.71	0.71	75
weighted avg	0.72	0.72	0.72	75

Figure 12: Classification Report

Given the size of the data set, 72% accuracy is acceptable (Barkved, 2022). Looking at the confusion matrix we could see that most instances were classified correctly. To visualise the confusion matrix, a heat map between the actual and predicted labels was used, which returns the following matrix:

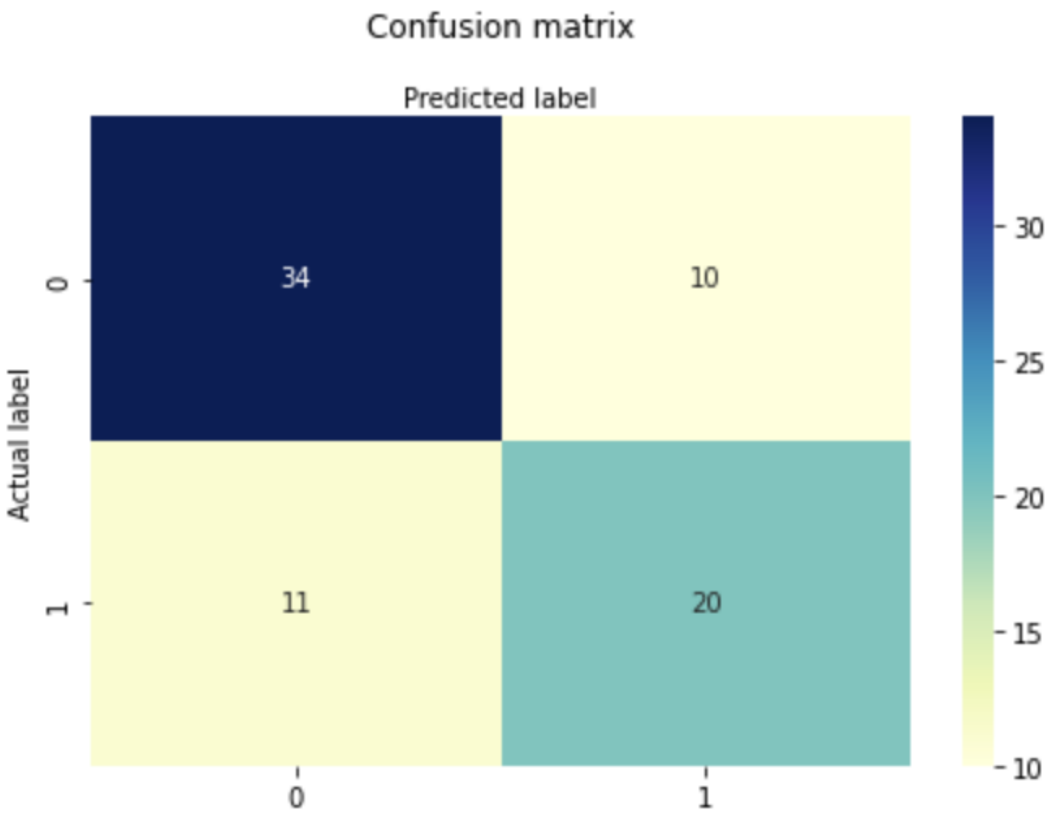


Figure 13: Confusion Matrix

3.3 Decision Tree

Decision trees (DT) are typically prone to over-fitting, especially when the data instances are very similar. However, as we will see, DTs perform well for our data set. First, let us take a look at the *criterion* functions to see which ones fits our data set best. Plotting the confusion matrices we get the following:

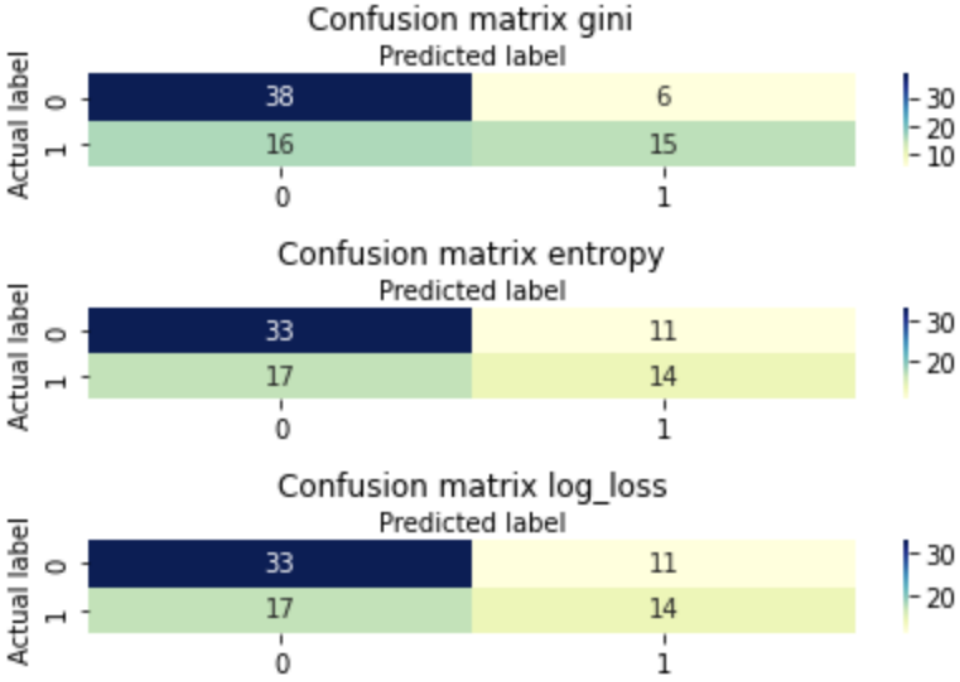


Figure 14: Confusion matrices for DT splitting functions

As we can infer from the plot, the *gini* criterion is better since it leads to less misclassifications.

criterion	accuracy
gini	0.7066666666666667
entropy	0.6266666666666667
log_loss	0.6266666666666667

From the table above it is clear that the *gini* criterion outperforms the others. another optimisation technique we could use is changing the minimum number of leaf samples (see the scikit documentaion). To figure out what the optimal number is, we could look at the accuracy. We try numbers from 1 to 50, which returns the following results:

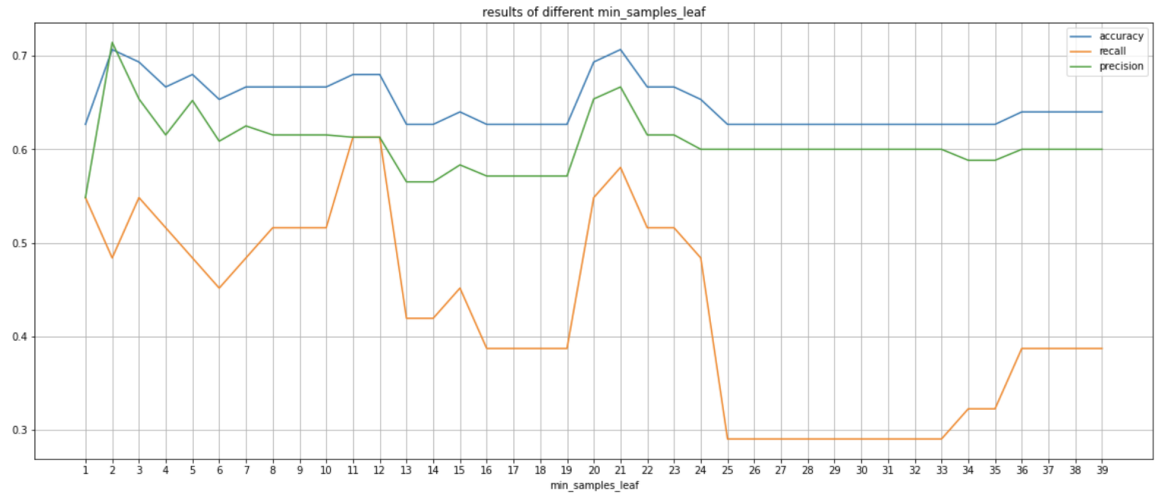


Figure 15: Accuracy, Recall, and Precision (DT)

Based on these results and the results from the previous experiment we will now run the decision tree classifier with the *gini* criterion (i.e., default) and a $min\_samples\_leaf$ equals 2, since we get a the highest accuracy at 2 while also maintaining a high precision and recall.

	precision	recall	f1-score	support
alive	0.70	0.86	0.78	44
dead	0.71	0.48	0.58	31
accuracy			0.71	75
macro avg	0.71	0.67	0.68	75
weighted avg	0.71	0.71	0.69	75

Figure 16: Classification report DT

An accuracy of 71% is acceptable given the the size of our data set, as we will see most algorithms we consider in this report have a similar accuracy score.

3.4 KNN

K-nearest neighbours (KNN) usually performs well on large amounts of data, however, even with a relatively small data set as ours, KNN could be optimised to predict $DEATH\_EVENT$ with a moderate degree of accuracy.

If we run with the default number of neighbours (usually 5), we get an accuracy of 65%. So not much less than Logistic regression and decision trees. However, changing the number of neighbours could

deliver better results. To find the optimal number of neighbours, we take a look at numbers from 1 to 50 and map out their accuracy.

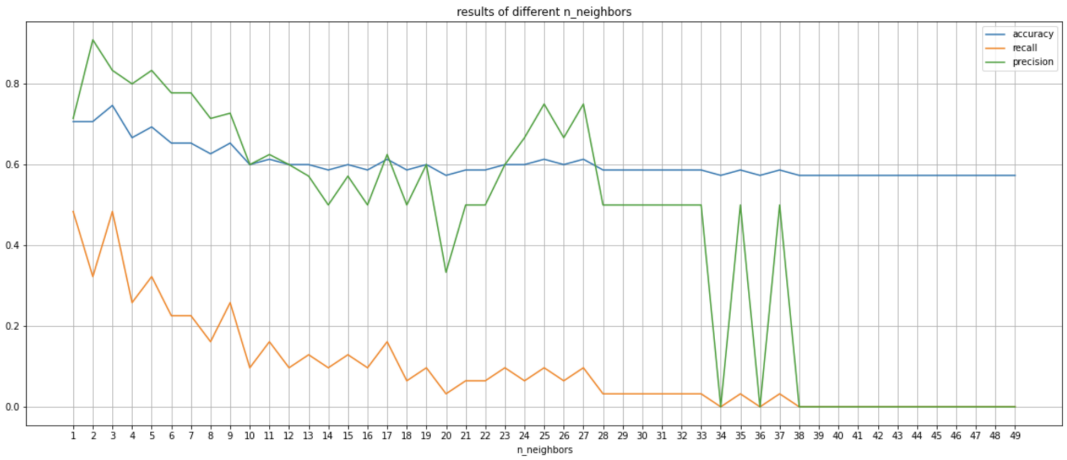


Figure 17: Accuracy, Recall, and Precision (KNN)

Since accuracy is the highest at $n\_neighbours = 3$. Moreover, recall and precision at that point are also relatively high, therefore, we will pick 3 as the number of neighbours. Apart from the number of neighbours, we could change the distance metric (p), which usually gives better results; see (Sharma, 2019). By changing the default Euclidean distance to the Manhattan distance the accuracy goes up from 68% to 75%. The classification report for the parameters described above looks as follows:

	precision	recall	f1-score	support
alive	0.72	0.93	0.81	44
dead	0.83	0.48	0.61	31
accuracy			0.75	75
macro avg	0.78	0.71	0.71	75
weighted avg	0.77	0.75	0.73	75

Figure 18: Classification report KNN

With 75% the KNN outperforms both the logistic regression and the decision tree classifier. In the next chapter we will discuss the differences between KNN, logistic regression and decision trees, and why they perform differently on the heart failure data set.

3.5 Discussion

As we saw in the previous chapter, not all classification algorithm perform the same. Logistic regression (LR) works fairly well on the data set, however, there is some room for optimisation (for example, by using grid search). Most likely, LR performs poorly because the data has non-linearities. In other words, the decision boundaries cannot be captured linearly (see this discussion for a detailed explanation). This becomes more clear when we compare LR to KNN which is not constrained by linear boundaries, and it does perform better on the data set. The decision tree, unlike LR, can work with non-linear data, however, it does not do well on continuous data (see this article). And since the data set has some non-binary columns, the performance of the decision tree is not optimal. The KNN does not suffer from either of the issues mentioned above. For this reason, it reaches a better accuracy. That being said, ideally a deep learning algorithm would be picked for classification[3].

[3] neural networks and deep learning are obviously outside the scope of this report

4 Unsupervised Analysis

The objective of unsupervised analysis is to find similarities in data that could explain the impact of each feature on the patients' health and ultimately, which features are useful in detecting heart failure. Unlike the supervised analysis where several algorithms were presented, here we will focus on one algorithm: the K-means clustering algorithm. K-means was chosen to avoid convergence issues.

4.1 Challenges

The data set does contain many similar patients but it is not always straightforward to predict whether they'll die or not. As a result, many similar instances might be labelled differently. Moreover, as was the case with the supervised analysis, our data set is relatively small, so it is a bit challenging to interpret the results.

4.2 K-Means

The objective of this section is to apply K-Means in order to unlock patterns in the data. This goal is achieved through clustering of the numerical columns, but it is not clear without visualisation. To visualise the data I mapped the labels obtained from K-Means for every two columns in the data set. However, the data did not always form clear boundaries across the features. Only plots that involved *serum_sodium* and *serum_creatinine* revealed two distinct clusters. The plots can all be seen in the accompanying code. But before looking at the plots, we will briefly discuss the quality of the clustering using three metrics: the silhouette score, the homogeneity score, and the completeness score.

```
silhouette:  0.1911067435014852
homogenity:  0.10573955498516314
completeness:  0.10800579679793926
```

Figure 19: scores

The silhouette score is positive and both the completeness and homogeneity are between 0 and 1. These are all modest ranges for these scores, which means that our K-Means solution is acceptable. Let us now direct our attention to the clusters themselves. As discussed above, we will limit ourselves to only two columns (*serum_sodium* and *serum_creatinine*) that allow the data to cluster nicely and thus assist us in finding patterns.

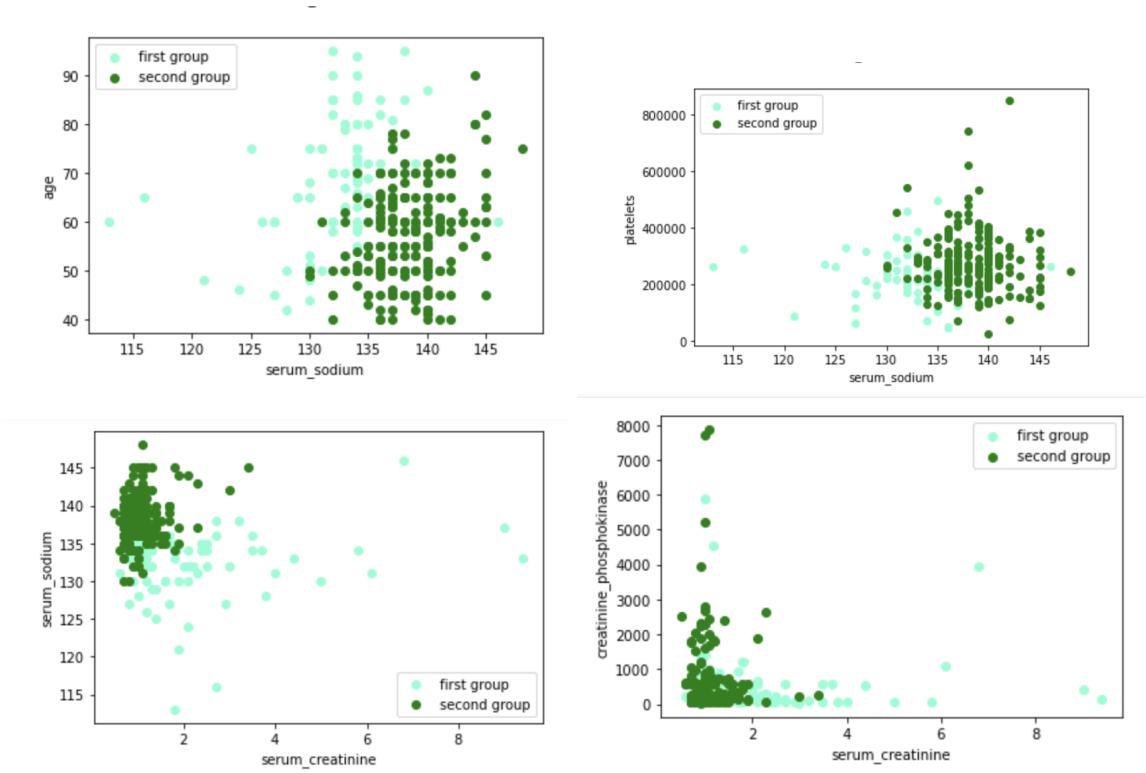


Figure 20: Cluster plots 1

As the plots show, there are two clusters, with minor overlaps. What is not clear is what these clusters are. The blue cluster (first group) represents people with high *serum_creatinine* and low *serum_sodium*, both of which increase the likelihood of dying in heart patients, as we explained in chapter 2. This is further corroborated by the number of deaths in each cluster. The first cluster (group one in the plot) has a relatively high percentage of deaths, compared to plot two where the death count is lower:

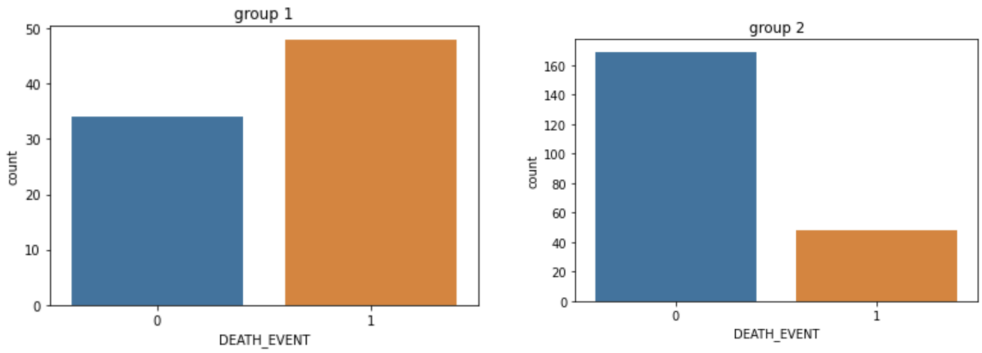


Figure 21: Dead vs Alive (group one & two)

Given our data, we had a *prima facie* reason to select two as the number of clusters. But is 2 mathematically the best number? To determine that we look at 3 metrics: silhouette score, homogeneity and completeness. For our purposes, the silhouette score holds more weight that the other two.[4]

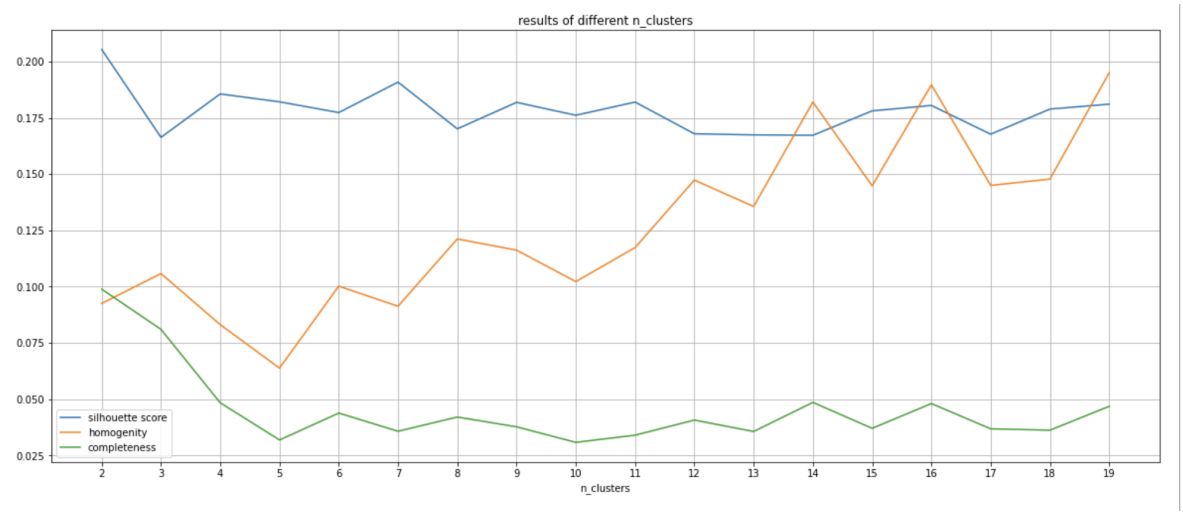


Figure 22: K-Means metrics

As the plot shows, the silhouette score is at its highest when $n_clusters = 2$, which proves our initial intuition.

4.3 Discussion

The difference results from the high low levels of sodium and high levels of creatinine. So it seems that one cluster represents people with symptoms that are more likely to lead to death while the other cluster represents the opposite, i.e., people with symptoms that are less likely to cause death.

5 Conclusion and Reflection

In retrospect, I should have chosen a larger data set with more instances. The data set at hand could, of course, be used to predict death, but the basic classification algorithms do not reach a high accuracy. In most examples on Kaggle, people used neural networks for classification and the results they achieved were invariably high (90%). But, despite that, I am not dissatisfied with my choice. The analysis in the report has revealed a few patterns between the columns, most of which are in line with medical findings.

[4]Homogeneity will eventually be extremely high since clusters will only contain persons that are almost identical. For more details see this article

References

A. H. Association. Ejection fraction heart failure measurement. 2022a. URL https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement.

A. H. Association. How high blood pressure can lead to heart failure. 2022b. URL https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-heart-failure.

K. Barkved. How to know if your machine learning model has good performance. 2022. URL https://www.obviously.ai/post/machine-learning-model-performance.

BHF. Facts and figures. 2022. URL https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures.

CDC. Diabetes and your heart. 2022. URL https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html.

MayoClinic. Anemia. 2022. URL https://www.mayoclinic.org/diseases-conditions/anemia/symptoms-causes/syc-20351360.

NHS. Cardiovascular disease. 2022.

NIH. Heart health and aging. 2022. URL https://www.nia.nih.gov/health/heart-health-and-aging.

N. Sharma. Importance of distance metrics in machine learning modelling. 2019. URL https://towardsdatascience.com/importance-of-distance-metrics-in-machine-learning-modelling-e51395ffe60d.

M. Sinai. Creatine phosphokinase test. 1997. URL https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test.

S. G. Wannamethee, A. G. Shaper, and I. J. Perry. Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke. 1997. URL https://pubmed.ncbi.nlm.nih.gov/9056611/.