HEART FAILURE PREDICTION

Exploration and investigation of heart failure using machine learning.

AUGUST 14, 2020

RMIT

Melbourne, Vic

Contents

1. Abstract

Heart failure is among the leading cause of death currently in the world. This study utilises data science visualisation methods to understand and validate the patient's conditions(features) and comparing them as the cause of death. Machine learning was also applied to predict the death events from the corresponding features which involve the application of the model classifier methods to validate the prediction from testing sizes of 50%, 40% and 20% respectively.

Broad consensus is evident in data science the use of data visualisation to understand data and comparing them appropriately. Appropriate model selection, and prediction can be executed from the sklearn classifier method estimate predictions and validate results.

The purpose of this study is to understand features(variable) leading to the cause of death of patients by data visualisation and to apply machine learning classification models to predict death event results as well as to validate them.

2. Introduction

The heart or cardiovascular disease generally is defined as the conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke (Heat Research Australia, 2018). Is chemic heart diseases and stroke account for about by 2016 making the world's leading cause of death. (WHO, 2016).

Heart failure is the state in which muscles in the heart wall enlarges and gets fade. As a critical organ essential for life. Heart failure study is a crucial subject to address causes of heart failure, management and improvement of the heart. Heart diseases, diabetes, high blood pressure, being overweight, high cholesterol and other diseases lead to heart failure.

Given the significance of the heart as a vital organ. Predicting its failure makes it a high priority to the patients, medical doctors, physicians and students with interest to the condition. This study helps us understand the if the several heart conditions listed in this report lead to death.

The purpose of this study is to utilise machine learning to understand each feature that may be characterised with heart failure and to examine the accuracy of machine learning methods to verify the heart conditions.

We hypothesised the appropriate pie, bar, boxplot, density and scatter graphs were favoured to visualise age, anaemia, creatinine phosphokinase, diabetes, serum creatinine, serum sodium, sex, smoking and time, to understand their distribution and comparisons. It was also hypothesised that the GaussianNB classification model was the most appropriate to determine the feature's likeliness to cause death by determining the model’s Accuracy, Confusion matrix, Classification error, Precision, Recall and F1-score.

3. Methodology

The study comprised of 299 patients of whom varied from the age of 40 to 95 years old. 194 of the patients were male and 105 were female. The follow up time was averaging 130 days, ranging from 4 days to 285 days. All the potential variables in table 1 Were considered valuable in predicting death caused by heart failure.

In this study, the heart failure clinical dataset was downloaded from the uci machine learning website, whereby the original was gathered by Government college university, Faisalabad, Pakistan. With the current being elaborated by Davide Chicco (Krembil Research institute, Toronto, Canada).

The dataset was downloaded and uploaded to jupyter notebook for analysis and prediction. The relevant pandas, NumPy, Seaborn, and matplotlib libraries were imported to perform the analysis and visualisation. For prediction, machine learning- sklearn libraries train_test_split, for training splitting and training models, Gausian for measuring similarity and KNN classifiers were imported.

Each variable was explored using the describe and value count functions, visualised using pie charts, bar graphs, density graphs and boxplots according to their more reasonable examining graphs.

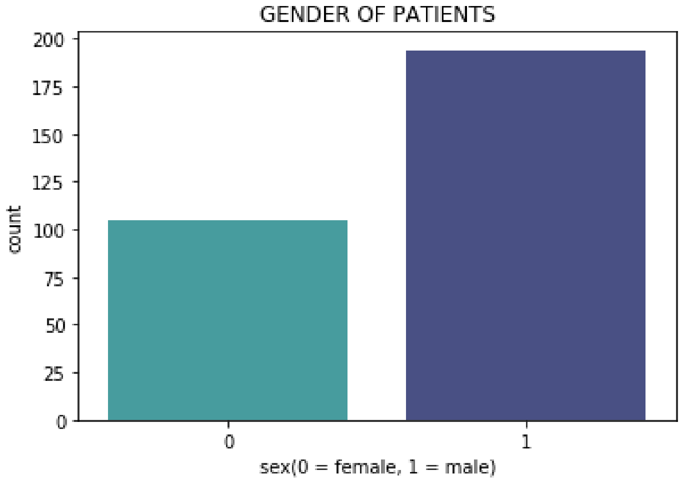
(a) Table1.

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,.., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0,1
High blood pressure	If a patient has hypertension	Boolean	0,1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	Mcg/L	[23,…,7891]
(CPK)			
Diabetes	If the patient has diabetes	Boolean	0,1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14,…,80]
Sex	Woman or man	Binary	0,1
Platelets	Platelets in the blood	Kiloplatelets/ml	[25.01,…,850.00]
Serum creatinine	Level of creatinine in the blood	Mg/dL	[0.50,…,9.40]

Serum sodium	Level of sodium in the blood	mE/ql	[114,…,148]
Smoking	If the patient smokes	Boolean	0,1
Time	Follow-up period	Days	[4,…,285]
(target) death event	f the patient died during the follow-up period	Boolean	0,1

4. Results

Figure 1.



There were 194 male and 105 female patients that participated in the study as presented in figure 1.

(a) Figure 2.

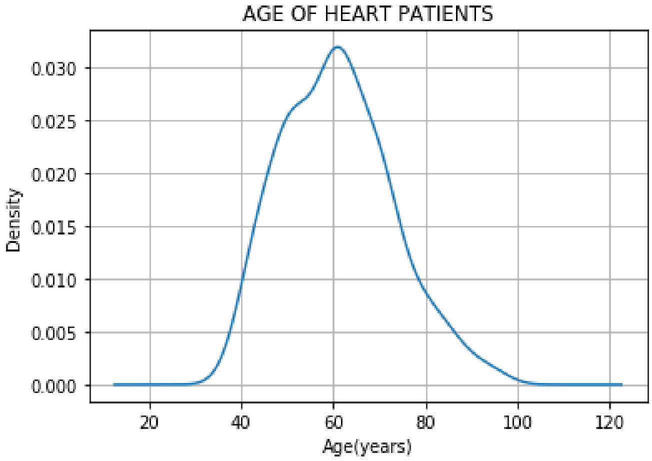


Figure 2 shows the range of patients age range from 40 to 95 with an average of 61 years old which is clearly illustrated.

(b) Figure 3.

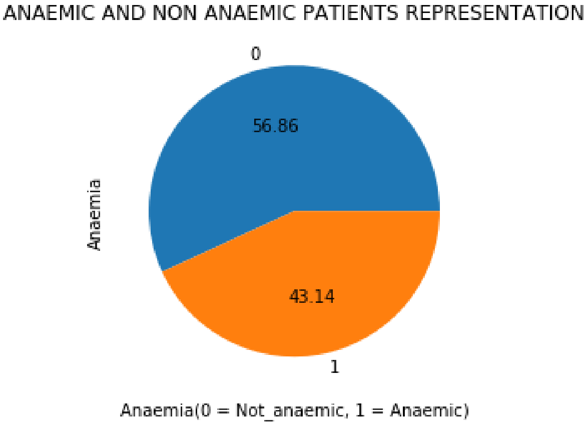


Figure 3 indicates that 43.14% of the patients were anaemic while the other 56.86% were not anaemic.

(c) Figure 4.

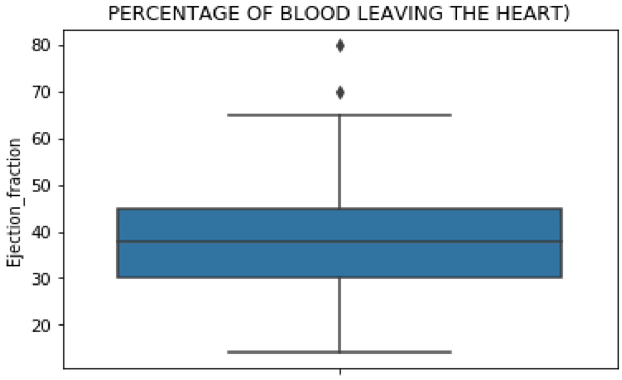
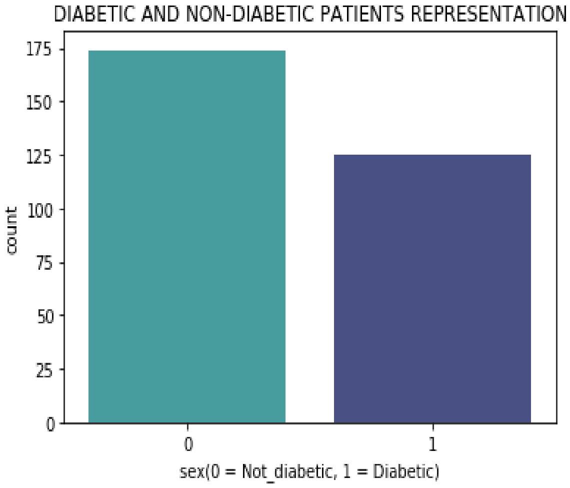


Figure 4 above shows that an average of 38% of blood leaving the heart while the minimum blood ejaculation was 14% and the maximum was 80%.

(d) Figure 5.



In figure 5, the boxplot shows the 174 nondiabetic patients against the 125 diabetic patients.

(e) Figure 6.

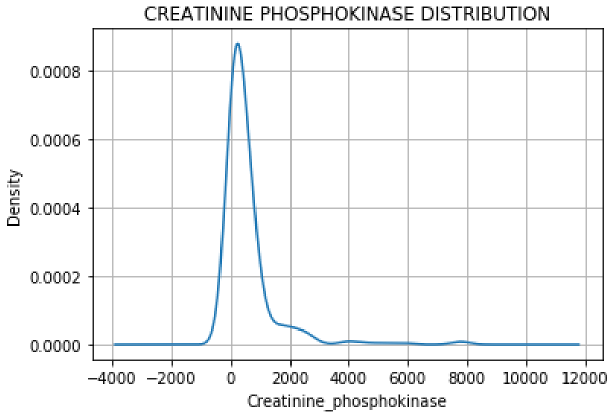
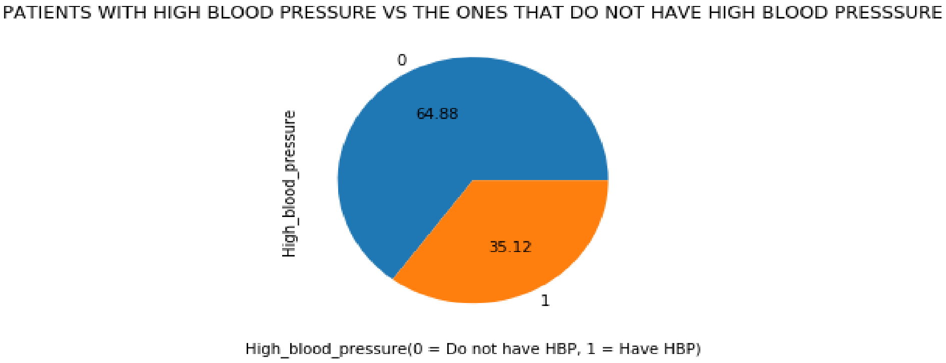


Figure 6 indicates the distribution of creatinine phosphokinase levels in patients.

(f) Figure 7.



The pie chart above represents composition of patients that have high blood pressure with about 35.12% and 64.88% that do not have high blood pressure

(g) Figure 8.

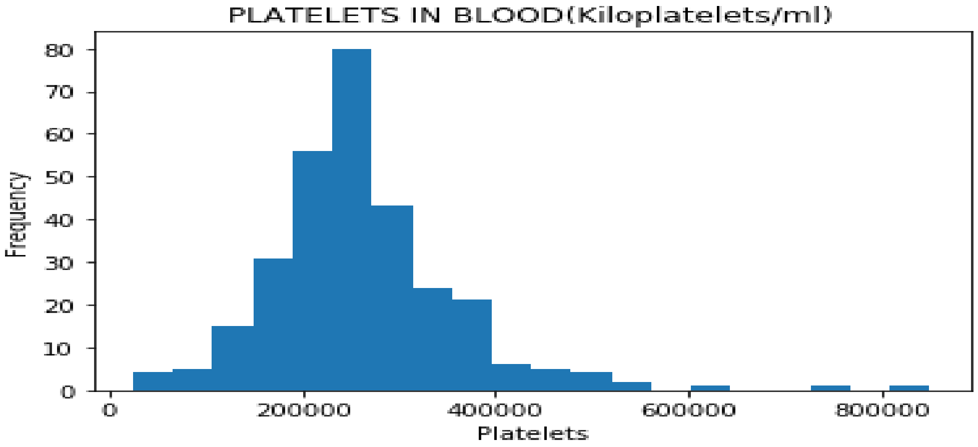


Figure 8 displays the platelets being nominally distribution in blood of all patients.

(h) Figure 9

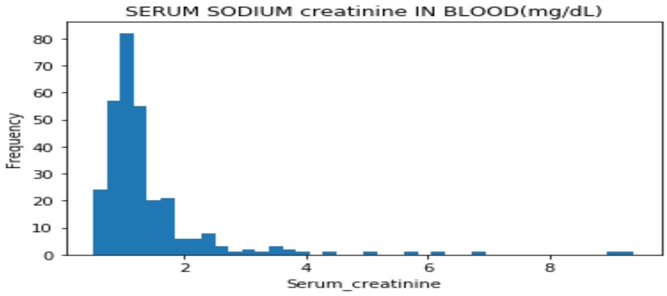
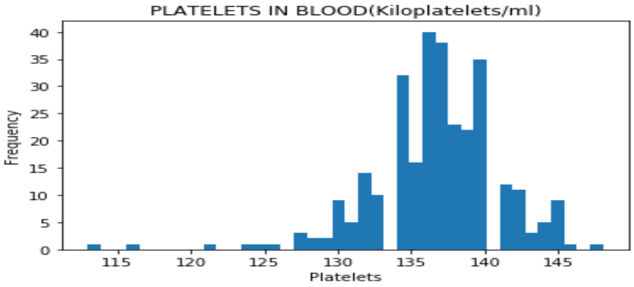


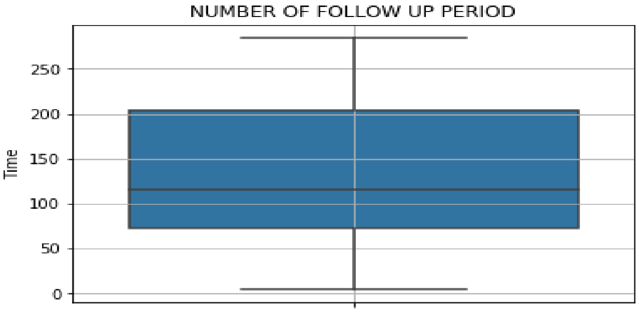
Figure 9 is serum creatinine levels distribution, skewed to the right. The average level is at 1.39 mg/dl, maximum at 9.4mg/dl and minimum at 0.5 mg/dl.

(i) Figure 10.



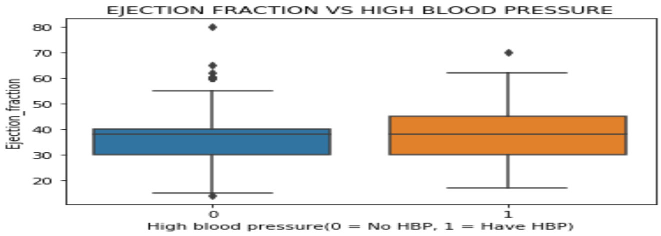
The bar graph presents the number of platelets frequency in blood of the patients in the study. The bar graph skewed to the left in appearance.

Figure 11.



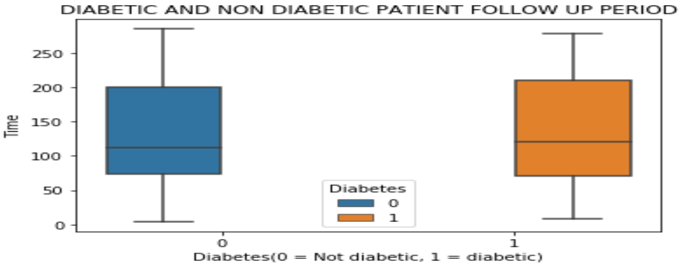
The above boxplot represents the patient follow up period in days. The average follow up time was 130 days, minimum was 4 days and maximum follow up period was 285 days.

(j) Figure 12



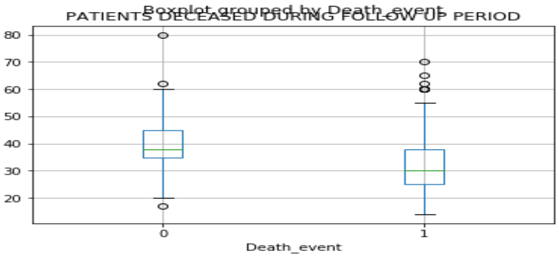
The boxplot in figure 15 indicates that more blood was ejaculated in patients with high blood pressure than for those with low blood pressure.

(k) Figure 13.



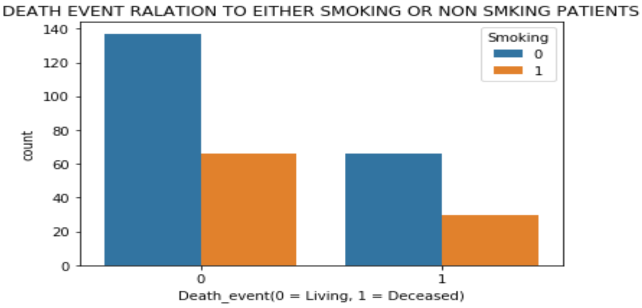
The boxplot in figure 16 above shows that both diabetic and non-diabetic patients had similar follow up periods.

(l) Figure 14.



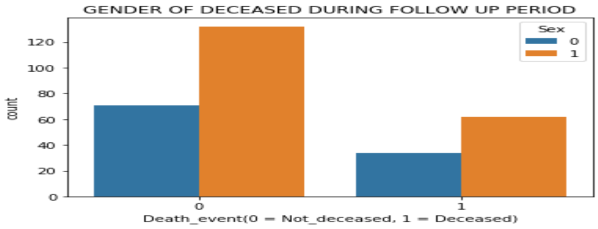
The boxplot above shows the comparison of the follow up period of the living and deceased patient.

(m) Figure 15.



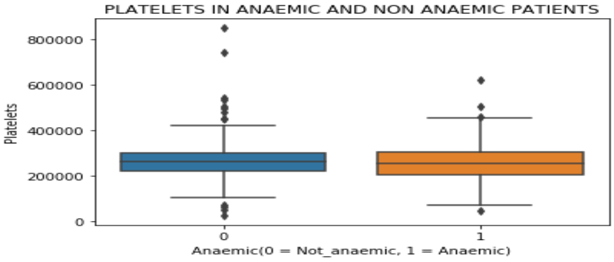
In the bar graph indicates that most of the patients most of the living patients were not smokers while the deceased had a history of smoking.

(n) Figure 16.



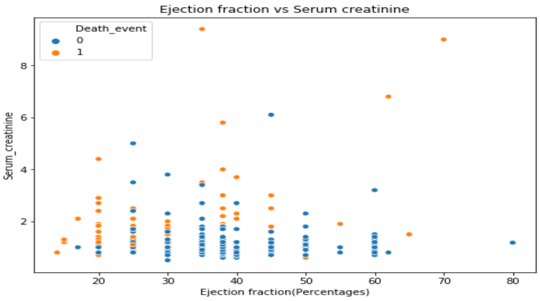
The bar graph above was used to compare the living and deceased patients by gender.

(o) Figure 17.



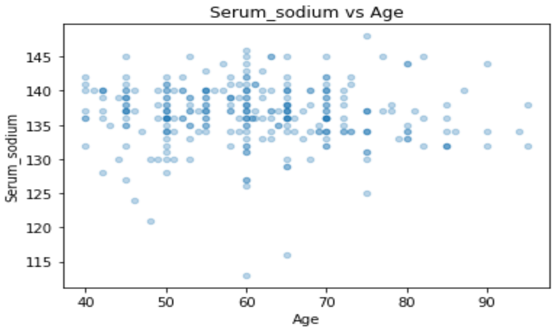
The boxplot comparison showed that there is not much difference in platelet amount in both anaemic and non-anaemic patients.

(p) Figure 18.



The scatterplot in figure 18 compares the numerical data of serum creatinine versus ejection fraction of blood of the patients.

(q) Figure 19.



(r) MACHINE LEARLING

For 50% training and 50% testing.

Prediction.

```
[0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 0

0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0

0 0 0 1 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0

0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0]
```

Accuracy	Classification error rate
79.3%	20.06%

Classification report

```
precision    recall  f1-score   support


0      0.77      0.97      0.86       95

1      0.90      0.49      0.64       55


accuracy                          0.79       150

macro avg      0.83      0.73      0.75       150

weighted avg      0.82      0.79      0.77       150
```

Confusion matrix.

```
[[92  3]

[28 27]]
```

For 60% training and 40 % testing.

Prediction

```
[0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 0

0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0

0 0 0 1 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0

0 0 0 0 0 0 0 0 1]
```

```
Accuracy.                                    Classification error rate.

76.66%                    23.33%
```

Classification report.

```
precision    recall  f1-score   support


0      0.72      0.99      0.83       70

1      0.96      0.46      0.62       50


accuracy                        0.77       120

macro avg      0.84      0.72      0.73       120

weighted avg      0.82      0.77      0.74       120
```

Confusion matrix.

```
[[69  1]

[27 23]]
```

For 80% training and 20% testing.

Prediction.

```
[0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 0

0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1]
```

Accuracy. Classification error rate.

73.33% 26.66%

Classification report.

```
precision    recall  f1-score   support


0      0.69      0.97      0.81       35

1      0.91      0.40      0.56       25


accuracy                        0.73       60

macro avg      0.80      0.69      0.68       60

weighted avg      0.78      0.73      0.70       60
```

Confusion matrix.

```
[[34  1]

[15 10]]
```

5. Discussion.

Each heart dataset variable and their comparison were explored appropriately using visualisation in python notebook and therefore confirms our hypothesis of using the right graphs to understand the features relation to leading to death. Categorical variables such as gender and diabetes were visualised using bar graphs as shown in figure1 and 5 respectively. The pie chart was also used to view proportion of patients with high blood pressure as shown in figure 7. The boxplot was the preferred graph to visualise comparison between numerical and categorical data of diabetic versus time and high blood pressure versus blood ejection fraction variables respectively as shown in figure 12 and 13. To compare two or more numerical data, the scatter plot graph clearly confirmed to be the preferred visualisation as show in figure 18 and 19 to compare serum creatinine versus ejection fraction and serum sodium versus age respectively. The GaussianNB from sklearn.naive-bayes was determined to be the appropriate machine learning classification model to measure and predict the features accuracy to causing death during heart failures.

Our results confirm our hypothesis in that the application of the appropriate visualisation generates better communication and understanding of the dataset. The right machine learning, classification model applied predicts using the data splitting method. The model's legitimacy was also confirmed by determining the sample test results by their accuracy, confusion matrix, classification error, precision, recall and F1-score. The greatest problem with the study was the limited number of patients and the age groups, younger patients than 40 years and more patients would have made the cases of the given features causing death more accurate.

In future, it is recommended that more patients from various regions to be tested for a longer period. Even then, the conclusion may be altered by age groups, regions and other pre-existing conditions/ diseases that the patients may have.

6. Conclusion.

Communicating results is a very important part of any study especially by using data science skills to investigate feature patterns that trigger the outcomes. Numerical data and their comparison are best understood using scatterplots, boxplots enable the visualisation of the categorical and numerical comparison as the bar graph was best used to understand categorical values.

The classification sklearn, model selection was used to select, train and split data to test 50%, 40% and 20% of the death event samples in relation to features. All samples were trained with the gnb.fit() model for prediction of events. The accuracy of 79%, 76% and 73% were determined respectively with validation by their classification error rate, confusion matrix and their classification report as per the result section of this study.

This study indicates that by considering the larger sample of 50% of the features for training, the results were more accurate and legitimate compared to the smaller samples. The credibility of the results does not suffer the obvious errors and can be improved larger samples for training and testing next time.

7. References.

archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository: Heart failure clinical records Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records# [Accessed 24 Aug. 2020].

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

In-text: (What is Heart Disease., 2020)

Your Bibliography: Heart Research Australia. 2020. *What Is Heart Disease.* [online] Available at: <https://www.heartresearch.com.au/heart-disease/what-is-heart-disease/> [Accessed 12 August 2020].

Worlld Health Organization, World Heart Day. https://www.who.int/cardiovascular_diseases/world-heart-day/en/. Accessed 8 August 2020.