Student ID: s3482450

Student Name: Matthew Bird

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 honour code by typing "Yes": *Yes*.

# Prediction of survivability of patients who experience heart failure using classification modelling.

**By Matthew Bird**

**Contact: s3482450@student.rmit.edu.au**

**16-05-2021**

**CONTENTS**

**1- cover**
**1-contents**
**2- summary**
**2- introduction**
**2- methodology**
**3- data composition**
**5- results**
**9- modelling**
**9- K- nearest neighbour**
**10- decision tree**
**11- discussion**
**11- conclusion**
**12- data attributes**
**12- references**

## Summary

The aim of this report was to investigate factors that could influence the survival rates of patients experiencing heart failure. These factors could in turn be used to predict survival rates. Analysis of the dataset showed that serum creatinine and ejection fraction could be used to predict the survival rates of patients with considerable accuracy. The findings reach a conclusion that despite the known impacts of habits such as smoking or previous medical conditions such as diabetes, these factors can be largely ignored when assessing the survivability of patients. Quantifiable measurements can be taken of the patient and used to accurately determine the risk factor of the patient in having a death event **following** a heart complication.

## Introduction

The survivability of a patient following a heart complication can be difficult to predict. Factors such as the patient health(smoking, age, diabetes) do not seem to be a reliable predictor of survivability. While these attributes certainly increase the risk of creating the conditions leading to heart complications, they are not useable as a predicable means of determining survivability. The purpose of this report is to investigate what factors can be used to provide an accurate means of predicting the likelihood of a death event following heart complications.

## Methodology:

A data set containing the results of 299 patients was used from the UCI machine learning repository[1]. This data was collected from various patients during follow up periods. The most important feature is the “DEATH_EVENT” feature which is the determining feature of if a patient lived or died following their complication. We will compare this feature against various other features to see if a link can be established between some features and a patient surviving.

The programming language used is Python 3, the data was checked for any errors, missing data or incorrect format. Any errors were corrected to ensure minimisation of inaccuracy in the models and portrayal of the data.

The features of interest were those that could have a provable relationship with a death event. Meaning we are not interested in features that increase the chance of a death event- but can definitively say if a patient will survive or die.

Two classification models were created. A K nearest neighbour and a Decision Tree. DEATH_EVENT was the obvious target feature and the features chosen in our model were [‘serum_creatinine’] and [‘ejection_fraction’]. These features were selected through a process of data exploration and the discovery of a relationship between these two features and DEATH_EVENT.

For the K nearest neighbour model, the K value was found by iterating through a range of 0-100. A comparison of the accuracy was done for each value. The best K value was stored and used in a final model. The best score achieved by the K-nearest neighbour was an accuracy of 87%. A decision tree was used but was less accurate, with a score of 75%.

The K nearest neighbour is the recommended model due to its higher score, with a K value of 12 and a p value of 1.

## Data composition

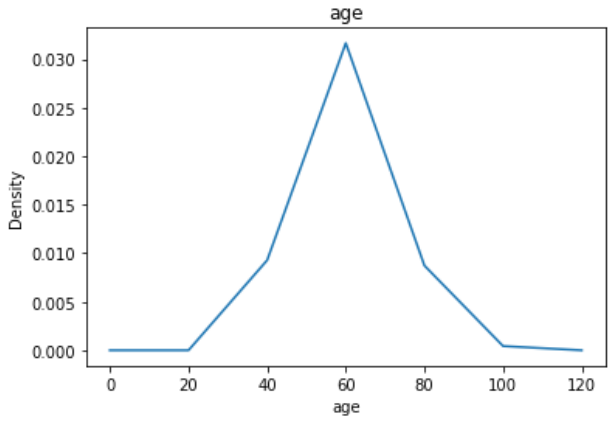


*Figure 1- age distribution*

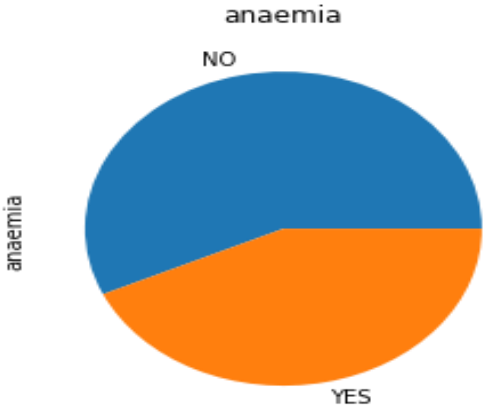


*figure 2- anaemia yes/no*

Participants in the data study were found to have a mean age of 60 years. 43% of them were found to have anaemia.

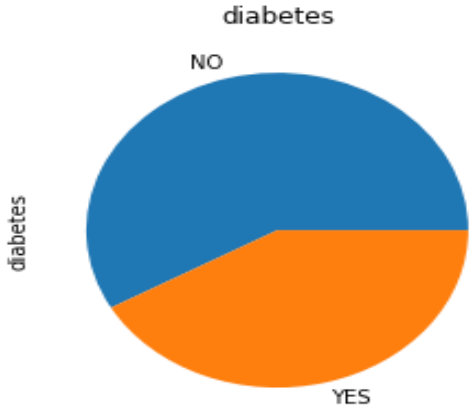


*Figure 3- diabetes yes or no*

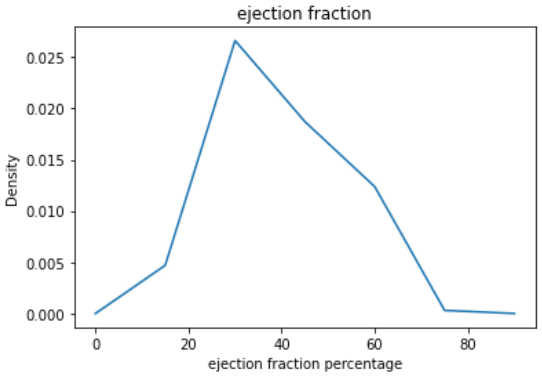


figure 4- distribution of ejection fraction levels

42% of participants were found to have diabetes.

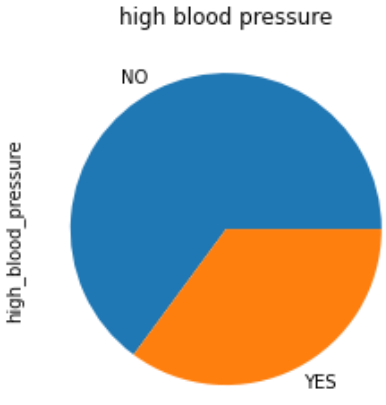


*Figure 5- high blood pressure*

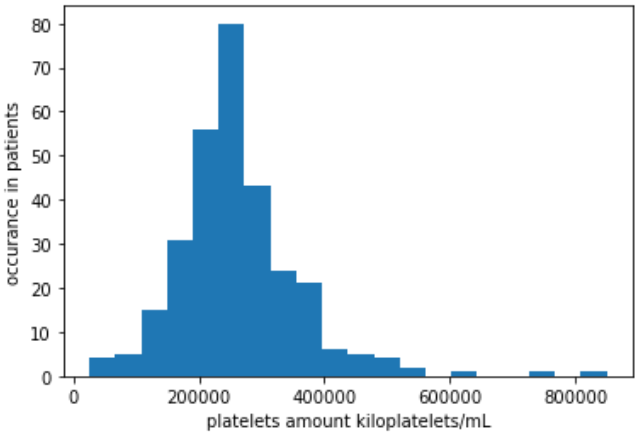


*figure 6- platelets distribution*

35% of participants were found to have high blood pressure.

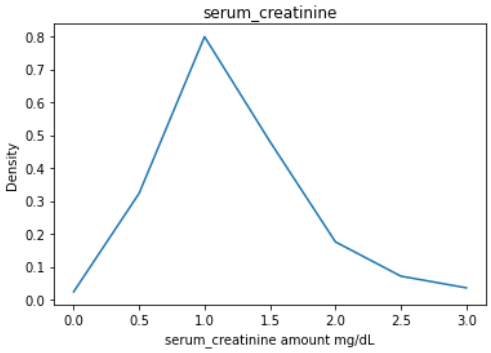


*Figure 7- serum creatinine levels distribution*

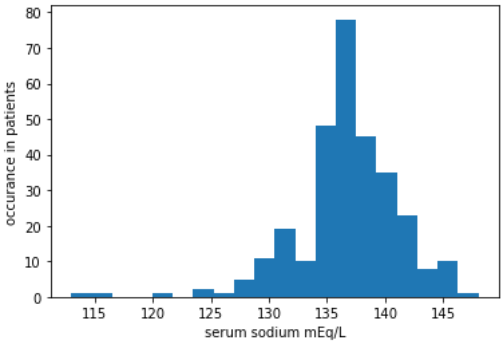


*figure 8- serum sodium levels distribution*

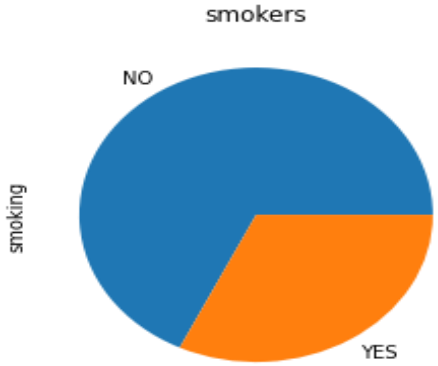


*Figure 9- smokers yes/no*

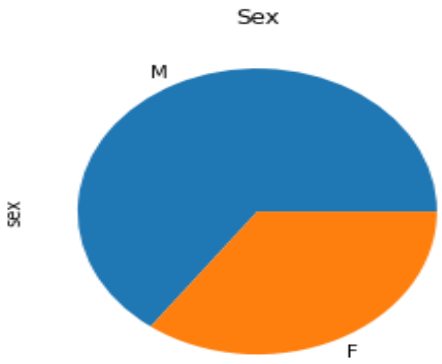


*figure 10- sex male or female*

32% of patients were smokers. 65% were male.

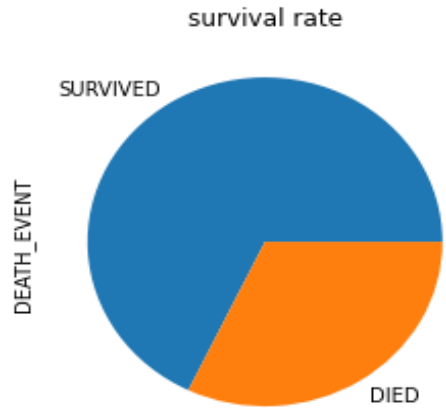


*Figure 11- death events.*

68% of patients survived and 32% experienced a DEATH_EVENT.

# RESULTS

**Assumption 1**- **It was assumed that patients with high blood pressure would be more likely to be diabetic**. A similar proportion of patients without diabetes but with high blood pressure was found as those with diabetes and high blood pressure, in relation to their respective non high blood pressure groups. We can see in figure 12 that the 2nd and fourth columns are of similar proportion to the first and third columns respectively.

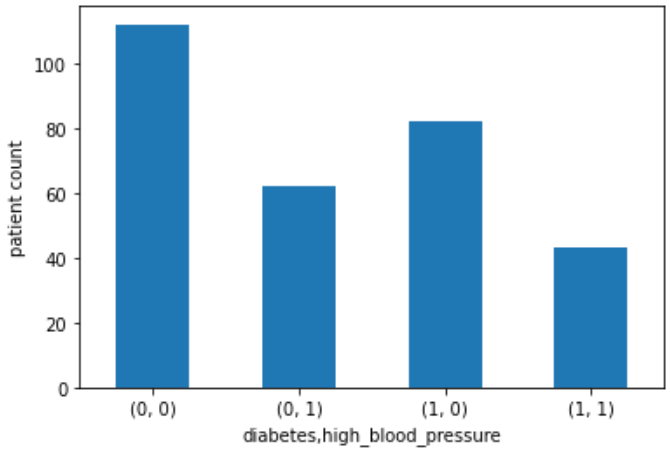


*Figure 12- diabetes and high blood pressure relationship*

**Assumption 2- Diabetes increases the chance of a death event-** it was assumed that diabetes would

Increase the chance of a death event. It was found that of the people without diabetes, 118 lived and 56 died, or approx 68% survival rate. Where as 85 people survived with diabetes and 40 died or aproximately 68% survival rate. There was no relationship found between having diabetes and experiencing a death event.

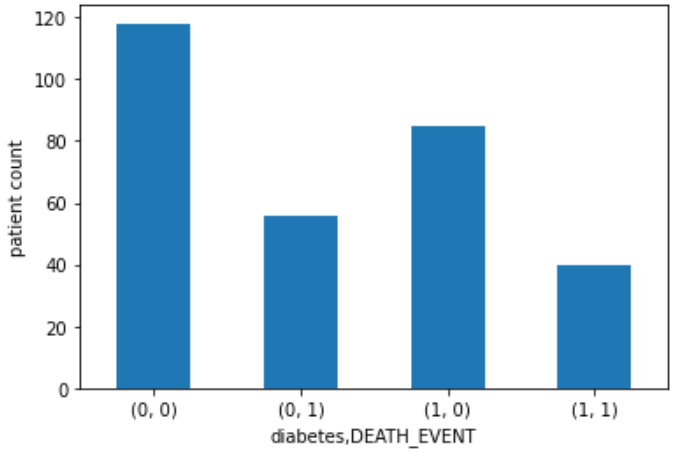


*Figure 13- relationship between diabetes and death event*

**Assumption 3- smoking increases chance of a death event- again a similar situation as assumption 2. No relationship or dramatic increase was found between death events and smoking**. 137 non smokers survived and 66 died. This is a 67% survival rate for non smokers. Of the smokers 66 lived and 30 died This is a 69% survival rate. This is a very small difference compared to non smokers. It can be assumed that there is no relationship between smoking and survival rates.

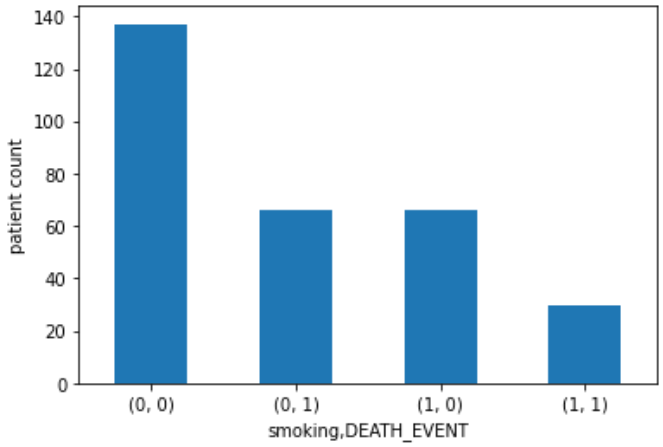


*Figure 14 - relationship between smoking and death events*

**Assumption 4- There is some kind of relationship between creatinine phosphokinase levels and platelet count that can be used to determine death event**. Looking at figure 15 we can see that there is no relationship between these two attributes that can be used to predict a death event. The relationship is somewhat random and unpredictable in terms of a death event.

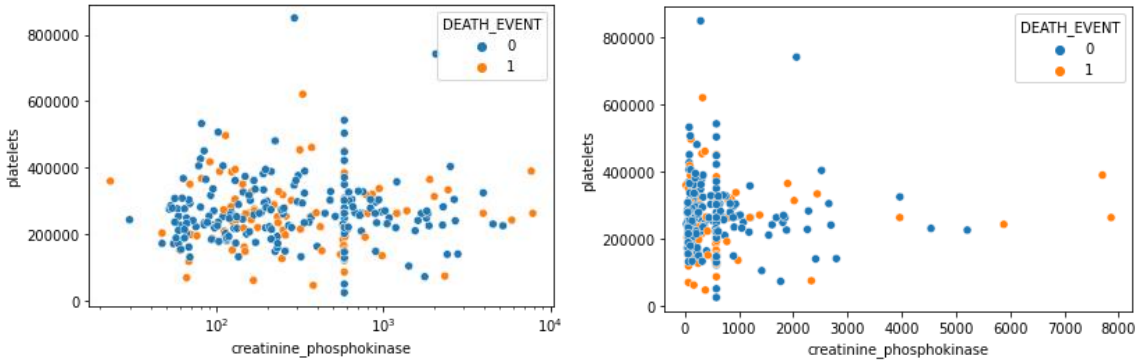


*Figure 15- platelet count and creatinine phos. Levels keyed by death event (left image is log scale for x-axis, right is linear scale)*

**Assumption 5- There is some kind of relationship between creatinine phosphokinase and ejection fraction that can be used to predict a death event**- Similar to assumption 4, we have no relationship between these two attributes. Figure 16 shows that the attributes can have a variety of values and still be either a death or survive event.

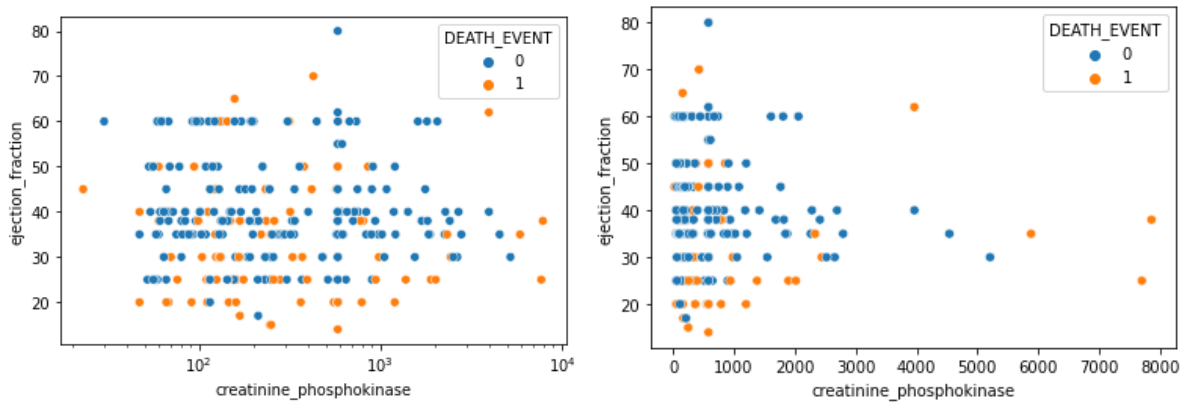


*Figure 16- relationship between creatinine phosphokinase and ejection fraction (left is log scale for x axis, right is linear)*

**Assumption 6**- high levels of creatinine phosphokinase increase chances of a death event- There is no relationship between creatinine phosphokinase levels and death event. We simply see a graph that shows similar information to figure 11. That more patients survived than died.

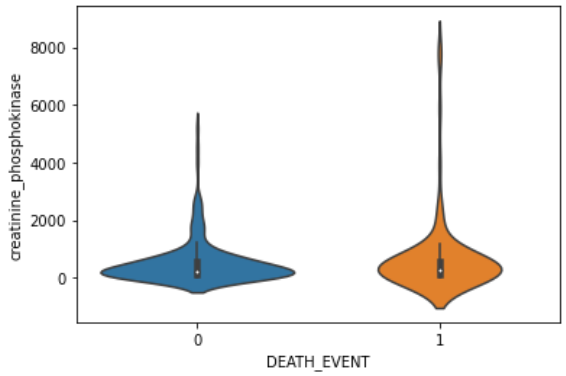


*Figure 17- creatinine phosphokinase and its relationship to death events*

**Assumption 7- high levels of ejection fraction increase chances of a death event-** We can see that there is some signs on figure 18 that ejection fraction can influence a death event. The surviving patients seem to have levels concentrated around 40 whereas the death event patients can have potentially lower levels. However, there are still considerable similarities between surviving and dying patients levels to be able to use ejection fraction alone to determine a death event.

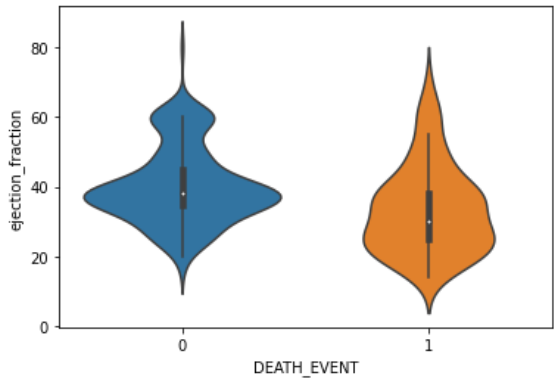


*Figure 18- relationship between death event and ejection fraction*

**assumption 8- high levels of serum creatine increase chances of a death event**-There is no noticeable relationship between serum creatinine levels and a death event. Other than a slightly broader range for death events, it alone could not be used to determine any death events.

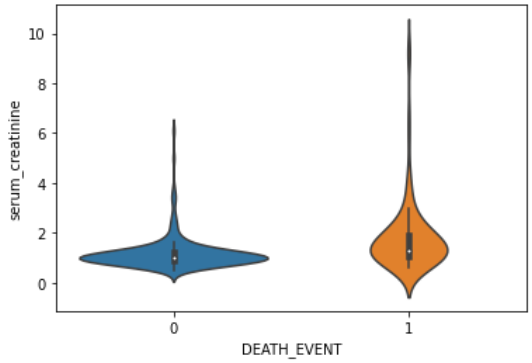


*Figure 19- serum creatinine levels and its influence on death events*

**Assumption 9- serum creatinine and ejection fraction can be used to predict death event-** This relationship so far has proved the most promising. We can see a clear area on our graph in figure 20, which can be thought of as the survival zone. Patients who have between 30 to 60 ejection fraction AND 0-2 serum creatinine levels demonstrated a very high survival rate. This shows that serum creatinine and ejection fraction are promising attributes for predicting survival rates in patients.

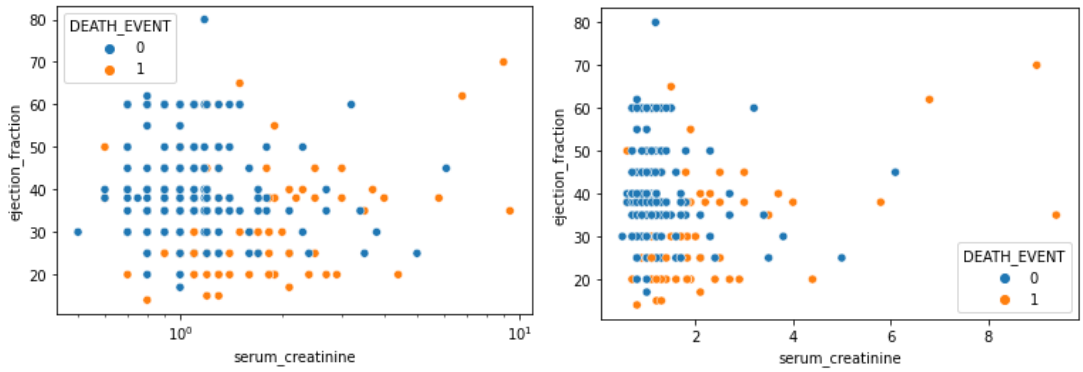


*Figure 20- the relationship between serum creatinine and ejection fraction(left x axis as log, right x axis as linear)*

**assumption 10- younger people are more likely to not experience a death event-** From figure 21 we can see a wide range of patients who survived tended to be under the age of 70. Whereas the patients who were above 70 were more likely to die. However age still does not provide a way to guarantee an ability to predict the survivability of patients, as patients from various age groups seem to be able to live or die.

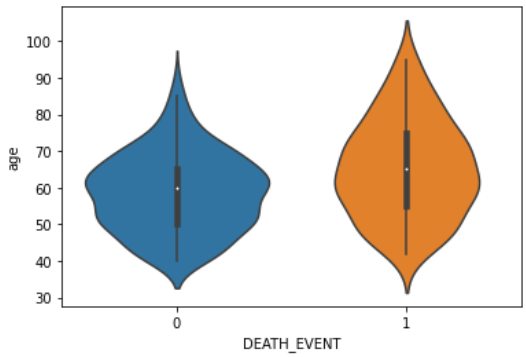


*Figure 21 relationship between age and DEATH-EVENT*

## Modelling

The data was modelled using two different classification models: K-nearest neighbour and a decision tree. All models were able to provide at least 70% accuracy but the best result was the K – nearest neighbour. The two attributes used in the models was serum creatinine and ejection fraction(See figure 20) as they proved to have the best potential for predicting death events.

## K- nearest neighbour

This model produced the best accuracy scores. Figure 20 was the only assumption to be proven true and provide a basis for how to investigate our model. The best model was the 20-80 split which provided a score of 88%(see 80-20 score in notebook). This is a good result for being able to predict patients who will be at high risk of future complications. The model did decrease down to 75% accuracy though once a 50-50 split train and test was conducted (see score in notebook). It is recommended that the k nearest model be used in future.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.98 | 0.93 | 46 |
| 1 | 0.89 | 0.57 | 0.70 | 14 |
| accuracy | | | 0.88 | 60 |
| macro avg | 0.89 | 0.77 | 0.81 | 60 |
| weighted avg | 0.88 | 0.88 | 0.87 | 60 |

*Figure 22 above. 80-20 split K nearest neighbour*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.90 | 0.86 | 84 |
| 1 | 0.70 | 0.53 | 0.60 | 36 |
| accuracy | | | 0.79 | 120 |
| macro avg | 0.76 | 0.72 | 0.73 | 120 |
| weighted avg | 0.78 | 0.79 | 0.78 | 120 |

*Figure 23 above 60-40 split K nearest*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.86 | 0.82 | 98 |
| 1 | 0.67 | 0.56 | 0.61 | 52 |
| accuracy | | | 0.75 | 150 |
| macro avg | 0.73 | 0.71 | 0.72 | 150 |
| weighted avg | 0.75 | 0.75 | 0.75 | 150 |

*Figure 24 above. 50-50 split K nearest*

## Decision Tree

The decision tree proved to be less reliable in all 3 variations of the train/ test split. It is not the recommended model for this data set. The best score was 74% from the 50-50 split(see notebook for score).

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.72 | 0.80 | 46 |
| 1 | 0.46 | 0.79 | 0.58 | 14 |
| accuracy | | | 0.73 | 60 |
| macro avg | 0.69 | 0.75 | 0.69 | 60 |
| weighted avg | 0.81 | 0.73 | 0.75 | 60 |

*Figure 25 above- 80-20 split decision tree*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.85 | 0.81 | 84 |
| 1 | 0.55 | 0.44 | 0.49 | 36 |
| accuracy | | | 0.73 | 120 |
| macro avg | 0.67 | 0.64 | 0.65 | 120 |
| weighted avg | 0.71 | 0.72 | 0.72 | 120 |

*Figure 26 above – 60-40 split decision tree*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.85 | 0.81 | 98 |
| 1 | 0.65 | 0.54 | 0.59 | 52 |
| accuracy | | | 0.74 | 150 |
| macro avg | 0.71 | 0.69 | 0.70 | 150 |
| weighted avg | 0.73 | 0.74 | 0.73 | 150 |

*Figure 26 above- 50-50 split decision tree*

## Discussion

There are two considerable findings in this report. The first is despite the fact that habits such as high blood pressure, diabetes, smoking and age are all contributors to increasing the risk of having heart problems[2], they do not seem to have a major impact on the actual survivability of the patient once the condition has come forward in the patient. Looking at figure 13 we can see that diabetes did not have a significant impact on the percentage of survival of patents vs those who did not have diabetes, both diabetic and non-diabetic patients had a 68% survival rate.

With figure 14 we can see that smoking did not increase the risk factor of a death event. Non-smokers had a 67% survival rate and smokers had a 69% survival rate.

Various other attributes also seemed to present no impact on the survivability of patients. Looking at figure 15 we can see no relationship between either platelets or creatinine phosphokinase.

The most interesting part of this report brings us to our second considerable finding. Looking at figure 18 ejection fraction does not seem to have a predictable pattern in terms of patient death events. And the same can be said for figure 19 in relation to serum creatinine. However, the combination of these two attributes produces a graph with extremely obvious patterns of survivability and death events (see figure 20). This demonstrates a clear relationship between ejection fraction and serum creatinine when it comes to prediction death events in patients experiencing heart complications.

## Conclusion

The findings of this report show that despite the negative health conditions or habits of patients that increase the risk of heart complications, these factors cannot be used to predict the actual death possibility of the patient following the heart attack. Many traditional health measurements also provide little in the way of predicting death. However, there is a clear relationship between serum creatinine and ejection fraction. This relationship shows promise in the ability to determine the risk a patient may be in following heart complications and these patients should be subject to stringent examination and medical care with the expectation of further heart complications if they show a predicted death event based on our model. The recommended model for the data set is the K-nearest model.

## Data attributes-

1- **Age**- integer value representing patients age in years
2- **Anaemia** – Boolean value representing if the patient has a deficiency of red blood cells
3- **Creatinine_phosphokinase-** an integer value representing these levels in patient
4- **Diabetes-** Boolean value representing if patient is diabetic
5- **Ejection_fraction-** integer value representing how well the heart is pumping
6- **High_blood_pressure-** Boolean value representing if patient has high blood pressure
7- **Platelets-** integer value representing platelet count for patient.
8- **Serum_creatinine –** float value representing creatinine to decliter of blood
9- **Serum_sodium –** integer value representing sodium count in patient
10- **Sex-** Boolean value representing if patient is male or female
11- **Smoking-** Boolean value representing if patient smokes
12- **Time-** integer representing amount of days follow up was with patient
13- **Death_event-** Boolean value representing if patient died.

## References

[1] UCI Machine Learning Repository: Heart failure clinical records Data Set

[2] https://www.cdc.gov/heartdisease/risk_factors.htm - What health conditions increase the risk of heart disease?