Results, Discussion and Conclusions

Jurrien de Jong

4/10/2021

Intro

The dataset analyzed for this research contains the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All patients had left ventricular systolic dysfunction which puts them on a higher risk of death, this needs to be kept in mind when looking at the data.

Heart failure is quite common, and thus affects a lot of people each year. The condition is the leading cause of hospitalization in people over age 65. One solution might be to prevent heart failure from happening by examining large datasets full of data which are known to relate (closely) to heart disease and/or heart failure. This research is trying to replicate this solution with a supplied dataset.

The question this research is aiming to give an answer to is: "Can a death event be predicted when blood serum and age data is given using machine learning techniques?"

Cleaning data

Before the data can be used by a Machine Learning algorithm, the data needs to be cleaned:

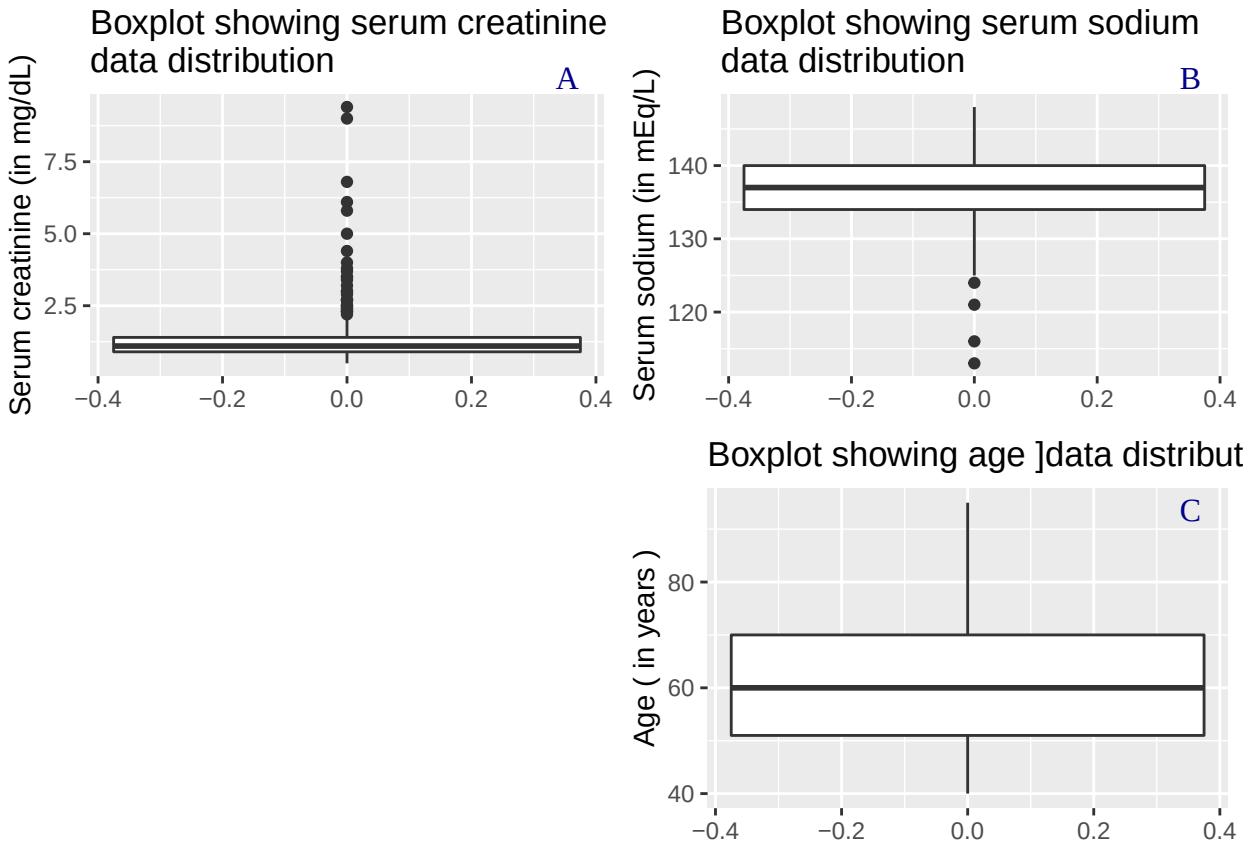
The datatypes, readability and reproducibility are very important in an EDA. This is why the function 'mutate' comes in handy. The 0,1 structure will be replaced by "False/True" to make the readability better and also increase the reproducibility. Finally, tibble will be used to give the 'head' of the data is table. As seen in the table below, the data has labels and no missing values at all.

```
## # A tibble: 299 x 13
##      age anaemia creatinine_phosphokinase diabetes ejection_fracti~ high_blood_pres~
##    <int> <fct>                      <int> <fct>                <int> <fct>
##  1    75 False                        582 False                   20 True
##  2    55 False                       7861 False                   38 False
##  3    65 False                        146 False                   20 False
##  4    50 True                         111 False                   20 False
##  5    65 True                         160 True                    20 False
##  6    90 True                          47 False                   40 True
##  7    75 True                         246 False                   15 False
##  8    60 True                         315 True                    60 False
##  9    65 False                        157 False                   65 False
## 10    80 True                         123 False                   35 True
## # ... with 289 more rows, and 7 more variables: platelets <dbl>,
## #   serum_creatinine <dbl>, serum_sodium <int>, sex <fct>, smoking <fct>,
## #   time <int>, DEATH_EVENT <fct>
```

Results

First, let's try to answer the research question by creating plots. An EDA process contains plenty of different methods of exploring the data. In this section those results will be shown.

It is very important to check if the data contains major outliers or maybe even typos. This could be dramatic to the results/conclusions of the research. By creating boxplots, outliers can be visible outside the 'box' as dots:



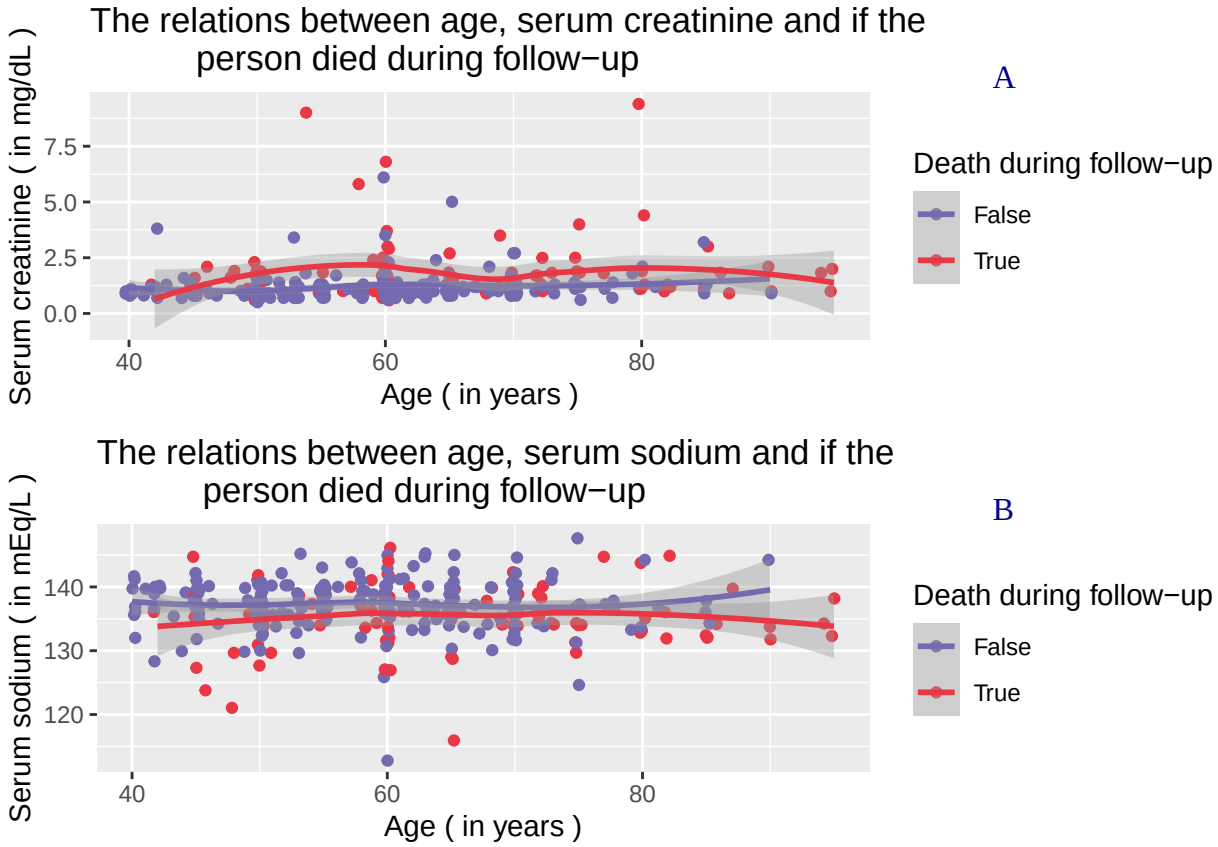
As seen in figure A above, the data contains many outliers. A normal serum creatinine value ranges between 0.59 and 1.35 mg/dL while the data contains values which are close to 10. Taken into account that the patients who have taken part in this research have underlying heart/renal conditions, these values are acceptable. This trend is also visible in figure B, while the range of the outliers is much smaller than figure A. The normal serum sodium level ranges between 135 and 145 milliequivalents per liter (mEq/L). As seen in figure B some outliers have much lower sodium values, this could be quite harmful to the heart because decreased sodium can cause muscle dysfunction and ultimately heart failure. The age boxplot seen in figure C shows normally distributed data. This can be supported by the fact that people get examined more as they get older because key body functions, like muscle- and renal function, will on average deteriorate over time.

Relationship between attributes

Dotplot

It is important to get to know what kind of relationship there is between the variables. Let's first 'cut out' the part of the data which will be used for the research. The serum and age data will be used, which are from columns 1,8 and 9 because the age, serum creatinine and serum sodium values have a great impact on

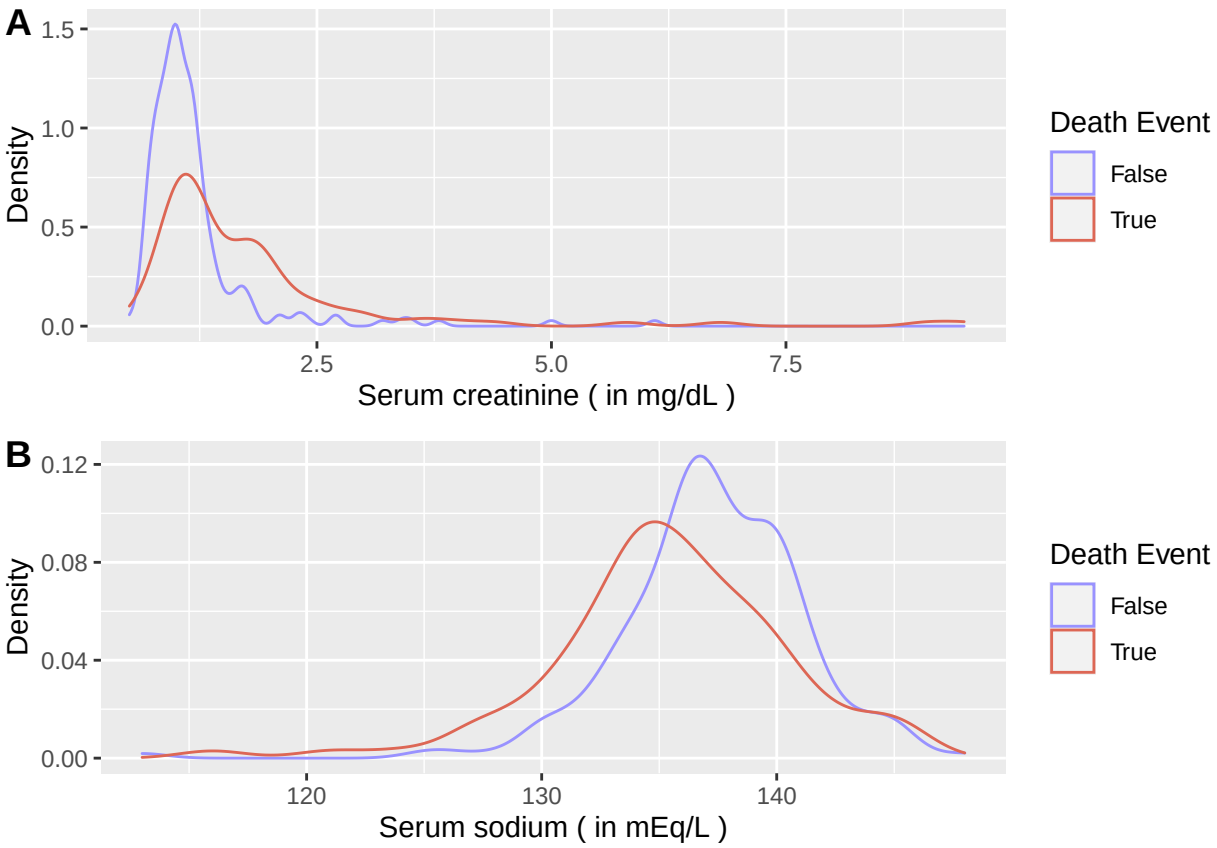
heart disease and renal dysfunction, as seen in "Survival analysis of heart failure patients: a case study". Below the relations are plotted in graphs:



As mentioned before, a higher serum creatinine count or a lower serum sodium count than normal can result in heart failure. The literature found was correct as seen in the plot above. Figure A shows an increased death chance if the creatinine value is higher than 0.125 mg/dL and figure B shows an increased death chance if the sodium drops below 135 mEq/L.

Density

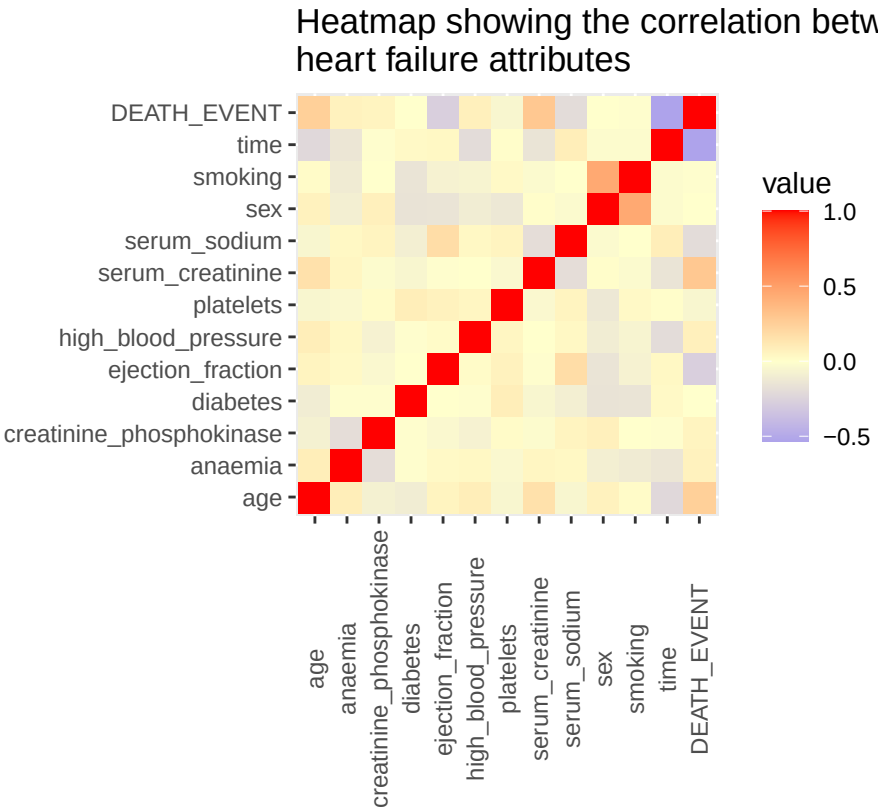
A density plot can be used to help display where values are concentrated over the interval. For this instance the interval is the amount of creatinine or sodium in the blood serum. To illustrate which concentration seems fatal, two density plots have been created, and colored based on the outcome of the follow-up period:



Just like the dotplot before, the density plot shows that a lower serum sodium value seems more fatal, as is a slightly higher serum creatinine value. The difference in frequency of figure B does seem more convincing than figure A.

Correlation

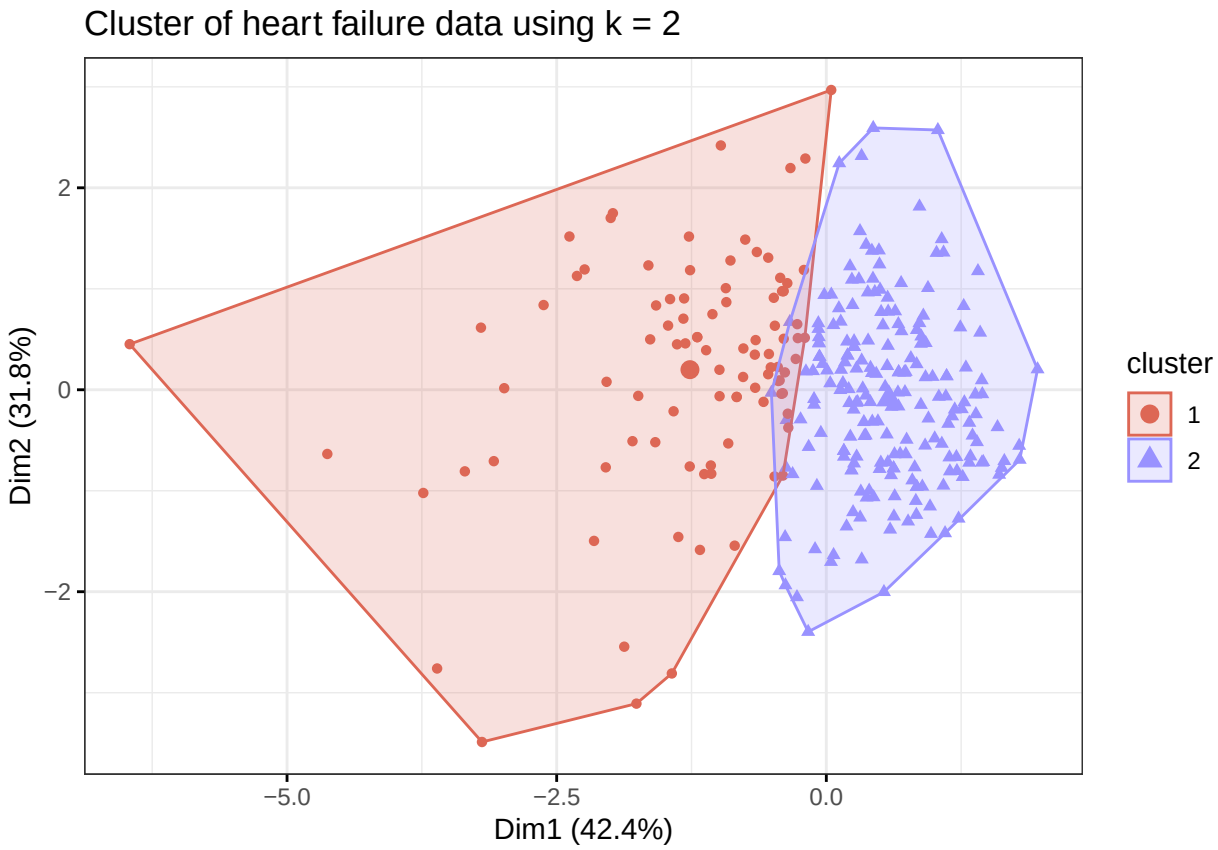
The correlation is an extremely important factor in machine learning because ML algorithms assume that all attributes are independent. When looking at the heatmap below, 3 attributes stand out. At first, the follow-up time has a very negative correlation to the class attribute. The serum sodium has a somewhat negative correlation while the serum creatinine shows that there is a positive correlation between it and death event. So concluded, both values have differing correlation with the death event attribute which means they are independent.



A heatmap visualizing the correlation between all heart value attributes. The red in the diagonal can be ignored because it represents a correlation of one variable instead of two. When the color of a tile is close to white there is no to near minimal correlation between attributes. If the tile is blue colored, there is a negative correlation present, while a red color indicates a positive correlation.

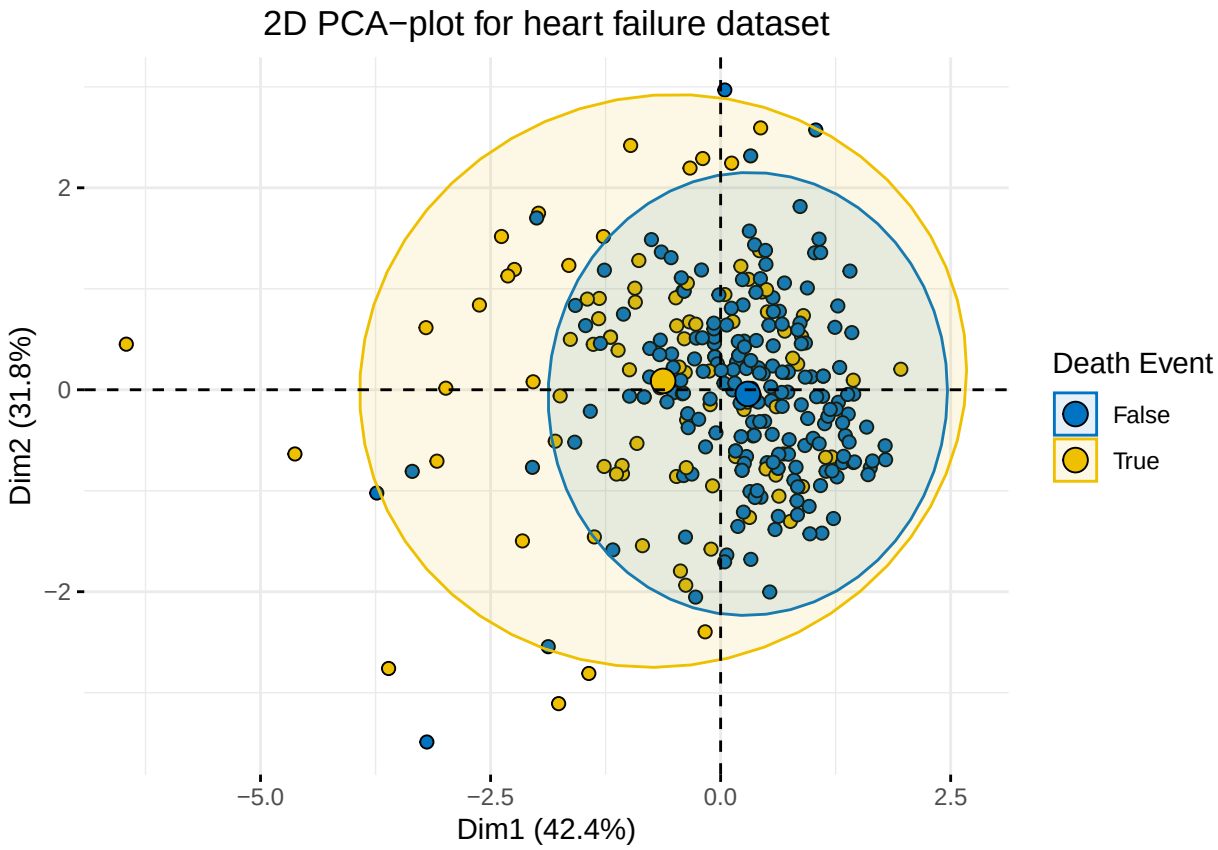
Clustering

Clustering can be a great tool to give a better understanding of the data by discovering patterns. Below, a k-means cluster of 2 will be performed. A k of 2 has been chosen because the death event attribute has 2^1 possibilities.



In the plot above there can clearly be seen that two independent clusters have been formed with the given data. The most right cluster in blue represents the patients who survived while the red cluster shows the patients which died during follow-up. There is however a small section visible where the two clusters overlay, which means that it is not certain what the result will be for patients inside that range.

Next a PCA cluster plot will be created. It is quite similar to the k-means cluster, so the result will most likely be the same. It shows clusters of samples based on their similarity, which gives a good indication for independent values. PCA is mostly used to decrease the amount of dimensions of dataset, but in this instance that is not the reason.



The PCA-plot seems to give the same result as the k-means cluster. A 'death' cluster seems to emerge when data points go to the left. There is however overlap between the false and true clusters, so the probability to predict a correct death event might not be very high.

Discussion

Most variables seem independable of each other, as seen in the heatmap created above. This, however, is a good sign because ML algorithms need no correlated values. One point of critique might be that also data of healthy patients must be taken to look at the differences between the data of healthy people and the patients with left ventricular systolic dysfunction. Some attributes which contribute to heart disease also need to be looked at in future research. Some key attributes are: daily physical activity, cholesterol levels, patient weight / BMI and most importantly for a bioinformatician, heredity. Because a lot of heart failure cases come from people who have heart disease in their family, possible heart disease needs to be caught early.

Conclusion

A death event is pretty hard to predict because so many variables have impact on the health of people. So to further improve this research, a whole lot of extra attributes need to be taken into consideration. A few attributes of heart data do not give a big picture of the whole situation. Also, the group size of 'only' 299 people is quite small to draw conclusions, so to improve, more patients need to be examined.