COSC2670 – Practical Data Science

Assignment 2

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

Affiliations: RMIT University

Date of Report: 23/05/2022

Student Name: Puneet Kaur Grewal

Student ID: s3900991

Student Email: 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

1. Table of Contents

2. Abstract

The aim of this report was to investigate if the survival of patients with heart failure could be predicted using clinical 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 survival using data from the medical records of patients. The Decision Tree models provides a higher accuracy to determine the chances of survival of patients using the features time, age, and serum creatinine. The report concludes that medical professionals can get a better understanding of their patients' chances of survival using data from medical records, specifically looking at time, age, and serum creatinine. It is recommended that medical professionals look at medical records and see potentially threatening data according to the Decision Tree model, so they can monitor and treat patients more carefully, hopefully leading to potential patient recovery.

3. Introduction

Heart failure is a type of cardiovascular disease which happens when the heart us unable to pump a substantial amount of 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 surrounding 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.

4. Methodology

The research outlined in this report used a dataset from 2015 which included information collected from 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, 2020*)*

The data was first checked to ensure it was clean. It was found that there were no missing or null values present in the dataset. The data type 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 was individually explored. The relationship between the features against death event was also explored. I then applied classification 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 feature selection was conducted using default parameters for the K-Neighbours Classifier. During the feature selection, the features 'age', 'diabetes', 'ejection fraction', 'platelets', and 'serum sodium' were removed. The parameters were then tuned using the new features obtained from feature select. A combination of possible values for 'k', 'weights', and 'p' were used to obtain the highest possible accuracy. The best possible values were found to be 'k=6', 'weights = uniform', and 'p = 1'. 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 creation as the Decision Tree is able to work well with both numerical and categorical features and does its own feature selection to get the best combination of features. Different maximum depths were tested, and it was found that a maximum depth of 1 and 2 resulted in the highest accuracies. A maximum depth of 2 was used instead of 1 to prevent underfitting of the model. From here, a Decision Tree model was created using a maximum depth of 2.

5. Results

5.1. Exploring Clinical Features

The 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 to investigate the type of patients who had heart failure. 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 were between around 51 and 70 years old.

Figure 1: Sex Count of Patients

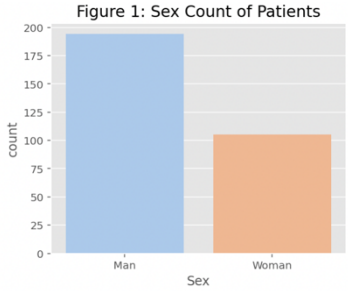
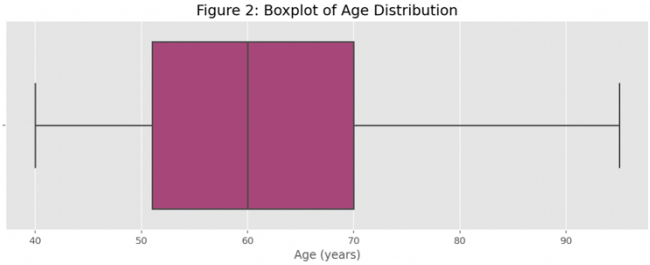
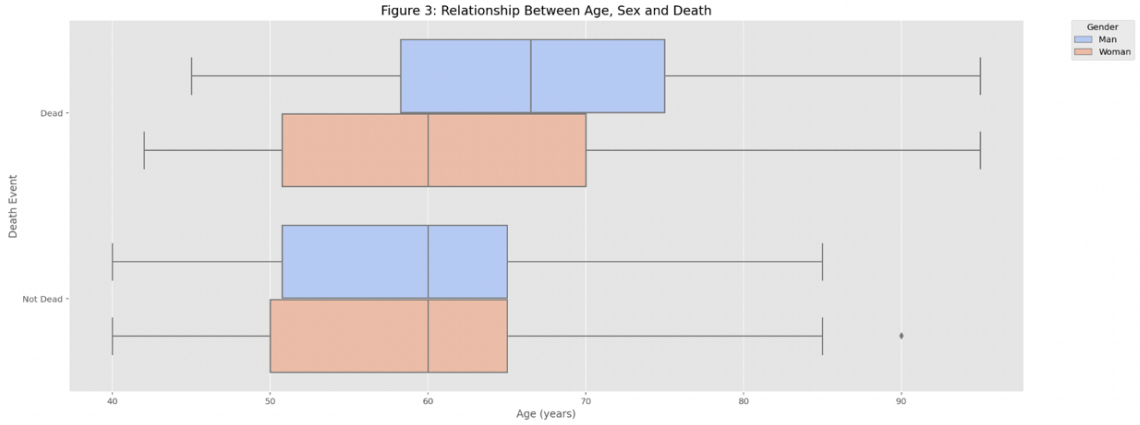


Figure 2: Boxplot of Age Distribution



I then explored the data to see if there was a relationship between the age and sex of a patient and whether they survived. From *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.

Figure 3: Relationship Between Age, Sex and Death



The relationship between clinical features ejection fraction and serum creatinine with death event was explored to see if there was a link between them. It was found that there was a strong relationship with these two features and 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*.

Figure 4: Ejection Fraction and Death

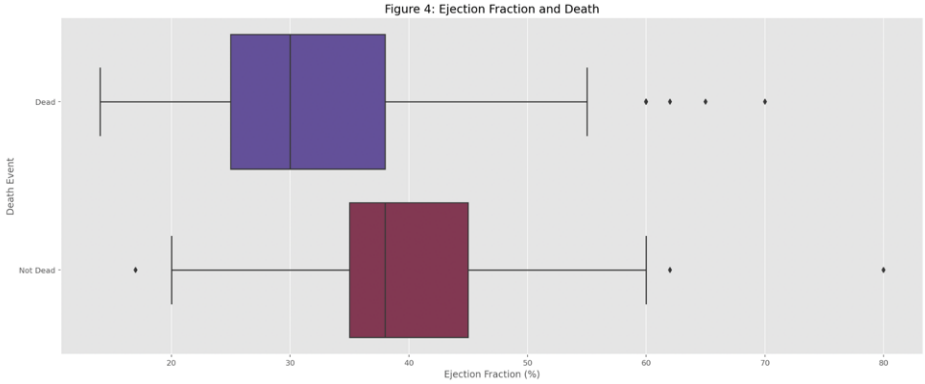
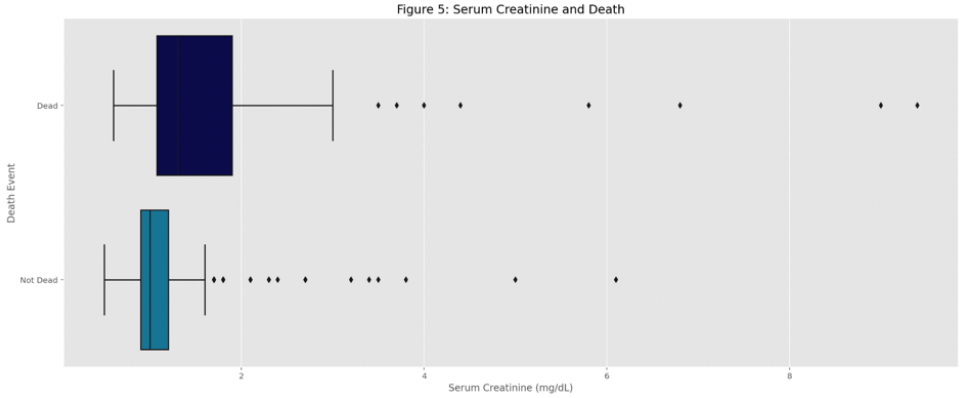
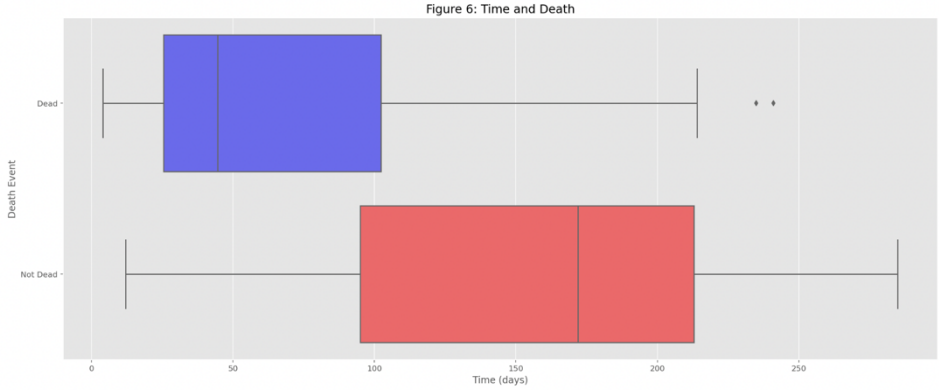


Figure 5: Serum Creatinine and Death



The 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 days, whereas for patients that survived was around 170 days. It was originally suspected that the 'time' variable could potentially cause a data leak as the target variable, 'death event' is if the patient died during the follow-up period. However, *Figure* 6 shows there is some overlap between the number of follow up days for both death and survival. This is particularly clear as the follow-up period for almost 25% of patients who died, is similar to the period where 50% of patients survived. For this reason, it was concluded that the feature 'time' is not a data leak.

Figure 6: Time and Death



5.2. K-Nearest Neighbours

After completing feature selection and parameter tuning, 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 patients were correctly predicted to be alive 78 times and 1 time an alive patient was predicted to be dead. A patient who was dead was predicted to be alive 20 times and was predicted correctly as being dead 21 times.

```
[[78  1]
 [20 21]]
```

Figure 7: Confusion Matrix for K-Nearest Neighbours

A classification report was created to gain a further understanding of how well the model can predict patient survival as shown in *Figure 8*, where '0' represents patients who survived, and the '1' represents patients who are deceased. The precision for patients who survived is 0.80 whereas for deceased patients it is 0.95. This means that of all the predicted instances, patients who survived were correct predicted 80% of the time, and deceased patients who predicted correctly 95% of the time. The recall of patients that survived suggests that of all alive patients in the dataset, it was successfully predicted 99% of the time. For deceased patients the recall was lower by almost half with only 51% of these patients being correctly predicted. As shown in the report, this model has an overall accuracy of 0.82.

	precision	recall	f1-score	support
0	0.80	0.99	0.88	79
1	0.95	0.51	0.67	41
accuracy			0.82	120
macro avg	0.88	0.75	0.77	120
weighted avg	0.85	0.82	0.81	120

Figure 8: Classification Report for K-Nearest Neighbours

5.3. 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 76 times, and incorrectly predicted as deceased 3 times. Deceased patients were correctly predicted 26 times, and incorrectly predicted 15 times.

```
[[76  3]
 [15 26]]
```

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, 84% were predicted correctly for patients were survived and 90% for patients who were deceased. The recall highlights that patients who survived were successfully predicted 96% of the time and deceased patients 63% of the time. The report shows that the model created had an accuracy of 0.85.

	precision	recall	f1-score	support
0	0.84	0.96	0.89	79
1	0.90	0.63	0.74	41
accuracy			0.85	120
macro avg	0.87	0.80	0.82	120
weighted avg	0.86	0.85	0.84	120

Figure 10: Classification Report for Decision Tree

A Decision Tree was then generated as shown in *Figure 11*. The Decision Tree uses the clinical features time, age, and serum creatinine. It splits time based on whether the follow-up period is less than or equal to, or greater than 73.5 days. Age is split at 66.5 years, and serum-creatinine is split at 1.45mg/dL. The tree has a maximum depth of 2.

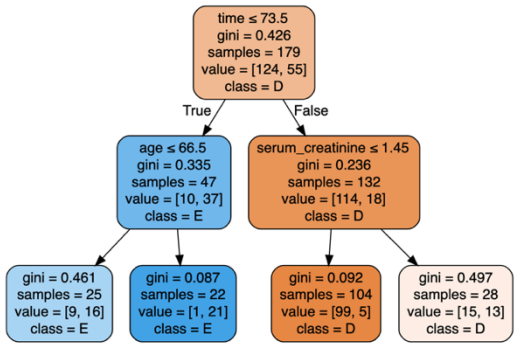


Figure 11: Decision Tree

5.4. Comparing the Two Models

The Decision Tree has an accuracy of 0.85 which is slightly higher than the accuracy of 0.82 generated by the K-Nearest Neighbours model. The f1 score which is the mean of precision and recall for both dead and alive patients is higher in the Decision Tree model suggesting that the Decisoin Tree has a better performance for predicting all patients compared to the K-Nearest Neighbours model. The K-Nearest Model uses a total of 7 features for its model compared to only 3 used by the Decision Tree.

6. Discussion

The results gathered from this study suggest that it is possible to create models to help medical professionals to predict survival of patients who have heart failure. It can be seen that the Decision Tree model allows for a higher accuracy of 85% compared to the 82% achieved from K-Nearest Neighbours. This could perhaps be due to the fact that K-Nearest Neighbours does not deal well with categorical features, which are present in this dataset. The decision tree was able to achieve an accuracy of 85% using only the features 'time', 'age', and 'serum creatinine'. When exploring the features in the data, it was found that 'time' and 'serum creatinine' had a strong relationship with the target feature, 'death event.' This could perhaps be why the Decision Tree used these features. 'Age' on the other hand showed a high relationship with 'death event' for only males as shown in *Figure 3*, whereas for females there was no clear distinction. It is also important to mention that 'age' was one of the features that was removed during feature selection for the K-Nearest Neighbours model. The feature was perhaps selected in the Decision Tree as there are more males in the data than females, and therefore age did help to predict the 'death event'. The Decision Tree model suggests that medical professionals are able to predict survival of patients using only the features used in the tree. However, it is important to remember that age seems to be a better indicator of survival for males.

It was unexpected that 'ejection fraction' was not selected as a feature in the Decision Tree or the K-Nearest Neighbours model as during the exploration of features, a strong relationship was found.

7. Conclusion

Ultimately, it was found that it is possible to predict survival of patients with heart failure using information found in medical records. However, it is difficult to generalise this research to all patients with heart failure. This is because the sample only consisted of 299 samples of patients from Punjab, Pakistan which is not representative of all patients all around the world with heart failure. It would be beneficial to replicate this research with datasets provided from hospitals from different countries around the world. Extraneous variables such as diet, physical measurements (height and weight), and past history of heart and lung related diseases could have all impacted an impact chance of survival. For this reason, including this information in the data would perhaps allow for a better prediction model.

8. References

Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20,** 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

UCI (2020) Heart Failure Clinical Records Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

WHO (2021) Cardiovascular Diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

NIH (n.d.) What is Heart Failure. Retrieved from https://www.nhlbi.nih.gov/health/heart-failure

NIH (n.d.) Causes and Risk Factors. Retrieved from https://www.nhlbi.nih.gov/health/heart-failure/causes

Ren, Y 2022, 'Classification' Powerpoint slides, COSC2670, RMIT University, Melbourne