# **Assignment 2:Data Modelling and Presentation**

**Prediction of Heart Failure Survival**

**19th May 2021**

Team members details

Student ID: s3841545

Student Name: Sai Ramishetty

Student email: s3841545@student.rmit.edu.au

Student ID: s3815294

Student Name: Manav Makkar

Student email: s3815294@student.rmit.edu.au

Lecturer:

Dr Yongli Ren

|  |
| --- |
| We certify that this is all our own original work. If we took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in our submission. We will show we agree to this honor code by typing "Yes": *Yes*. |

# Table of contents

1. **Abstract/Executive Summary**………………………………………………………….

2. **Introduction**…………………………………………………………………………………….

3.**Methodology**…………………………………………………………………………………….

4.**Results**………………………………………………………………………………………………

5.**Discussion**………………………………………………………………………………………….

6.**Conclusions**………………………………………………………………………………………..

7.**References**………………………………………………………………………………………….

**Abstract/Executive Summary**

Heart failure occurs when the blood pumped by the heart is not enough to satisfy the needs of a person. In North America, approximately 379,800 people have died in 2018 due to heart failure. There are several factors that can show impact on how our heart operates. These factors can be helpful in identifying potentially high risked heart failure patients. Machine learning is a tool that can be utilised to predict whether a patient can survive a heart failure or not. We will use two features, ejection fraction and serum creatinine, to predict whether a patient will survive a heart failure or not. However, we will also build models on the whole dataset and compare the accuracy to see if ejection fraction and serum creatinine are enough to predict survival or death. For this study, we will be using two prediction classification models, kNN and Decision Tree, and determine which model will be used for making accurate predictions.

**Introduction**

Heart failure is a deadly condition that accounts for about one in 50 deaths in Australia. Despite the fact that twice the men experience heart failure than females, it is reported that females are more vulnerable to die from heart failures([Key Statistics: Heart Failure | The Heart Foundation](https://www.heartfoundation.org.au/activities-finding-or-opinion/key-statistics-heart-failure)). On the whole, around 30,000 Australians with heart failures are diagnosed on average every year. There are many symptoms associated with heart failures, such as irregular heartbeat, swelling in legs, dizziness, etc([Heart failure - treatment, causes, living with it and more | healthdirect](https://www.healthdirect.gov.au/heart-failure)). However, there are other health factors that contribute to those symptoms. In fact, those factors are the ones which can determine whether a patient can survive a heart failure or not. Through the reports above, we have gone through the general facts, but for medical treatment, it would be beneficial for the doctors to know whether there are particular health factors, such as ejection fraction and serum creatinine, which are enough to tell whether a patient will overcome a heart failure or not.

**Methodology**

***Data***

The dataset being used is from [UCI Machine Learning Repository: Heart failure clinical records Data Set](https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records) . Originally, the dataset was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan). However,  Davide Chicco (Krembil Research Institute, Toronto, Canada) elaborated on the dataset and sent it to University of California Irvine Machine Learning Repository. The data consists of 299 patients in total. The survival is indicated by the variable DEATH\_EVENT. The number 0 means the patient survived, while 1 means the patient died.

***Data Analysis tools***

The Integrated Development Environment selected for the study is Jupyter Notebook. Packages such as pandas, matplotlib, numpy, seaborn, sklearn and math were used to conduct the analysis. The first four packages were mainly used in Data preparation and Data exploration, while sklearn was used for Data Modelling.

***Classification Models***

**KNN(k-Nearest Neighbours Classifier)**

The KNN classifier classifies a data point based on how its neighbour is classified. The letter k represents the number of neighbours near to the new data point. Below is an image displaying how KNN classification process works.Diagram

Description automatically generated

*The star in this image has many neighbours around it, however, when classifying it, the k value determines whether it is class A or class B. When k=3, the star is classified as class B, but, when k=6, it is classified as class A.*

KNN classifier is excellent to use for this study. This dataset has labelled data for the target feature DEATH\_EVENT, which works well for KNN model. Another thing to note is our dataset size is 299, which fairly small.

The perfect value of k is dependent on two things

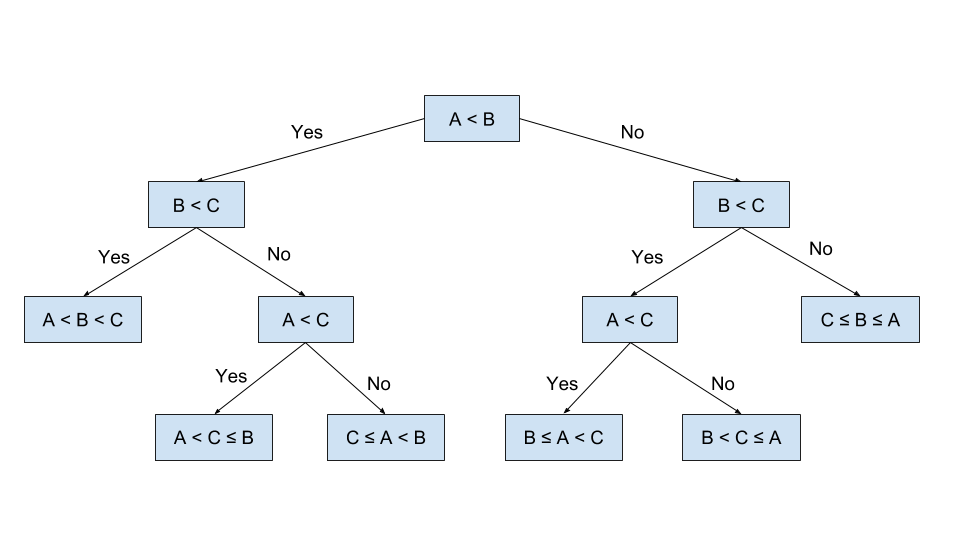
* √(size of y test)[square root of y test size].
* The k value has to be odd as confusion can be avoided

This model will be trained using K-folds validation and Leave-1 out. The parameter n\_splits for K folds will be 5 since our dataset is relatively small and it is a default value. On the other hand, Leave-1 out validation will check the absolute error of predicting actual observations. Both these validations will give an idea on whether the model is able to accurately make predictions. Finally, the model will be broken into two, standardised and non-standardised. This is to make sure we analyse how the presence of outliers affects the accuracy of the KNN model making predictions.

**Decision Tree Classifier**

Decision trees are the one of the commonly used machine learning models. They are non-parametric models which learn by recursively split the predictor space according to the best feature till the tree reaches a reserved depth. Ahead the subsets contain the elements of one class only. The best thing about decision tree model is that if it is used to discriminate samples, it allows us to get the best possible according to the measure (Gini index).

where 𝑝 is the ratio between number of samples of class 𝑗j and total number of samples.



*Here it can be clearly seen for decision tree model works on simple conditions as per dataset.*

The decision trees are simple to understand and can handle both numeric and categorical data. In addition, they require less effort for data preparation and are not affected by any non-linear relationships between the parameters.

**Results**

***Data preparation***

During this stage, we aimed to go through the dataset and filter out any unwanted errors or fill missing values. The dataset consisted of 299 records and 13 columns/features. The features include age, anaemia, creatinine phosphokinase, etc. The target feature of the dataset is Death event and this variable shows whether a patient survived or died. Table 1 will present data of numerical features and Table 2 will do the same for categorical features.

**Table 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **SD** | **min** | **Q1** | **median** | **Q3** | **max** |
| Age | 60.83 | 11.89 | 40 | 51 | 60 | 70 | 95 |
| Creatinine Phosphokinase | 581.84 | 970.29 | 23 | 116.5 | 250 | 582 | 7861 |
| Ejection Fraction | 38.08 | 11.83 | 14 | 30 | 38 | 45 | 80 |
| Platelets | 263358.03 | 97804.24 | 25100 | 212500 | 262000 | 303500 | 850000 |
| Serum Creatinine | 1.39 | 1.03 | 0.5 | 0.9 | 1.1 | 1.4 | 9.4 |
| Serum Sodium | 136.63 | 4.41 | 113 | 134 | 137 | 140 | 148 |
| Time | 130.26 | 77.61 | 4 | 73 | 115 | 203 | 285 |

Except for Age and Time, it appeared there were invalid errors within the data of these features. Though our focus is ejection fraction and serum creatinine, we want to make sure the whole data is clean. Moving on, we were surprised to notice the lowest number of platelets being 25100 as it is generally way higher than that. Thus, we went through each feature’s data in detail by extracting the outliers present. Then, we did some research regarding the average range of each feature’s count. It was concluded that those patients who were outliers were serious cases and they can not be excluded from the dataset as they can provide further insight to our study. Other errors, such as missing values, were not noticed in these features’ data.

**Table 2**

|  |  |
| --- | --- |
| Anaemia | Yes:43.14% No:56.86% |
| Diabetes | Yes:41.81% No:58.19% |
| High Blood Pressure | Yes:35.12% No:64.88% |
| Gender | Male:64.88% Female:35.12% |
| Smoking | Yes:32.11% No:67.89% |
| Death Event | Yes: 32.11% No:67.89% |

As it can be noticed, majority of the patients in the data set do not possess any serious health problem or have bad health habits, therefore, most of them are able to survive a heart failure. But, through data exploration, we can identify whether ejection fraction and serum creatinine have a strong relationship with death event. There were no missing values or any other errors during the preparation of this data.

**Data Exploration**

Let us look at individual columns first. The important features to look at are ejection fraction and serum creatinine. Later, two box plots will be presented to explore relationship between two features.

Chart, histogram

Description automatically generated

As it can be noticed, this histogram is approximately symmetrical since the two sides that come before and after the peak are almost the same. It also indicates the majority of the patients have ejection fraction levels between 35-40%.

*Chart, histogram

Description automatically generated*

This histogram clearly presents that the data is rightly skewed with some outliers. Therefore, majority of the patients have serum creatinine levels between 0 and 2 mg/dl. The outliers must indicate that patients have serious health problems, therefore, this will be useful for our data model.

Chart, box and whisker chart

Description automatically generated

The hypothesis we believed is that if a patient has more ejection fraction, then he or she will have a greater chance of surviving. However, the box plot on the left provided surprising results which were not expected. Of course, it was clear that a patient with higher ejection fraction survived(median of survived is close to 40% while median of died is just above 30%), but it must be noted that the patient who had close to 60% ejection fraction died from heart failure. Overall, there is a clear relationship between the death event and ejection fraction.

Chart, box and whisker chart

Description automatically generatedThe hypothesis regarding this relationship was if a patient had higher amount of serum creatinine, then it is unlikely to survive a heart failure. The box plot on the left does prove this, but like the previous box plot, it throws surprises. The outliers on the survived box plot shows that patients with higher serum creatinine somehow managed to survive. Such outliers can be beneficial for our model as it can make it broader, which can help in giving better predictions. Overall, there is an existing relationship between Death event and Serum creatinine.

Both box plots have outliers, which can be a problem with the KNN model, therefore, we can create two different models(standardised and unstandardised). Along with this, other features will also be used and the overall accuracy difference will be compared.

***Classification Models***

The first type of model built is KNN classifier. We have taken 4 scenarios in total and these include

* Total dataset with outliers
* Total dataset standardised(no effect of outliers)
* Ejection Fraction, Serum creatinine as independent features with outliers
* Ejection Fraction, Serum creatinine as independent features standardised(no effect of outliers)

We started off by creating the data and target variables. Afterwards, it was decided to do the parameter tuning for K value (number of neighbours near data point) only while others would be done manually. The reason for this is because the remaining parameters weights and p have only two values each(uniform, distance for weights and 1,2 for p) while the K number can have values from 1 to number of samples, though a lower K value can improve the accuracy of making predictions. The way to select the K value was simple. There are two conditions to it. The first one is it has to be an odd number(this was outlined in the Methodology) while the second one is just square rooting the number of target or data points in the dataset. This was done several times throughout the scenarios as we wanted to maintain consistency throughout the construction of each model.

After this, the focus was put on training the models. We used K-folds and Leave-1 out validation to check the accuracy of our data models. While K-folds tells the overall accuracy in making predictions, Leave-1 out validation gives the mean error of predicting observations. Table 3 will display the scores for each scenario.

**Table 3**

|  |  |  |
| --- | --- | --- |
| Scenario | K-folds validation score | Leave-1 out validation score |
| Total dataset with outliers | 83.29% | 0.30 |
| Total dataset standardised | 82.17% | 0.30 |
| Only Ejection Fraction, Serum Creatinine with outliers | 74.87% | 0.37 |
| Only Ejection Fraction, Serum Creatinine standardised | 73.75% | 0.37 |

As it can be seen, there was a decline in the K-folds validation score when we only used Ejection Fraction and Serum Creatinine. The margin of error increased, however, the amount it increased is not that significant to come to immediate conclusion. It was also observed that outliers do not really show any affect on the predictions as the scores with or without standardisation are almost similar. Since the models were trained, it was ready to make predictions. Table 4 will provide the overall performance for each case.

**Table 4**

|  |  |
| --- | --- |
| Scenario | Test score |
| Total dataset with outliers | 62.5% |
| Total dataset standardised | 77.5% |
| Only Ejection Fraction, Serum Creatinine with outliers | 74.17% |
| Only Ejection Fraction, Serum Creatinine standardised | 77.5% |

The scores attained here were dependent on the parameter tuning. Interestingly, the suitable k value for the first two scenarios was 11, while it was 9 for the latter two. The other two parameters, weights and p, were kept consistent with uniforms and 2 throughout the model building. This highlights the fact that these parameters can rather be manually tuned while K value should be tuned automatically. Interesting aspect from these results was Ejection Fraction and Serum Creatinine achieving the same score as the total dataset when both are standardised. The big surprise was Ejection Fraction and Serum Creatinine recording a higher score than total dataset when both had outliers present.