RMIT UNIVERSITY

COSC2670 Practical Data Science with Python
Assignment 2 – Data Modelling and Presentation

Prediction of Survival of
Patients with Heart Failure from Different Clinical Features

Joyal Joy Madeckal - s3860476 (s3860476@student.rmit.edu.au)

Prawal Lohani - s3853474 (s3853474@student.rmit.edu.au)

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*.

Affiliated to: School of Computing Technologies, RMIT University

Date: 29th May 2021

RMIT UNIVERSITY

Table of Contents

RMIT UNIVERSITY

Executive Summary

The aim of the report was to investigate clinical features affecting the heart failures in the patients and their survival rate. Pre-historical data of 299 heart failure patients were available for the project. Overall, it is found that among the clinical features serum creatinine and ejection fraction has a major role in determining the survival rate of the patients. As a conclusion, patients who have suffered from heart failures, the serum creatinine and ejection fraction levels must be monitored closely.

Introduction

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels including, coronary heart disease (heart attacks), cerebrovascular diseases (strokes), heart failure (HF), and other types of pathology (Chicco & Jurman, 2020). A lot of people are dying around the world due to heart diseases and heart failures. Once a person has suffered from heart failure, then the survival of the person is heavily dependent on the different clinical features of the person. Clinical features of a person include hormone levels, heart contraction rates, sodium levels etc. This report is trying to have an insight into the clinical features of a person and thereby trying to deduce the survival rates of person who have suffered heart failures.

Methodology

Data Preparation

The dataset for the assignment contains the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015 (Chicco & Jurman, 2020). The data is loaded using **pandas**. Once the data is loaded, head of the data (First 5 observations) was printed out and compared with CSV file to ensure data import is proper. Number of observations is also validated against the CSV data. Datatypes of all the columns were validated.

Following sanity checks have been carried out on different columns of the dataset:

Columns	Sanity Checks
All columns	Maximum and minimum values of the columns were checked with the related research paper and ensured all the column values are falling within the ranges. Missing values are also checked for all the columns.
Age	Floating ages were converted to integers.
Platelets	The platelets count in the CSV file is converted to kiloplatelets/mL to be consistent with research paper used for the assignment - Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone by Chicco and Jurman.

Feature Summary:

Column Names	Minimum	Maximum	Data Type	Type of Data
age	40	95	int64	numerical
anaemia	0	1	int64	categorical
creatinine_phosphokinase	23	7861	int64	numerical
Diabetes	0	1	int64	categorical
ejection_fraction	14	80	int64	numerical
high_blood_pressure	0	1	int64	categorical
platelets	25.1	850	float64	numerical
serum_creatinine	0.5	9.4	float64	numerical
serum_sodium	113	148	int64	numerical
sex	0	1	int64	categorical
smoking	0	1	int64	categorical
time	4	285	int64	numerical
DEATH_EVENT	0	1	int64	categorical

Data Exploration

Feature Exploration

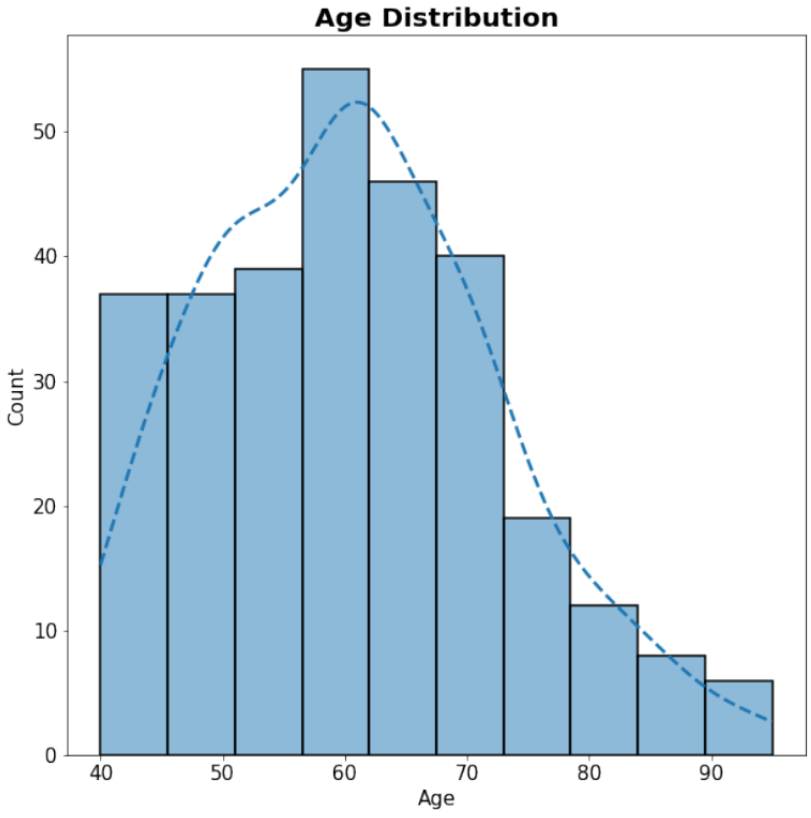

Figure 1: Age Distribution

According to Figure 1, most patients in the data are from the range 55 to 70 which can be observed from the graph in figure 1. The distribution is slightly skewed to right. From this we can understand majority of the heart failure patients are middle aged.

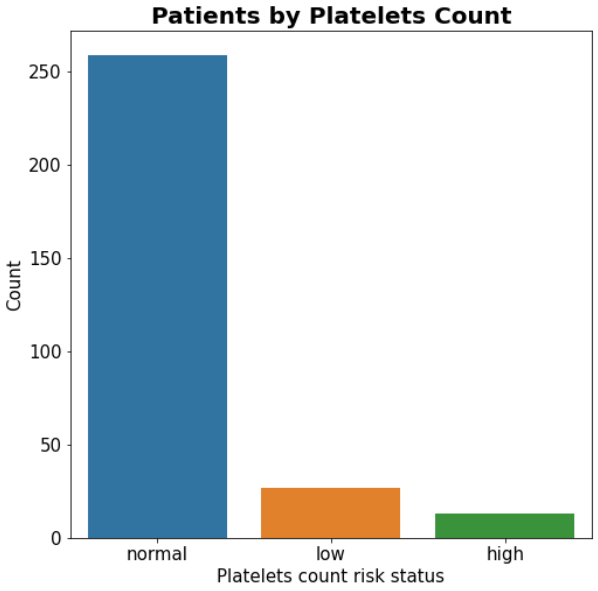

Figure 2: Platelets Count

The normal range of platelets counts is in the range of 150 to 450 kiloplatelets/mL (Johns Hopkins Medicine, 2021). Counts above or below these range low and high platelets count respectively.

It is clear from Figure 2 that majority of the sample patients (80% above) in the group are having normal platelets count. Number of patients with low platelet counts is almost double the patients with high platelet count which gives us an indication that platelet count may not be a variable of interest for the analysis for predicting the survival rate of heart failure patients. But as it is a clinical feature it will be considered for modelling part.

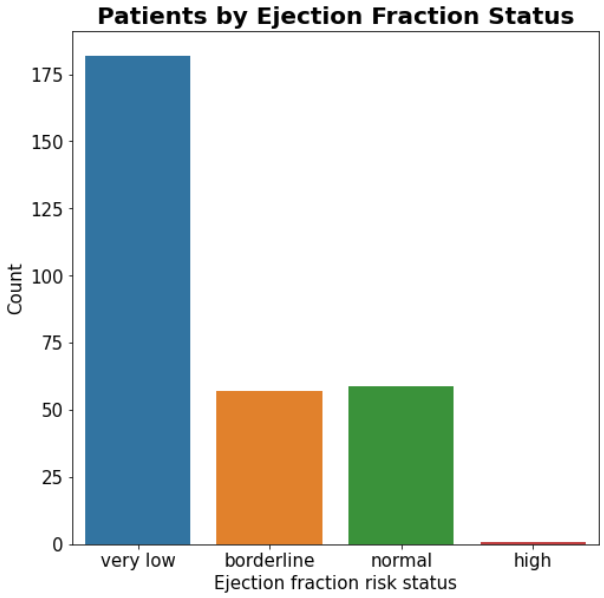

Figure 3: Ejection Fraction

The normal value of ejection fraction (which is expressed in percentage) lies in the range of 50% to 70%, while low is considered below 40%. The range between them i.e., 40% to 49% is considered as borderline while above 70% it can be categorized as high (Chicco & Jurman, 2020)

Figure 3 is saying that around 60% of the patients are having very low values for ejection fraction and leads us to the fact that ejection fraction has a huge impact on heart failure patients as this data set consists

of heart failure patients only. People with normal and borderline values are almost in equal numbers in the data set.

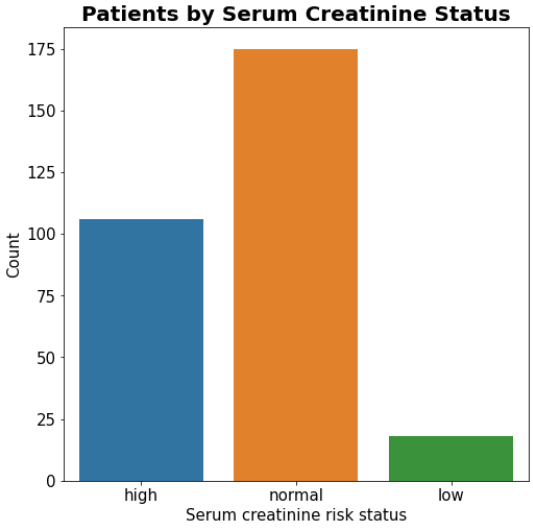

Figure 4: Serum Creatinine

The normal range for serum creatinine for male and female is 0.74 to 1.35 mg/dL and 0.59 to 1.04 mg/dL (Mayo Clinic, 2021).

From the visualisation graph, we can observe that 35% of patients are having high levels of serum creatinine and 58% are having normal levels. Very few are having low levels of serum creatinine. High levels of serum creatinine are also a primary reason for heart failures.

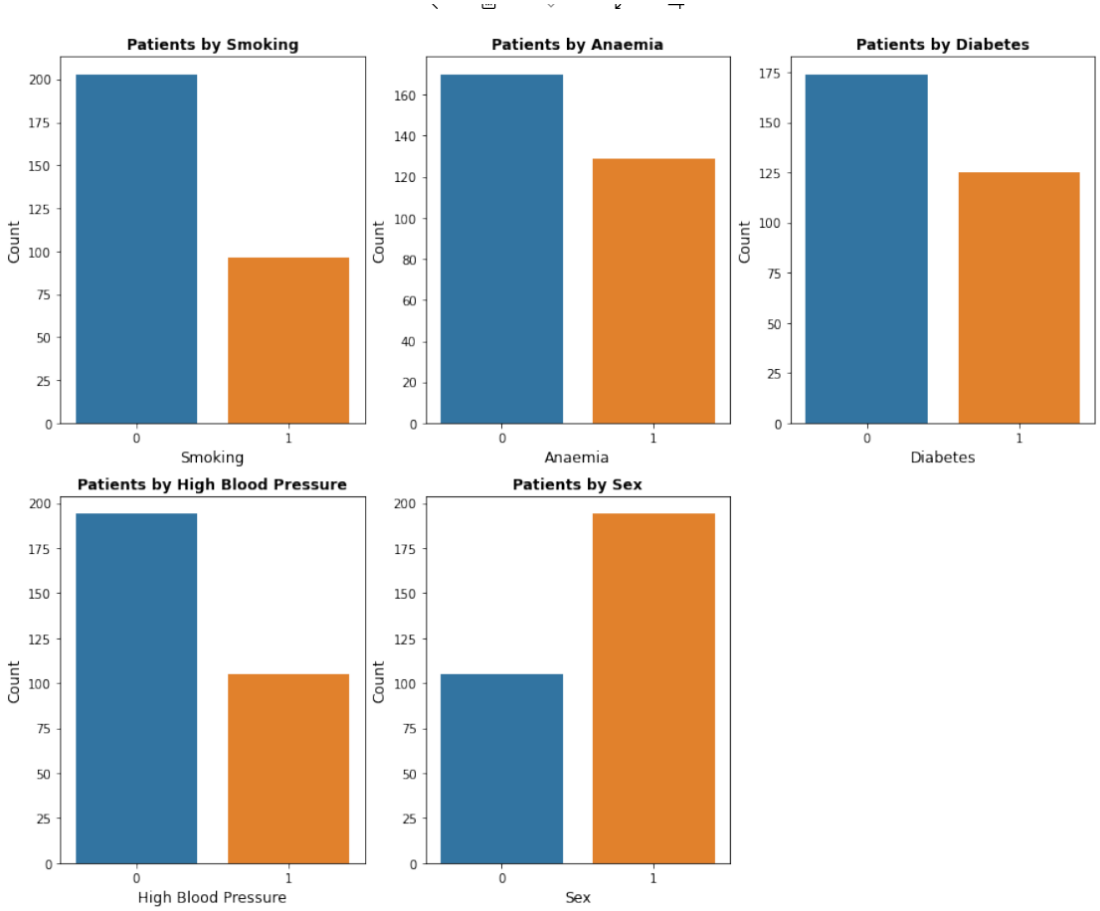

Figure 5: Categorical bar plots of Smoking, Anaemia, Diabetes, Blood Pressure and Sex

From these visualisations we can infer the following:

1. About majority of the patients are non-smokers, not anaemic, no high blood pressure, not diabetic

2. Majority are men in the data sample

Feature Relationships

Relationship between various categorical features and Deaths

Table in Figure 6 is showing categorization of persons died under various categorical features. 10 people who are dead due to heart failure are not suffering from diabetes, high blood pressure or smoker. All in this category are men and have anaemia. At the same time if we see the second row, we find 8 people dead are not having any of anaemia, diabetes, high blood pressure or smoker. These statistics are not leading us anywhere regarding the connection between these variables and deaths. Reading the observations in 3rd and 4th rows (almost complementary to each other), which is the third highest in terms of deaths, telling us these categorical variables are either not having any influence in deaths due to heart failure or importance of these variables in determining death by heart failures are less.

anaemia	diabetes	high_blood_pressure	smoking	sex	counts
1	0	0	0	1	10
0	0	0	0	1	8
1	1	1	0	0	6
0	0	0	1	1	6
1	1	0	0	0	6
0	0	0	0	0	5
1	0	1	1	1	5
0	1	1	1	1	4
		0	1	1	4
1	0	1	0	0	4
0	0	1	0	1	4
	1	0	0	1	4
1	1	0	0	1	3
0	0	1	1	1	3
	1	0	0	0	3
		1	0	0	3
1	0	1	0	1	3
0	1	1	0	1	3
1	0	0	1	1	3
	1	0	1	1	2
	0	0	0	0	2
0	0	1	0	0	2
	1	1	1	0	1
1	1	0	1	0	1
	0	1	1	0	1

Figure 6: Table showing death count by different categorical features

Blood Pressure and Sex

Hypotheses: Both males and females suffering from heart failure should be having high blood pressures.

Among the heart failure patients, it is seen that about 2/3rd of the males is having normal blood pressure whereas in case of females there is no comparable difference between normal and high blood pressure as seen from Figure 7. The plot is showing that the hypotheses we have assumed does not hold good for both males and females. Majority of males and females are having normal blood pressures. This is showing that high blood pressure may not be a good feature that can be relied upon for the prediction of survival rate of heart failure patients.

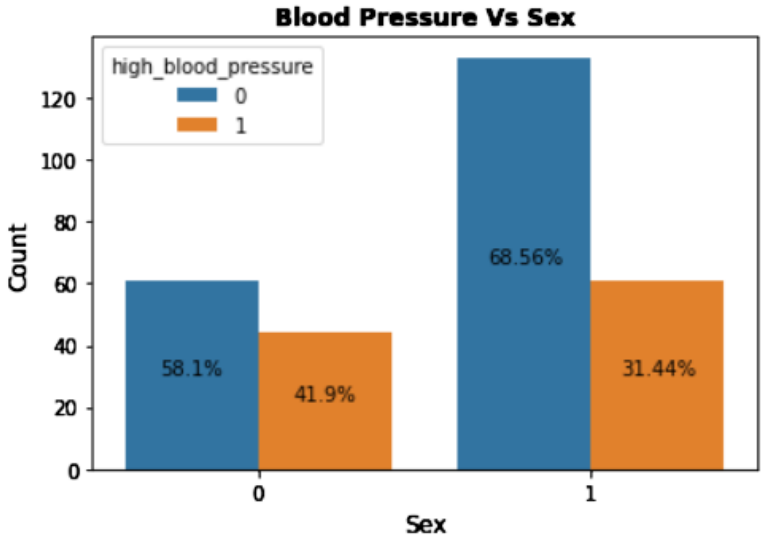

Figure 7: Blood Pressure vs Sex

Categorical features Vs Deaths

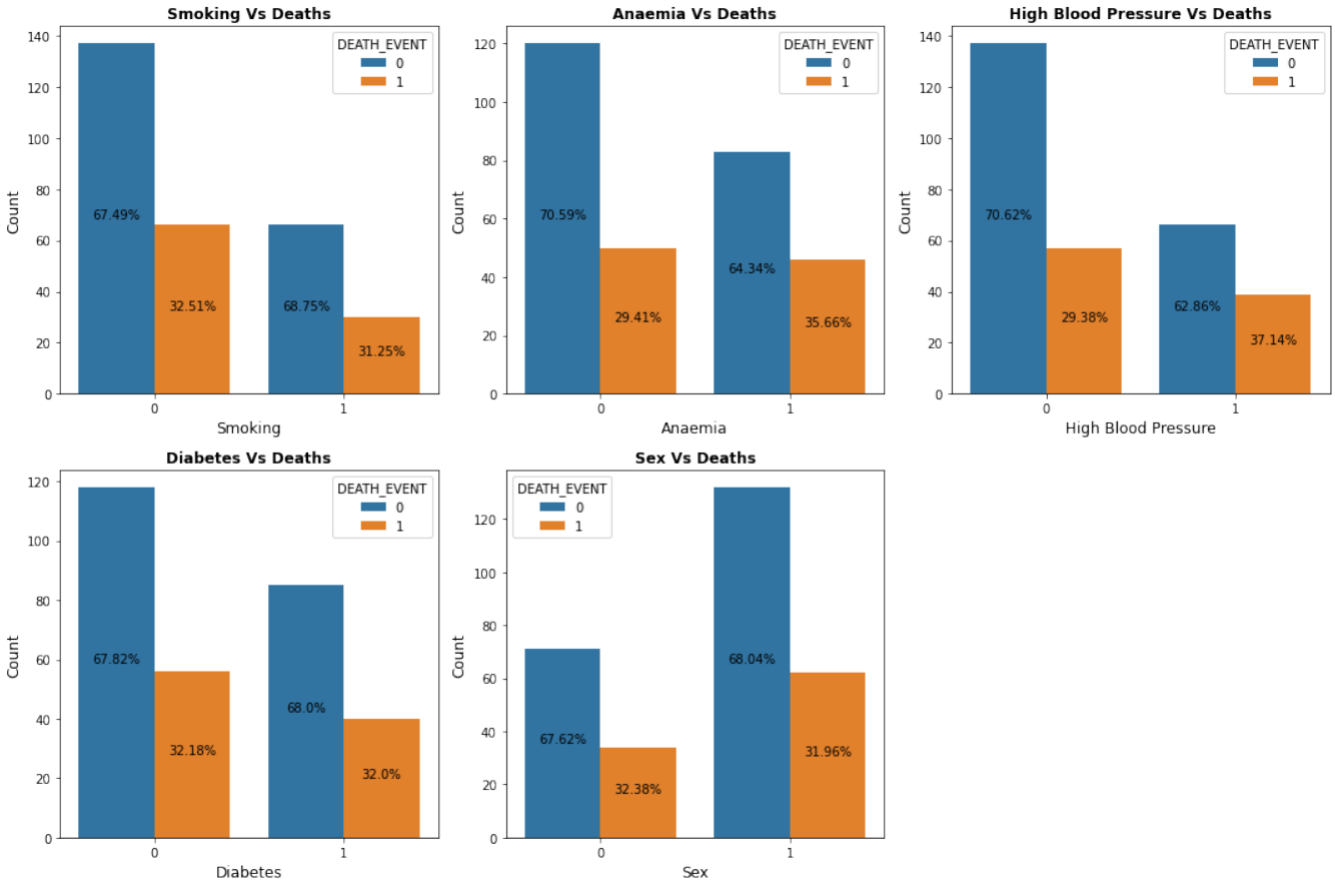

Figure 8: Smoking, Anaemia, High Blood Pressure, Diabetes and Sex Against Death Event

Hypotheses: Death rate should be more with people who smokes, are anaemic, are having high blood pressure, are diabetic irrespective of their sex.

From Figure 8, for the sample of heart failure patients available, death event is not showing any relationship between categorical variables mentioned here. The percentage split is almost the same for the positive and negative case (Death and survival). For example, among non-smokers survival rate is more than double the death rate which also holds good for smokers and all the other categorical variables. This indicates that the hypotheses we have formulated is not followed when we have visualised the data.

Serum Creatinine, Ejection Fraction Vs Deaths

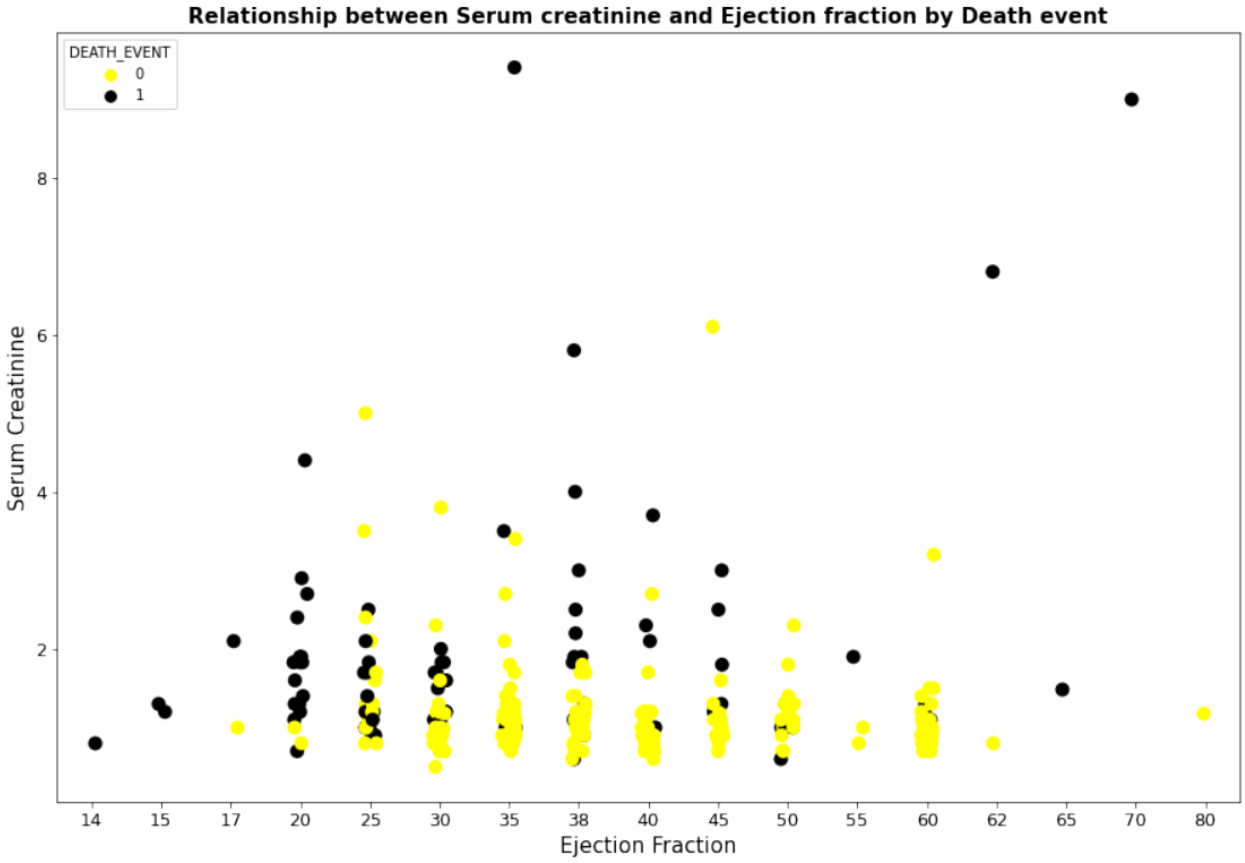

Figure 9: Serum Creatinine, Ejection Fraction Vs Deaths

Hypotheses: Death rate will be high with patients having lesser serum creatinine and ejection fraction

Deaths caused by heart failures are highly influenced by ejection fraction and serum creatinine. From the visualisation (Figure 9), we can see that, deaths are more concentrated when ejection fraction is lesser than 35 and serum creatinine less than 4. But, usually higher values of serum creatinine drives to death and the conclusion from here is, when serum creatinine is lower, deaths by heart failures is more affected by lower values of ejection fraction. When we move towards right side of the graph, we can see the black dots (deaths) are going to higher values and this indicates when the ejection fractions are high the death event is mainly caused by higher serum creatinine. The probability of death is most for persons with high serum creatinine and low ejection fractions. There are some outliers from the graph, but majority of the data points follows the same. The hypotheses we have formulated is partly correct for ejection fraction and partly wrong for lower values of serum creatinine.

Data Modelling

From the exploration of the different features of the data set it is found out that the clinical features are mainly helping in deducing the survival rate of the heart failure patients. We have treated the problem as a classification task and using the two classification algorithms – K-nearest neighbours and Decision tree.

Initially we have done the feature selection for the modelling by using the hill climbing technique. We have considered only the clinical features mentioned below as it was evident from the exploration task only the clinical features were more important. Clinical features – anaemia, 'creatinine phosphokinase', 'ejection_fraction', 'diabetes', 'high_blood_pressure', 'platelets', 'serum_creatinine', 'serum sodium'.

1. K-nearest neighbours

   For the application of the algorithm to the data set first we needed to do the feature selection process. So, we have adopted steep ascent hill climbing technique for feature selection from the list of the clinical features mentioned above. After the hill climbing algorithm, we have understood the best results are obtained by selecting 3 of the clinical features which can be used for the prediction.

   Once the hill climbing algorithm part is done, next the training and testing of the model with the KNN algorithm started. The KNN algorithm was performed with neighbours ranging from 2-6. We have considered uniform weights and minkowski metric with power 2. Reason for choosing these params for KNN is, the data set consists of only 299 observations, about 50% of the features are Boolean values (0 or 1). The p-value of KNN should be 2 here since if increased further it can heavily reduce the effects of the other features on the model and the model can become biased.

2. Decision Tree Classifier

   Another classification algorithm used was the Decision Tree. The following parameters value were selected.

   - criterion='entropy': The value of criterion determines how the data will be divided for the next node or in other words 'the quality of split'. The value entropy is selected for the information gain.
   - max_features='sqrt': This parameter determines how many features should we consider for further classifying the node. 'sqrt' means simply taking the square root of all the available features.
   - min_samples_split=6: This parameter defines the minimum number of data points that should be present inside the node for further split. This is to avoid overfitting of the model.
   - min_samples_leaf=3: The integer value assigned determines there won't be further division of the node and will be considered as a leaf node. This is to avoid overfitting of the model.
   - max_depth=None: This defines the maximum depth of the decision tree. Here, we have kept it as none because we have other constraints defined for min_samples_leaf and min_samples_split.
   - max_leaf_nodes=None: This means that tree can have unlimited number of leaves.

   As part of executing the decision tree algorithm we found that again, 3 features are mostly able to make good predictions for the survival rate of the heart failure patients by using the hill climbing algorithm. The decision tree formed, is given below.

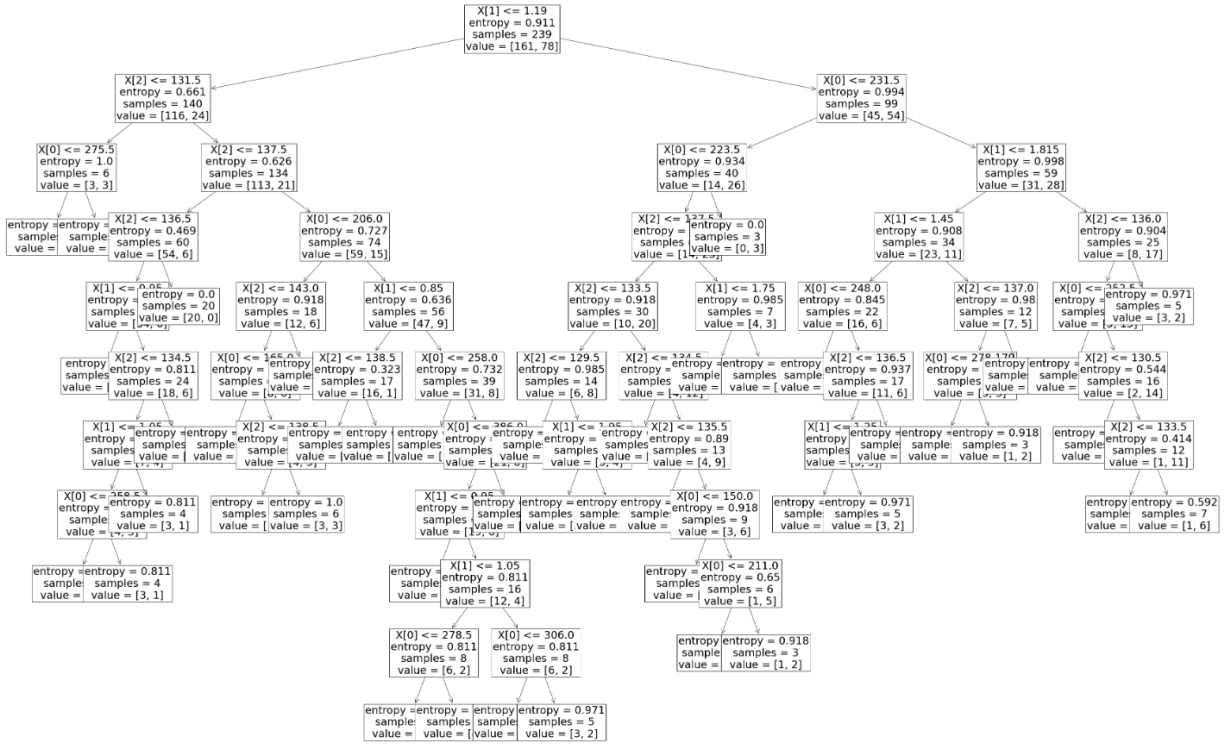

Figure 10: Decision Tree plot

Results

From the exploration of the data set provided, we have found that clinical features play a major role in determining the survival rate of the heart failure patients. On application of the different classification algorithms we could understand not all but some important clinical features are having more effect or have more predictability of the death event of the heart failure patients. The important features have been identified as serum_creatinine and ejection_fraction as per both the algorithms applied.

After the KNN algorithm was performed the following shows the classification report:

	precision	recall	f1-score	support
0	0.82	0.88	0.85	42
1	0.67	0.56	0.61	18
accuracy			0.78	60
macro avg	0.74	0.72	0.73	60
weighted avg	0.78	0.78	0.78	60

Figure 11: Classification Report for KNN

As a part of performing the algorithm, it is found that the clinical features 'ejection_fraction', 'high_blood_pressure', 'serum_creatinine' are the most prevalent in determining the survival rate of the heart failure patients.

After the iterative execution of the decision tree algorithm features 'ejection_fraction', 'diabetes', 'serum_creatinine' came to be the important features important in the prediction of the survival rate.

The following shows the classification report for Decision tree classifier:

	precision	recall	f1-score	support
0	0.75	0.76	0.75	62
1	0.44	0.43	0.44	28
accuracy			0.66	90
macro avg	0.60	0.59	0.59	90
weighted avg	0.65	0.66	0.65	90

Figure 12: Classification Report for Decision Tree

Discussion

KNN algorithm's classification report is healthier than the classification report of Decision tree algorithm on first look. KNN algorithm was giving more consistent results than compared with decision tree for the best features to be considered. As majority (about 6) of the features in the data set is Boolean (0 or 1) KNN would be a better choice for modelling. Homogeneity was not easier to be found using the decision tree classifier whereas KNN considering the neighbouring data points can be a better algorithm for this scenario. Both the algorithms have better prediction rate for survival rather than death.

KNN algorithm when performed multiple times on the data set with set parameters, the results obtained were all the same every time whereas when we ran the Decision tree classifier algorithm the results obtained were fluctuating between different features and hence it cannot be completely relied for a good prediction on an unseen data set.

Taking all these factors into consideration, it can be suggested that KNN based model is best for making the predictions for heart failure patients' survival.

Conclusion

There are lot of people in the world suffering from cardiovascular diseases and succumbing to death after heart failures. So, machine learning can be employed to predict the survival rate of the patients who have suffered heart failures.

From the analysis and modelling done as part of this project, we could find out some of the clinical features which have more influence on the heart failure person's survival rate. The two clinical features serum_creatinine and ejection_fraction are most important features in determining the survival rate of heart failure patients. By carefully monitoring these clinical features of the patients we can save some people from deaths. By employing machine learning techniques as mentioned in this report high risk patients (survival chance is less) can be identified at an earlier stage and proper treatment can be provided.

References

- Yongli, R 2021, 'COSC2670 -> Modules', PowerPoint slides & Pre-recorded lectures, COSC2670, RMIT University, viewed from 20th May to 28th May2021, < https://rmit.instructure.com/courses/79792/modules>
- Johns Hopkins Medicine 2021, *What are platelets and why are they important?*, viewed on 25th May 2021, < https://www.hopkinsmedicine.org/health/conditions-and-diseases/what-are-platelets-and-why-are-they-important#:~:text=A%20normal%20platelet%20count%20ranges,150%2C000%20is%20known%20as%20thrombocytopenia.>
- Mayo Clinic 2021, *Creatinine Tests*, viewed 25th May 2021, < https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646#:~:text=The%20typical%20range%20for%20serum,52.2%20to%2091.9%20micromoles%2FL)>
- 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>