**COSC2670 – Practical Data Science**

**Assignment 2**

Title: Can the survival of patients with heart failure be predicted?

Affiliations: RMIT University

Date of Report: 19/05/2022

Student Name: Puneet Kaur Grewal

Student ID: s3900991

Student Email: [s3900991@student.rmit.edu.au](mailto:s3900991@student.rmit.edu.au)

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 honor code by typing "Yes": Yes

INCLUDE A READ ME FILE FOR THE CODE.

# Table of Contents

[1. Table of Contents 2](#_Toc104068616)

[2. Abstract 2](#_Toc104068617)

[3. Introduction 2](#_Toc104068618)

[4. Methodology 3](#_Toc104068619)

[5. Results 4](#_Toc104068620)

[5.1. Exploring Clinical Features 4](#_Toc104068621)

[5.2. K-Nearest Neighbours 6](#_Toc104068622)

[5.1. Decision Tree 7](#_Toc104068623)

[6. Discussion 8](#_Toc104068624)

[7. Conclusion 8](#_Toc104068625)

[8. References 8](#_Toc104068626)

# Abstract

\*\* Make changes.

The aim of this report was to investigate if the survival of patients with heart failure could be predicted using different features. The data used in this study was obtained from a dataset of medical records of patients with heart failure. The K-Nearest Neighbours and Decision Tree classification techniques were used to create models to help predict the survival of patients with heart failure. The results of the models show that it is possible to be able to predict the survival rate using data from the medical records of patients. The Decision Tree models provides a higher accuracy to determine the chances of survival of patients. The report concludes that medical professionals are able to get a better understanding of their patients’ chances of survival using data from medical records. It is recommended that medical professionals utilise the Decision Tree model when reading through medical records of patients with heart failure to gain a greater understanding of their patient’s risk death.

If medical professionals are able to look at medical records are see potentially threatening data according to the model, they will be able to monitor and treat patients more carefully, hopefully leading to potential patient recovery.

Help to forecast heart-failure related events.

# Introduction

\*\* have in text referencing throughout this because there is no way I know this information off the top of my very intellectual mind. Also, plagiarism check – especially that long sentence where I list the causes.

Heart failure is a type of cardiovascular disease occurring when the heart does not pump enough blood around the body. The risk of heart failure can be increased due factors such as aging, family history, unhealthy lifestyle habits (such as a poor diet, smoking, use of illegal drugs, alcohol abuse and lack of exercise), having heart of blood vessel conditions, lung disease, infections such as HIV. (NIH, n.d.). Cardiovascular diseases were responsible for approximately 17.9 million deaths in 2019 and was the cause of 32% of all deaths around the world (WHO, 2021). The high death rate surrounded heart related diseases due to the heart being such a vital organ highlights why it is so important for medical professionals to find a way to predict heart failure is patients. This report will use classification techniques such as K-Nearest Neighbours and a Decision Tree to create a model to use data from medical records to predict the chances of survival of patients with heart failure.

# Methodology

The research outlined in this report used a dataset from 2015 which information collected for 299 patients with heart failure using their medical records. The data was collected from medical records of patients at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad from the province of Punjab in Pakistan. This was between the months of April and December. The clinical features collected from medical records can be seen in *Table 1*.

|  |  |
| --- | --- |
| **Clinical Feature** | **Description** |
| Age | Age of the patient (years) |
| Anaemia | Decrease of red blood cells or haemoglobin (0: false, 1: true) |
| High Blood Pressure | If the patient has hypertension (0: false, 1: true) |
| Creatinine Phosphokinase (CPK) | Level of CPK enzyme in the blood (mcg/L) |
| Diabetes | If patient is diabetic (0: false, 1: true) |
| Ejection Fraction | Blood leaving heart at each contraction (percentage) |
| Platelets | Platelets in the blood (kiloplatelets/mL) |
| Sex | Man or woman (0: woman, 1: man) |
| Serum creatinine | Level of serum creatinine in blood (mg/dL) |
| Serum sodium | Level of serum sodium in blood (mEq/L) |
| Smoking | Whether patient smokes (0: false, 1: true) |
| Time | Follow-up period (days) |
| Death Event (target) | Whether patient was deceased during the follow-up period (0: false, 1: true) |

*Table 1: Clinical Features (UCI, n.d.)*

The data was first checked to ensure it was clean. It was found that there were no missing values or or null values present in the dataset. The data types for each clinical feature was then checked. All features had the correct datatype except for age which showed a float. As age is represented in years, a float age was not possible and therefore it was changed to be an integer. Despite Booleans being categorical, because they were represented as a ‘0’ or a ‘1’, they were left to be integers. All the values of the features were then checked by creating a table of numerical features, which included Boolean features. All features had the correct range of values, and therefore no data needed to be manipulated or removed.

Before any models were created, the distribution of different clinical features individually was explored. The relationship between the features against the target variable, death event was also explored. I then applied classification machine learning techniques such as K-Nearest Neighbours and the Decision Tree to predict whether patients with heart failure would survive. During the K-Nearest Neighbours technique cross validation was used to find the best possible value of ‘k’. This was found to be 10. After this, feature selection was conducted in order to achieve the highest possible accuracy of predicting patient survival from the model. From here, a K-Nearest Neighbours Model was created.

The Decision Tree model was then created for the heart failure data. All features were used in the model as the Decision Tree is able to work well with both numerical and categorical features. Different maximum depths were tested, and it was found that a maximum depth of 1 and 2 resulted in the highest accuracies. For this reason, a maximum depth of 2 was used. From here, a Decision Tree model was created using a maximum depth of 1.

# Results

## Exploring Clinical Features

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedThe clinical features in the dataset were analysed and explored to get a better understanding of them. I first explored the sex and ages of the patients. It was found that a majority of the patients with heart failure were men as shown in *Figure 1*. The boxplot exploring age distribution highlights that the mean age of patients was 60 years old with the youngest patient being 40 and the oldest 95. It can also be shown that 50% of patients had between around 51 and 70.

The relationship between age and sex was then compared with the target variable, death event as shown in *Figure 3*. It can be seen that more men than women with heart failure in the dataset were deceased. Men who pass away from heart failure also tend to be older compares to women who passed away from heart failure. The mean age for men passing away is around 65 whereas for women it is 60. A higher number of women also get heart failure at a younger age compared to men.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedThe clinical features ejection fraction and serum creatinine also showed a strong relationship with death event. A higher chance of death can be associated with a lower percentage of blood leaving the heart at each contraction as shown in *Figure 4*. Higher levels of serum creatine in the blood also tends to be linked with death as shown in *Figure 5*.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedThe time of the follow-up period was another feature that showed a strong relationship with death. As shown in *Figure 6*, it can be seen that a shorter follow-up period suggests a higher chance of the patient being dead. The mean time for a follow-up for patients that passed away was around 45, whereas for patients that survives was around 170.

## K-Nearest Neighbours

After completing cross-validation and feature selection, a K-Nearest Neighbours model was created for data. A confusion matrix was created based on the model as shown in *Figure 7*. The confusion matrix demonstrates that the patient was correctly predicted to be alive 77 times and 2 times an alive patient was predicted to be dead. A patient who was dead was predicted to be alive 21 times and was predicted correctly as being dead 20 times.

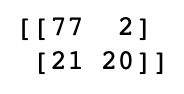


Figure 7: Confusion Matrix for K-Nearest Neighbours

A classification report was created to gain a further understanding of how well the model is able to predict patient survival as shown in *Figure 8*, where ‘0’ represents patients who survived, and the ‘1’ represented patients who are deceased. The precision for patients who survived is 0.79 whereas for deceased patients it is 0.91. This means that of all the predicted instances, patients who survived were correct predicted 79% of the time, and deceased patients who predicted correctly 91% of the time. The recall of patients that survived suggests that of all alive patients in the dataset, it was successfully predicted 97% of the time. For deceased patients the recall was lower by almost half with only 49% of these patients being correctly predicted. As shown in the report, this model has an overall accuracy of 0.81.

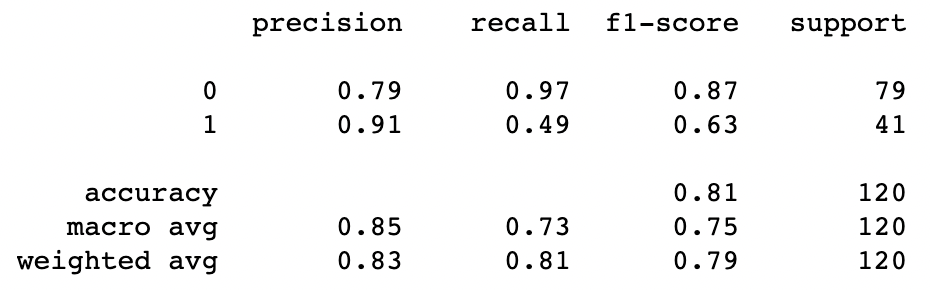


Figure 8: Classification Report for K-Nearest Neighbours

## Decision Tree

After finding the optimal maximum depth of the tree, a model was created using the data. A confusion matrix was then created as shown in *Figure 9*. The matrix shows that patients who survived were correctly predicted 60 times, and incorrectly predicted as deceased 4 times. Deceased patients were correctly predicted 16 times, and incorrectly predicted times.

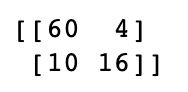


Figure 9: Confusion Matrix for Decision Tree

As shown in *Figure 10*, as classification report was generated for the Decision Tree. The precision suggests that of all predicted instances, 86% were predicted correctly for patients were survived and 80% for patients who were deceased. The recall highlights that patients who survived were successfully predicted 94% of the time and deceased patients 62% of the time. The report shows that the model created had an accuracy of 0.84.

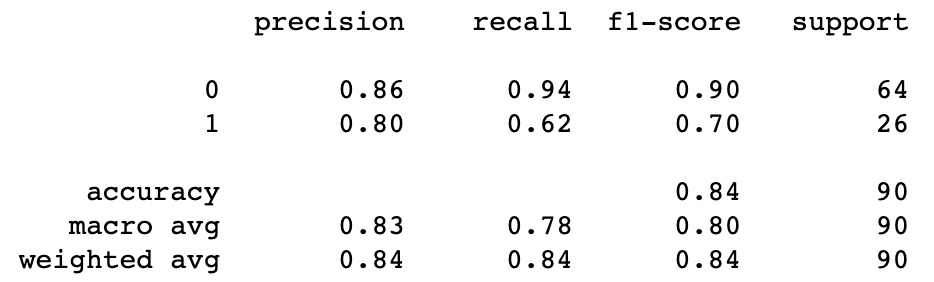


Figure 10: Classification Report for Decision Tree

A Decision Tree was then generated as shown in *Figure 11*. Based on the first one the model is able to predict using time, serum sodium and serum creatine. Based on the second one it can predict using just the time.

\*\* Decide which to use and also explain gini.

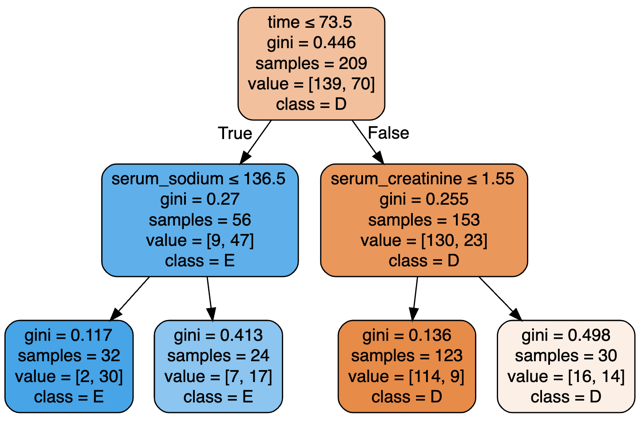


Figure 11: Decision Tree

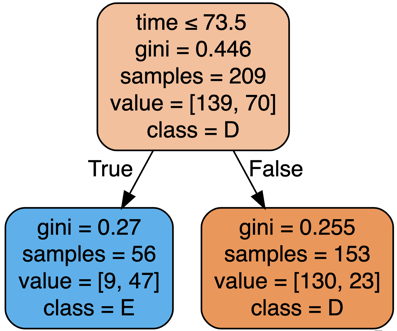


Figure 11: Decision Tree

# Discussion

\*\* Analyse the confusion matrix and the classification report for both models

# Conclusion

Could we only use one or two features to predict heart failure instead of the 10?

# References

Add my own references, like to dataset and the actual research paper

Also to the slides

Open Universiteit. (2018). Software Architecture. Retrieved from https://www.ou.nl/documents/40554/791670/IM0203\_03.pdf/30dae517-691e-b3c7-22ed-a55ad27726d6

Sarwar, T. (2022). Software Engineering Fundamentals - Object-Oriented Structural. RMIT. Retrieved from https://rmit.instructure.com/courses/101453/files/23520182?wrap=1

Sarwar, T. (2022). Software Engineering Fundamentals - Requirements and Elicitation. RMIT. Retrieved from https://rmit.instructure.com/courses/101453/files/23392441?wrap=1

Sarwar, T. (2022). Software Engineering Fundamentals - Scrum Process & Parts. RMIT. Retrieved from https://rmit.instructure.com/courses/101453/files/23402916?wrap=1

Sarwar, T. (2022). Software Engineering Fundamentals - Software Processes. 20-27. RMIT. Retrieved from https://rmit.instructure.com/courses/101453/files/23399023?wrap=1

Svirca, Z. (2020, May 30). Everything you need to know about MVC architecture. Retrieved from Towards Data Science: https://towardsdatascience.com/everything-you-need-to-know-about-mvc-architecture-3c827930b4c1

<https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5>

<https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>

<https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>

<https://www.nhlbi.nih.gov/health/heart-failure>

<https://www.nhlbi.nih.gov/health/heart-failure/causes>