Student ID: S3851265 + S3785169

Student Name: Brian Machleish Rabino + Yanbo He

|  |
| --- |
| 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*. |

**How machine learning can predict the survival rates of patients with heart failure through medical statistics**

Author(s): Brian Rabino, Yanbo He  
Affiliations: Royal Melbourne Institute of Technology  
Contact Details: s3851265@student.rmit.edu.au, s3785169@student.rmit.edu.au  
Date of Report: 23 May 2021

**Table of Contents**

- Abstract

- Introduction

- Methodology

- Results

- Discussion

- Conclusion

- References

**Abstract**

The purpose of this report is to investigate the ability for machine learning to be applied to medical scenarios regarding heart failure, predicting death and survival rates through given medical information. We analysed a dataset of 299 recorded heart failures at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April-December 2015. According to the results, machine learning models can, to an extent, predict survival and death rates with some degree of accuracy, using some amount of patient information provided by hospitals and using specific classifier models.

**Introduction**

Cardiovascular diseases are extremely prevalent globally, especially heart failures. Heart failures occur when the heart is not able to pump enough blood to meet the needs of the body. By using medical technologies which quantifies patient information, such as symptoms, features, status, and performing analysis with machine learning, it is possible to perform modelling techniques that can produce predictive results using the patient(s)’ medical records.

**Methodology**

The dataset was first retrieved from the original data curators Ahmad and colleagues, who made their dataset freely available to the public. We left the data shape as is, with 299 rows and 13 columns. All 299 patients had histories of heart failure and were aged between 40 to 95, with 105 women and 194 men. For our analysis, we decided to retrieve all 13 attributes and perform data exploration visually and statistically before proceeding with the modelling portion of the report.

To choose the variables to be used in data exploration, we must understand briefly what each variable in the dataset means. Creatine phosphokinase (CPK) is the level of CPK enzyme in the blood, produced when muscle tissue is damaged, which may indicate heart failure. Ejection fraction is the percentage of blood that leaves the left ventricle of the heart. Serum creatinine is waste product of creatinine. Serum sodium is a measurement of sodium in the blood, where abnormally low levels may indicate heart failure. Death event is our target in this project and indicates whether a patient lived or died before the end of the follow-up period. Anaemia, high blood pressure, diabetes, sex, and smoking were other categorical variables provided by the data.

Before any data exploration, we must first pre-process the data such that it is largely free from errors, and we are able to get consistent and clear results. To do this, we applied a number of checks before any visualization or modelling, such as checking for correct datatypes, value counts, missing null values, impossible values, or bad lines. After ensuring the data is processed properly and ready for exploration, we then proceeded to data exploration.

In this particular task, it is because the data has been presented with clear labels and that the output of the algorithm is binary that Classification models were chosen to approach this task. K-Nearest Neighbours is selected as the first algorithm to treat our model. It is simple to implement and powerful with regards to the searching space and categorizing observation into classes. As the dataset is based on non-linear data, Decision Trees is selected as the second algorithm for its high effectiveness with non-linear data. A simple hill climbing method was also used to identify the amount of features that should be used from the dataset.

**Results**

**Data Exploration**

Through exploration of attributes in the dataset, we are able to depict a clearer idea of what kind of information the dataset contains, as well as what the likely relevant features are when applying machine learning features later on.

Chart, pie chart

Description automatically generated

Simply confirming the values of the ‘sex’ attribute with Figure 1 pie chart and value count, with a 64.88% to 35.12% proportion, male to female respectively. There were more men than women who had heart failures in this dataset.

Chart, histogram

Description automatically generated

The histogram chart depicts the age range of all patients in the dataset. We see that most patients were between ages 40-70, with most patients being around age 60.

It is now time to analyse the more medically related attributes, which will create a clearer depiction on the patients in the dataset.

The following Figures 3, 4, 5, and 6 visualize the proportions of patients who have a certain condition, either smoking, anaemia, or diabetes.

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, pie chart

Description automatically generated

The data shows that the majority of patients are non-smokers (67.89%) compared to smokers (32.11%). There is also a majority of non-anaemic patients (56.86%) compared to anaemics (43.14%), as well as non-diabetics (58.19%) to diabetics (41.81%). High blood pressure is also a non-majority, with around 190 patients not affected, and around 110 affected.

Figure 7 is a simple value count for the number of deaths that occurred due to heart failure in the dataset, with roughly 200 patients survived compared to around 100 patients died.

Chart, histogram

Description automatically generated

Figure 8 depicts a histogram of serum creatinine levels, rightly skewed, showing the majority of patients with levels below roughly 1mg/dL.

Chart, line chart

Description automatically generated

The density plot for Ejection Fraction in Figure 9 is quite symmetric, peaking at around 40% Ejection Fraction. This tells us that the amount of heart failures is most common when Ejection Fraction is around 40%.

Chart, box and whisker chart

Description automatically generated

Figure 10 is a boxplot graph visualizing statistical data such as a mean of around 250,000 kiloplatelets/mL (k/mL), with the upper whisker around 400000 k/mL and the lower whisker of 100000 k/mL. There are a few notable outliers, such as the three patients above 600000 k/mL.

Now that we have an idea of what the data is showing us, we must begin to explore the relationships between these attributes. We can determine how certain attributes affect others easily through some more complex visualizations in the form of the following graphs.

Note: As we are basing the model depending on the ability to predict survival rate based off of serum creatinine and ejection fraction, we must create a mask of the patients that did and did not survive to form a deeper understanding of specific attributes and their relation to survival rate. We will refer to death events as deceased patients interchangeably.

Chart, line chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

Hypothesis – The older the patient, the less likely to survive heart failure.

This density plots in Figures 11 and 12 shows us the age range of survived and deceased patients. We see that the most common ages in the dataset hover around the ages of 60. Figure 12 especially show us that around the ages of 70-95, there is a significant increase in deceased patients comparative to the same age range in the survived patients, thus supporting our hypothesis.

Chart, box and whisker chart

Description automatically generated

Figure 13 more accurately shows the relationship between the attributes of death events and age. As hypothesised, the older patients, the less likely they are to survive heart failure.

**Chart, box and whisker chart

Description automatically generated**

Hypothesis – The higher the levels of serum creatinine, the more likely a patient will die to heart failure.

Figure 14 shows a boxplot which supports this hypothesis, as the interquartile range, or the majority, of patients have higher serum creatinine levels in their systems resulting in heart failure and death. The upper and lower whiskers of this graph, meaning the minimums and maximums of serum creatinine, are also higher in deceased patients compared to survived patients. There are also significantly more outliers of high serum creatinine levels when the patients survive, meaning it is less likely. Thus, the hypothesis is supported.

**Chart, box and whisker chart

Description automatically generated**

Hypothesis – The lower ejection fraction levels are, the more likely a patient will die to heart failure compared to those with higher levels of ejection fraction.

Figure 15 portrays how ejection fraction levels affect the patient’s survivability. In those that survived, the IQR of ejection fraction are around 35% to 45%, whereas in those that did not, values of around 25% to 38%. Thus, we can confidently draw from the graphs that those with lower ejection fraction levels are more likely to die to heart failure.

A picture containing timeline

Description automatically generated

Hypothesis – The higher the levels of creatinine phosphokinase, the more likely a patient will die to heart failure compared to those with lower levels of creatinine phosphokinase.

Figure 16 depicts that the majority of creatinine phosphokinase levels are below 2000 mcg/L in BOTH survived and deceased patients, meaning that there is actually little significant impact on survival rate. Although, the graph for deceased patients do contain outliers of CPK, notably above 4000 mcg/L, and a lack of outliers of CPK above 3000 mcg/L in the survived patients’ graph. This hypothesis is not supported by the data.

Chart, box and whisker chart

Description automatically generated

Hypothesis – Lower levels of serum sodium will likely result in death to heart failure compared to those with higher levels of serum sodium.

Figure 17 displays how lower levels of serum sodium does result in death to heart failure, as seen by the slightly lower IQR in the deceased patients’ boxplot, around 133 to 137 mEq/L, compared to the survived patients’ boxplot IQR of around 135 to 140 mEq/L of serum sodium. Although there are notable outliers, such as a patient with around 113 mEq/L of serum sodium, yet surviving, they are exceptions to the norm, and the hypothesis is largely supported by this visual data.

Chart, histogram

Description automatically generated

Hypothesis – Abnormal levels of platelets (below 150000 k/mL or above 450000 k/mL) will likely result in death to heart failure compared to those with normal levels of platelets.

Figure 18 demonstrates how too little or too many platelets in a patients’ system affects the likelihood of a resulting heart failure compared to those with normal platelet levels. As seen in the survived patients’ graph, there are notably more patients with platelets below 150000 k/mL, as well as above 400000 k/mL, compared to the deceased patients’ graph, using the y-axis scale as rough measurements. Therefore, this hypothesis is not supported according to the data drawn from Figure 18.

Chart, pie chart

Description automatically generated

Hypothesis – An individual with high blood pressure is more likely to result in death to heart failure compared to an individual without high blood pressure.

Figure 19 portrays the proportions of which survived patients and deceased patients do or do not have high blood pressure. Using this graph, we see that there is a larger proportion of patients who died to heart failure while having high blood pressure (40.62%) compared to those who survived while having high blood pressure (32.51%). From the dataset, we can support the hypothesis that patients with high blood pressure are more likely to result in death to heart failure compared to those without high blood pressure.

Hypothesis – A patient who smokes is more likely to die to heart failure than a patient who does not smoke.

Figure 20 contains two pie charts which present how the proportion of survived patients who smoke and the proportion of deceased patients who smoke is incredibly similar. Such that, there is a marginal difference in those who smoked and died (31.25%) and those who smoked but survived (32.51%). Thus, we can conclude from the data that the hypothesis is not supported, and that patients who smoke are NOT more likely to die to heart failure than patients who do not smoke.

Hypothesis – Heart failure is not more prevalent in one sex compared to the other.

Figure 21 depicts how the proportion of sex within the patients’ dataset are similar in value to each other, as seen in the survived males (65.02%) and deceased males (64.58%) as well as survived females (34.98%) and deceased females (35.42%). Thus, the hypothesis is supported in that heart failure is not more prevalent in one sex to the other, and that heart failure occurs in roughly the same proportions, as seen in the visualized data in Figure 21.

Chart, pie chart

Description automatically generated

Hypothesis – Patients who are anaemic are more likely to result in death to heart failure than patients that are non-anaemic.

Figure 22 displays that the proportion of anaemics who survived (40.89%) is less than the proportion of anaemics who died (47.92%), and that there are more non-anaemics who survived (59.11%) than non-anaemics that died (52.08%). Thus, the graph supports our hypothesis that patients who are anaemic are more likely to die from heart failure compared to patients who are non-anaemic.

**Data Modelling**

**K-Nearest Neighbours Classification**

When identifying what features go into our models, the relation between each column from the dataset was first evaluated through different visualizations in order to find the appropriate features. Then a simple hill climbing method was used to identify the number of features that should be used from the dataset.

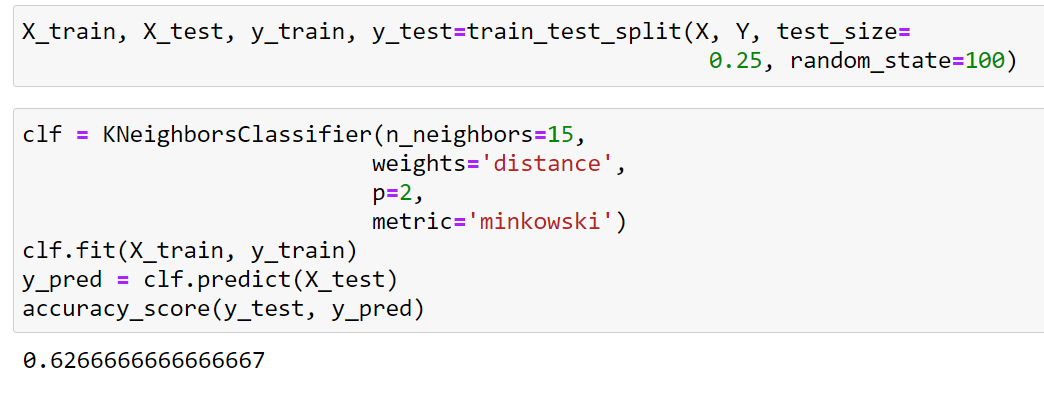


As shown in the image, 10 features will be selected. And from the previous data exploration findings, we have decided to use the following 10 features: age, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking and anaemia.

N-neighbors: Through data exploration, some outliers were found but to a reasonable amount. However, the overall dataset is clean. As a result, “13” was set as the parameter for n-neighbors because it can not only suppress the effects of “noise” but also remain the boundaries of different classes.

Weights: Since the issue of outliers was addressed, “distance” was set as the parameter for weights as it weights points by the inverse of their distance. By giving the closer neighbors more influence it helps reducing the impact of outliers data.

Metric and p: Standard Euclidean metric are used for the distance metric, where metric is set to “minkowski” and p is set to “2”. Having a lower p also help reducing the influence of a high difference in some dimension, though in this particular task it is not too big of a problem.



**Decision Tree**

When identifying what features go into our models, a simple hill climbing method was first used to identify the number of features that should be used from the dataset. Then the relation between each columns from the dataset was evaluated through different visualizations in order to find the appropriate features.

Graphical user interface, text

Description automatically generated

As shown in the image, 8 features will be selected. And from the previous data exploration findings, we have decided to use the following 8 features: age, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium and sex.

Criterion:

Since the training dataset is not big, “entropy” was selected for the criterion as it can provide a slightly better result comparing to “gini” criterion.

Random\_state:

A random state of “100” is given for the model to establish the same tree every time.

Max\_depth:

In order to control over-fitting issue with the model, a maximum depth of tree is set to 5.

Min\_samples\_split:

A minimum sample for a node split can be used to control over-fitting and prevent a model from learning relations which might be highly specific to the particular sample selected for a tree. It has been set to “12”.

Min\_sample\_leaf:

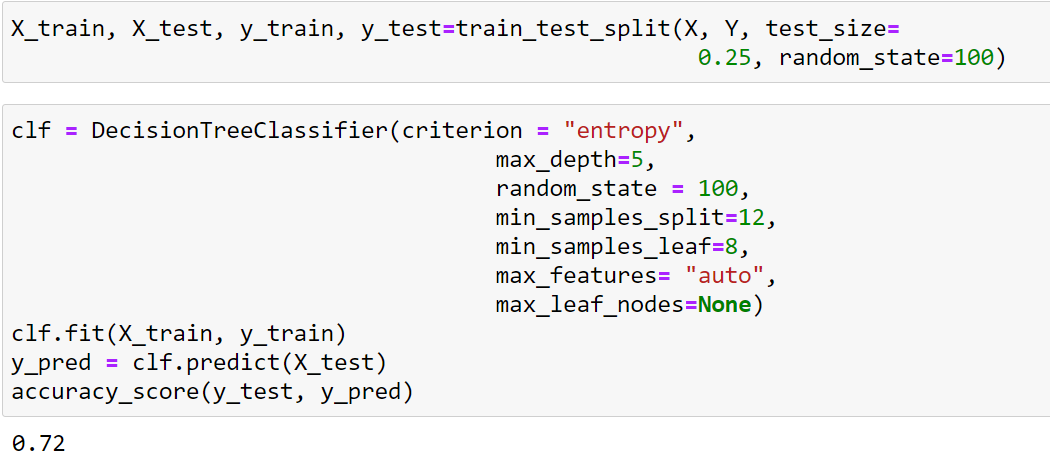
It is because the dataset contains mostly imbalanced classes, a minimum sample for a terminal node was set to a lower value, “8”, in order to control over-fitting.

Max\_features:

The maximum number of features to consider when looking for the best split is set to “auto”.

Max\_leaf\_nodes:

The Maximum number of terminal node is set to “None” as no limit was considered.

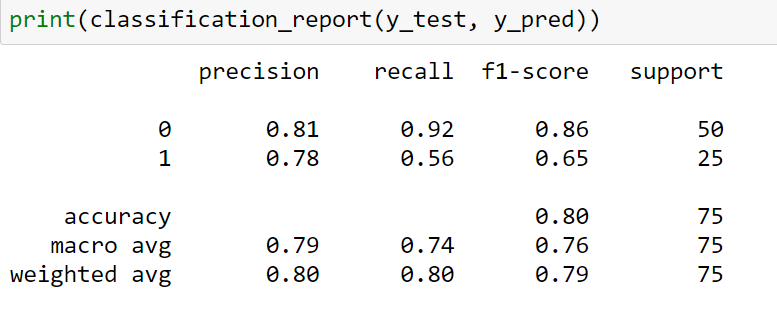


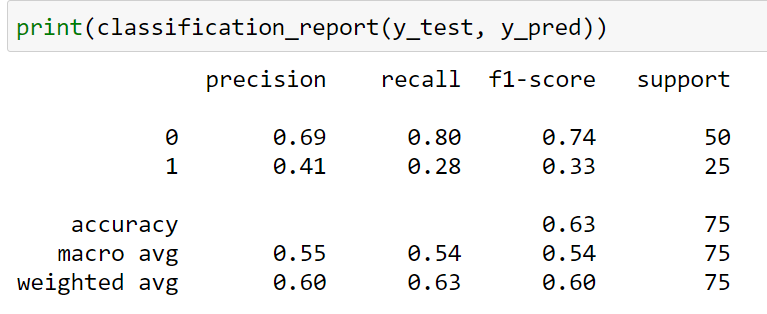
**Discussion**

Exploring the relationships between the pairs of attributes helps us establish how the model has prioritized certain features. One such relationship would be age, which we hypothesised would have a larger impact on the death events compared to what was actually visualized in the graph (Figure 11 and 12). Although there is a noticeable difference in death rates between the older and younger patients, we expect that with a larger sample size, this difference would be more prominent and represented in the density plots.

As with creatinine phosphokinase, we also assumed that the data would display a greater difference between the levels of CPK in deceased patients comparatively to survived patients, as CPK is released under stimulus of muscle and tissue damage. The two histograms were largely similar, with the x-axis scale being larger due to a few outliers in the deceased patients graph compared to the survived patients’ graph. The model may have also seen these results and decided on CPK as a feature with less importance and impact on the resulting predictive power.

When comparing the 2 models: KNeighbors and Decision Tree, a classification report table was generated in order to evaluate how each model preforms. Since the task is to predict heart failure, the recall value is what we considered the most important metric as it indicates how many of the positive classes the model is able to predict correctly. And from the result we can say that the model of Decision Tree has better outcome and thus it is a more suitable model for this task.

Decision Tree: 

KNeighbors: 

**Conclusion**

After pre-processing the data, exploring the attributes and relationships, as well as testing and evaluating the dataset with two of our classification models, we confirm that the ability for machine learning to implement predictive power through our trained models, something that cannot be done by a human, does provide relevant results to an extent. Using only the given electronic health records from the dataset, and with limited attributes, it is possible for machine learning to predict the survival rates of patients.

One obvious limitation is the lack of a larger sample size in order to fully develop and train the model to it’s potential. As well as a lack of sample size, reproducibility using datasets from countries globally would largely benefit to the machine learning capabilities of our work. Due to limited attributes, we believe that adding more detailed and extensive attributes that distinguish important features related to heart failure would only further improve the potency of our model.

**References**

Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. PLoS ONE. 2017; 12(7):0181001.

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

<https://quantdare.com/decision-trees-gini-vs-entropy/#:~:text=The%20Gini%20Index%20and%20the,is%20%5B0%2C%201%5D.&text=The%20gini%20index%20has%20also,which%20are%20not%20very%20significant>

<https://corporatefinanceinstitute.com/resources/knowledge/other/decision-tree/>

https://towardsdatascience.com/how-to-best-evaluate-a-classification-model-2edb12bcc587