# Data Science Techniques and Applications Coursework 01

Student: Rohan Khanolkar Student ID: 13199041

## Phase 1:

For the purpose of this project I have created an account in Kaggle (www.kaggle.com) and selected the [1] " Heart Failure Prediction " dataset published by the user Larex (a database grand master in Kaggle). This dataset is available on Kaggle from August 2020.

This dataset intends to show how cardiovascular diseases (CVD) leads to a heart failure which can result in increased risk of death. According to the World Health Organisation [3], around 31% of global deaths (i.e. 17.9 million people) occur due to CVD and these are manifested primarily as stroke or heart attacks. Usually, [2] when the heart is unable to supply or pump adequate amount of blood required by the body to meet its needs is known as 'Heart Failure' (HF). [4] HF is caused by conditions due to which heart muscles are overworked and damaged or advanced age, which leads to release of certain proteins and/or substances that have toxic effect on the flow of blood and eventually affect the heart. [4] Diabetes and High Blood pressure are some of the risk factors of heart failure.

This topic is of a personal interest to me as I had a close family friend who was recently detected with CVD and lost his life a few days after he received the reports. This topic would not be a part of my MSc. project, but I will use my learning from this coursework in my future endeavours.

This dataset seems to be having a similar degree of magnitude to the Iris flower dataset taught in the lab sessions. There is a published machine learning research paper using this dataset which can predict the survival of heart failure patients by D. Chicco and G. Jurman, G. [5] .

## Phase 2:

### Overview of dataset:

This dataset available on UCI website [2], Kaggle website mentions the dataset with 13 features out of which there is a target feature. There are not many details mentioned about the dataset on the Kaggle website, but the UCI website and research paper has a detailed description of the same.

(on next page)

| Sr. no. | Feature name | Range | Measurement scale | Type of feature |
|---|---|---|---|---|
| 1 | Age | 40 to 95 | Years | Numeric (Continuous) |
| 2 | Anaemia | 0 or 1 | Boolean | Categorical |
| 3 | High Blood Pressure | 0 or 1 | Boolean | Categorical |
| 4 | Creatine Phosphokinase (CPK) | [23 to 7861] | mcg/L (Micrograms per litre) | Numeric (Continuous) |
| 5 | Diabetes | 0 or 1 | Boolean | Categorical |
| 6 | Ejection fraction | 14 to 80 | Percentage (%) | Numeric (Continuous) |
| 7 | Sex | 0 or 1 | Boolean | Categorical |
| 8 | Platelets | 25.01 to 850.00 | Kiloplatelets/mL | Numeric (Continuous) |
| 9 | Serum Creatinine | 0.50 to 9.40 | mg/dL (Micrograms per decilitre) | Numeric (Continuous) |
| 10 | Serum Sodium | 114 to 148 | mEq/L (milliequivalents per liter) | Numeric (Continuous) |
| 11 | Smoking | 0 or 1 | Boolean | Categorical |
| 12 | Time | 4 to 285 | Days | Numeric (Continuous) |
| 13 | Death event (Target) | 0 or 1 | Boolean | Categorical (Target) |

Out of the 13 features there are five categorical features and seven continuous features along with one target variable which are narrated in the research paper. Features like high blood pressure, smoking, sex and diabetes are binary. A patient is said to have anaemia if the level of haematocrit levels were less than 36%. The original dataset does not give any information about the definition of high blood pressure. The Creatinine Phosphokinase (CPK) blood enzyme is an important factor to be considered inf heart failure, high level of CPK in blood increases chance of heart failure, [7] CPK spreads in blood as a result of a damaged muscle tissue. The ejection fraction is a [8] measurement of blood that is pumped out from the left ventricle of the heart for each contraction. [9] Creatine in the body generates waste product called serum which is used to checks the kidney function and as this serum increases it cause a renal dysfunction (i.e., kidney failure). Sodium is essential in every human, [10] it helps the appropriate functioning of nerves and muscles, so when the level of sodium decreases the chances of heart failure increases. The death event and time states whether the patient survived (death event = zero) or died (death event = 1) in the follow-up period of average of 130 days [6].

## Data Collection Background:

The dataset consists of medical records of 299 patients who suffer from heart failure. [6] This dataset was collected by T. Ahmad et al. between April to December 2015 at Faisalabad Institute of Cardiology, Pakistan along with allied hospitals nearby and this dataset was last updated on 24.04.2020. The UIC website cites two relevant papers about this, by T. Ahmad et al [6] and D. Chicco and G. Jurman, G. [5].

## Challenges / Task, Discussion and Activity:

There are two challenges proposed for this dataset, one being predicting heart failure which has almost 121 submissions and the other challenge is to create models of logistic regression and random forest which has less submission [1]. In the discussion forum, members have discussed about issues they faced when working on this dataset. Apart from that, there were 407 of solutions provided in the code section where the members have submitted various solutions and their point of views (this section had a greater number of member votes).

## Relevant Literature papers:

T. Ahmad et al. [6] are noted as original study investigators of this dataset published in the year 2017. In this paper there is a wide discussion on process of data collection and proposed algorithms towards predictions of heart failure. The UCI website refers to it as [2] “Original dataset version”.

The UCI website refers D. Chicco and G. Jurman [5]. as the “Current dataset version on the UCI ML Repository:” published in the year 2020. This paper give an in-depth analysis on how they selected the feature “serum creatinine “ and “ ejection fraction ” that can predict an accurate heart failure and it also mentions how their model can be encouraging in hospital and laboratory settings which can help doctors to predict patients survival just by analysing the two said features. This model is not implemented in any real-life situation, but I believe it can prove to be a potential step in saving someone’s life in future.

# Phase 3 :

## Dimensional Analysis:

The Heart Failure Prediction dataset is has 299 observations and 13 features. This dataset is an imbalanced dataset as the Target feature (i.e., Death Event) is not balanced (203 observations are "0" and 96 observations are "1").

To observe the dataset I have divided the its features into two types, numeric (continuous) and categorical – I have plotted a scatter matrix plot from Pandas and Heatmap from the Seaborn library for numeric (continuous) variables, for categorical variable, I have plotted column graphs.

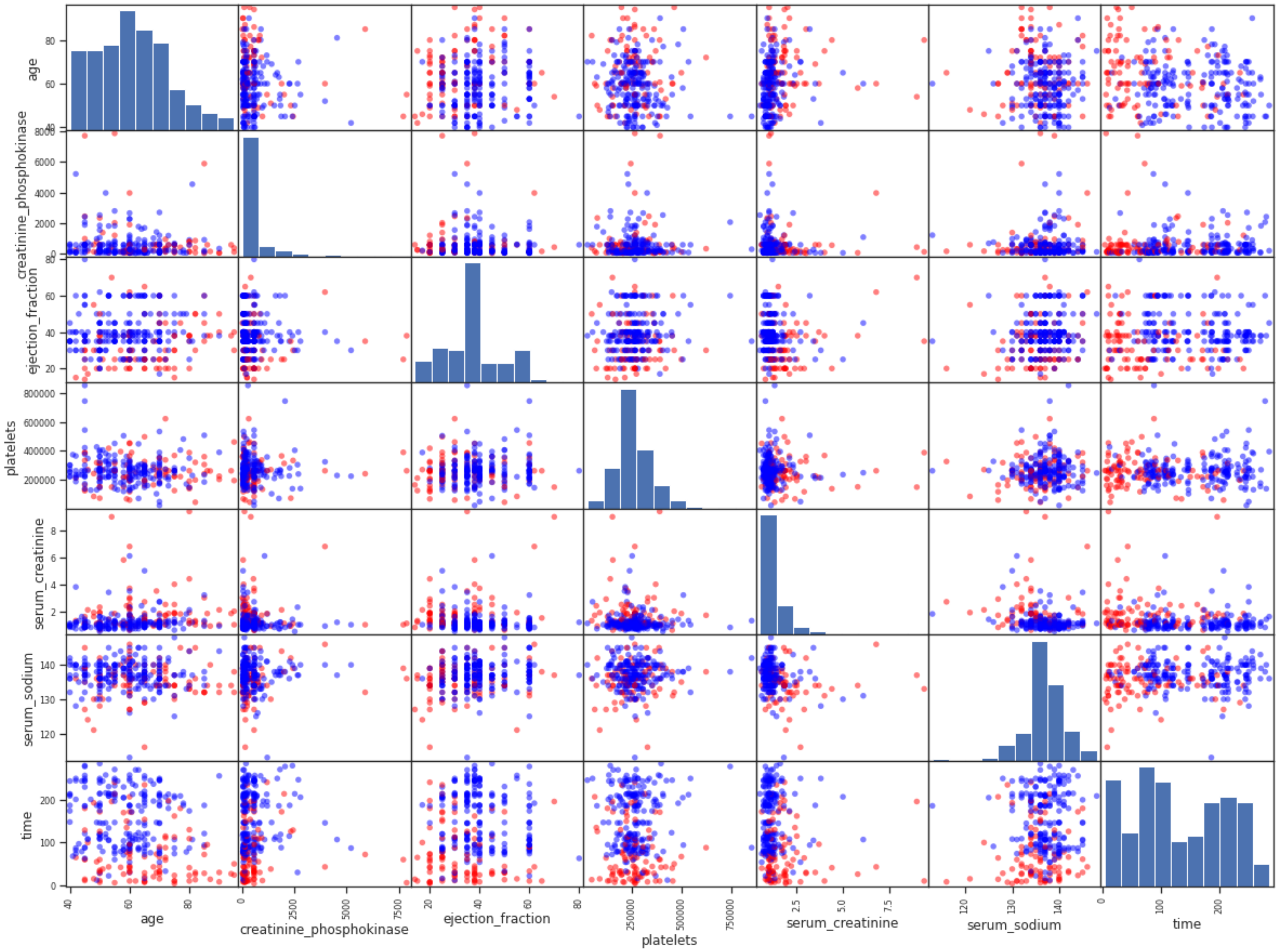
## Numeric (continuous features) Variables:

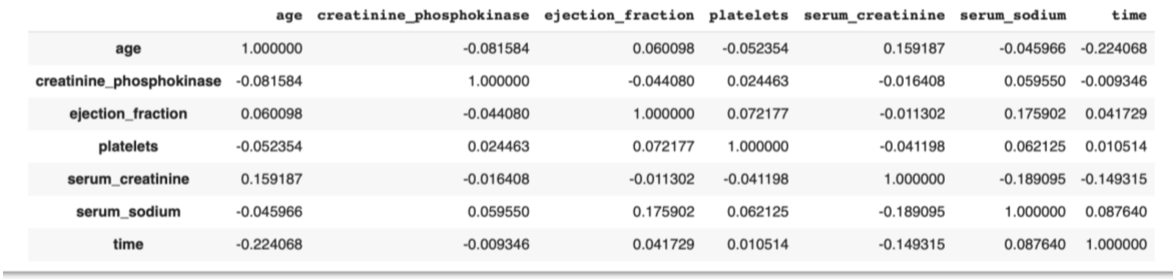
As mentioned above, to observe the numeric variables we have plotted the below scatter matrix (graph 1) and it do not show any clear separation of classes. Few plots looks somewhat showing separation:

- 'time' against 'age'
- 'serum_sodium' against 'serum_creatinine'
- 'time' against 'serum_sodium'
- 'time' against 'platelets'

(Plot on Next page)

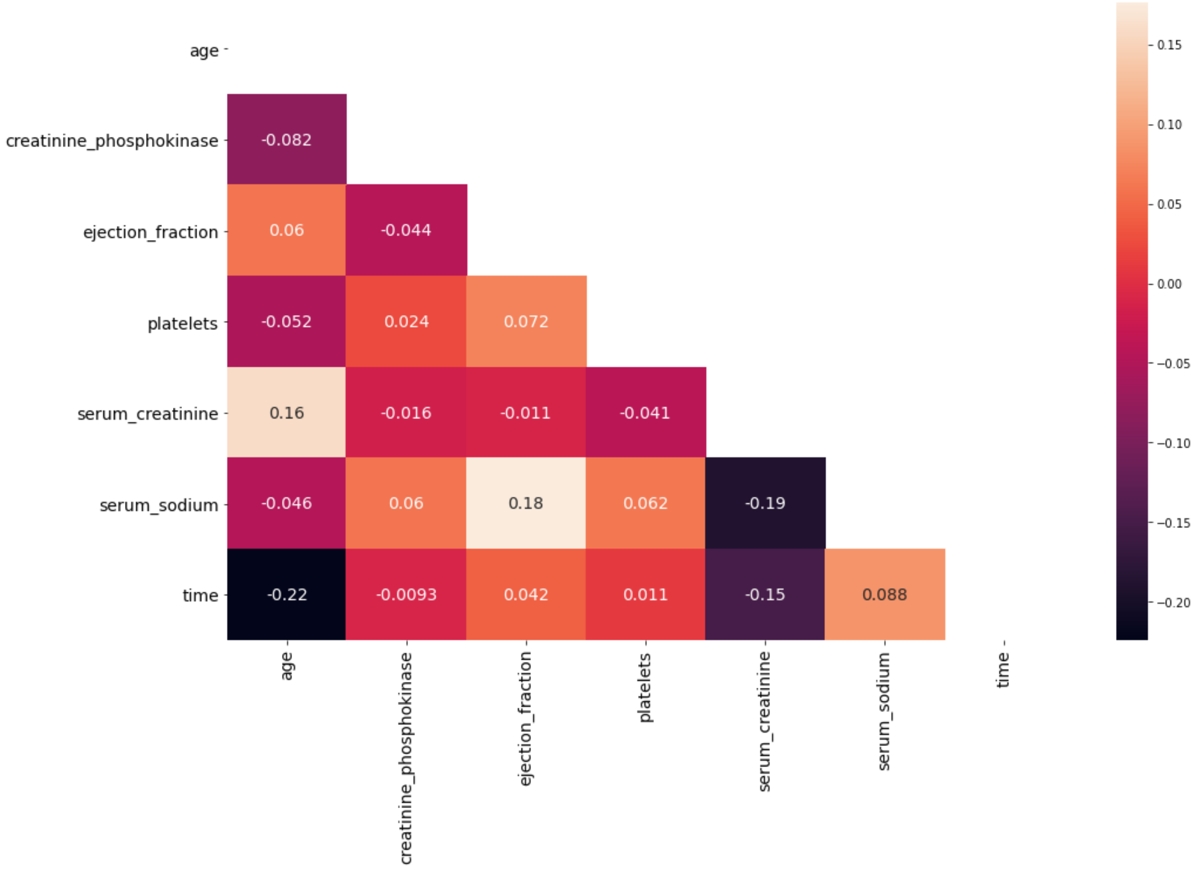
*Graph 1 : Scatter Matrix Plot (Numeric features)*





| | age | creatinine_phosphokinase | ejection_fraction | platelets | serum_creatinine | serum_sodium | time |
|---|---|---|---|---|---|---|---|
| **age** | 1.000000 | -0.081584 | 0.060098 | -0.052354 | 0.159187 | -0.045966 | -0.224068 |
| **creatinine_phosphokinase** | -0.081584 | 1.000000 | -0.044080 | 0.024463 | -0.016408 | 0.059550 | -0.009346 |
| **ejection_fraction** | 0.060098 | -0.044080 | 1.000000 | 0.072177 | -0.011302 | 0.175902 | 0.041729 |
| **platelets** | -0.052354 | 0.024463 | 0.072177 | 1.000000 | -0.041198 | 0.062125 | 0.010514 |
| **serum_creatinine** | 0.159187 | -0.016408 | -0.011302 | -0.041198 | 1.000000 | -0.189095 | -0.149315 |
| **serum_sodium** | -0.045966 | 0.059550 | 0.175902 | 0.062125 | -0.189095 | 1.000000 | 0.087640 |
| **time** | -0.224068 | -0.009346 | 0.041729 | 0.010514 | -0.149315 | 0.087640 | 1.000000 |

*Table 1: Correlation Matrix (Numeric Features)*



***Graph 2: Heatmap (Numeric Features)***

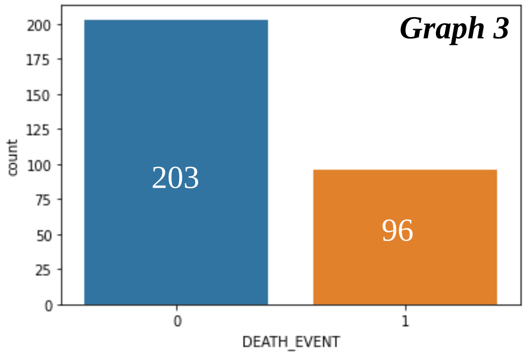
The above Correlation Matrix and Heatmap (by seaborn library) (graph 2 and table 1) show positive and negative corelation between variables :

- ‘ time ‘ and ‘age’ - They are negatively correlated as age of the patient do not relate to the number of days between the follow-up (days).
- ‘serum_sodium’ and ‘serum_creatinine’ – They are negatively correlated with each other because ‘serum_sodium’ is helps the function of muscles and ‘serum_creatininine’ is helps in the functioning of the kidney to flush all the toxins to the urinatory bladder and removes it from the body [11].
- ‘serum_sodium’ and ‘ejection_fraction’ – They are highly correlated with each other as the ‘ejection_fraction’ defines the percent of blood pumped out at each contraction of heart [8] , these contractions use the muscle tissue to pump the blood out of the heart. These function of muscle tissue are maintained by ‘serum_sodium’ [10].
- ‘serum_ creatinine’ and ‘age’ – It is a known fact that as the age of a patient increases the functioning of kidney decreases [9] and thus these variables are highly correlated to each other.

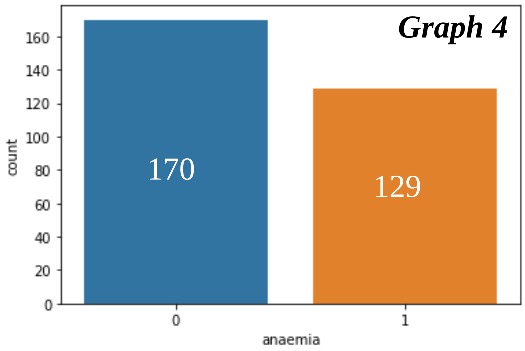
**Categorical Variables (features):**

The column graphs below narrates the discreet features of dataset which also include the target variable ‘DEATH_EVENT’ and its observations :

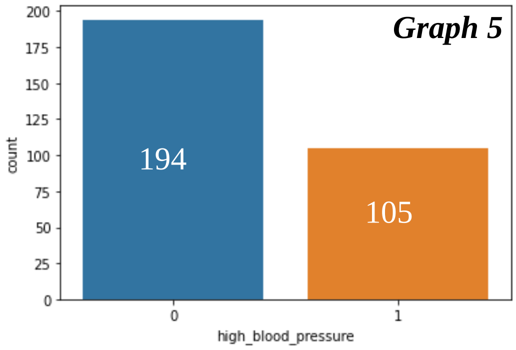
***Graph 3 – 8 : Bar Chart for Categorical Variables***



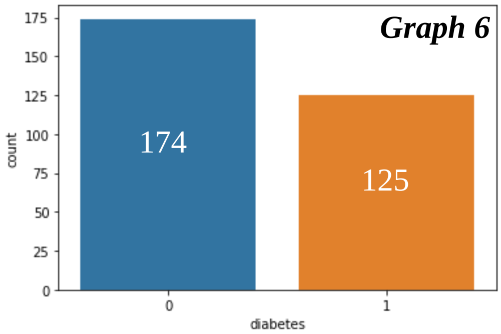
**'DEATH_EVENT'** – This feature is the target variable of the data set. There are 203 patient (marked in blue) who survived and 96 patient (marked in orange) who died a death event (please refer graph 3).This show that the data is imbalanced.



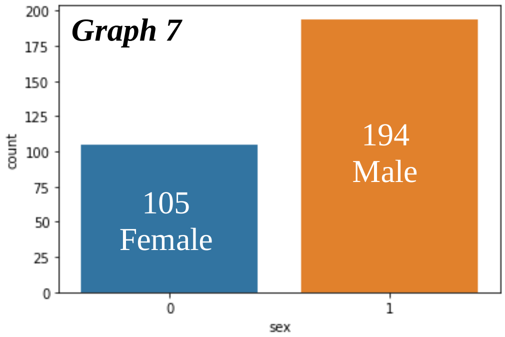
**'anaemia'** - A patient is considered as to have 'anaemia' if the haematocrit level in the blood are lower than 36% . [12] A haematocrit is a comparison of done during blood test, which examines the volume of red blood cells to the total volume of blood (i.e., it compares the plasma of the blood to the red blood cells) . There are 170 patient (marked in blue) who are not detected with 'anaemia' and 129 patient (marked in orange) who are detected with 'anaemia'(please refer graph 4).



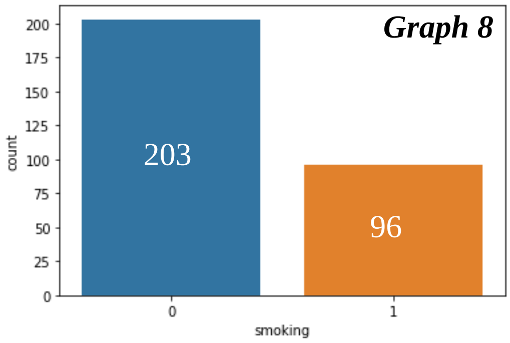
**'high_blood_pressure'** – A patient is considered as to have 'high_blood_pressure' when there is plaque accumulated in the coronary arteries of the heart [14]. There are 194 patient (marked in blue) who are not detected with 'high_blood_pressure' and 105 patient (marked in orange) who are detected with 'high_blood_pressure' (please refer graph 5).



**'diabetes'** - A patient is considered as to be 'diabetes' when their blood sugar is to high [14]. There are 174 patient (marked in blue) who are not detected with 'diabetes' and 125 patient (marked in orange) who are detected with 'diabetes' (please refer graph 6).



**'sex'** – There were 194 male patients (marked in orange) and 105 female patients (marked in blue) in the entire dataset (please refer graph 7).



**‘smoking’** - There were 96 patients (marked in orange) who are smokers and 203 patients (marked in blue) who are non-smokers in the entire dataset (please refer graph 8).

**Shortlisted Key Dimensions / Features:**

The D. Chicco and G. Jurman [5] paper discusses how they the ‘traditional univariate biostatistics analysis i.e., they used various different methods where they tested each feature of the dataset with the target variable. [5] Mann-Whitney *U* test, Pearsons correlations coefficient, chi squared and many more were used to shortlist key dimensions. They concluded the feature selection by selecting only two features, namely ‘eection_fraction ‘ and ‘serum_creatinine’ to further build machine learning models on it.

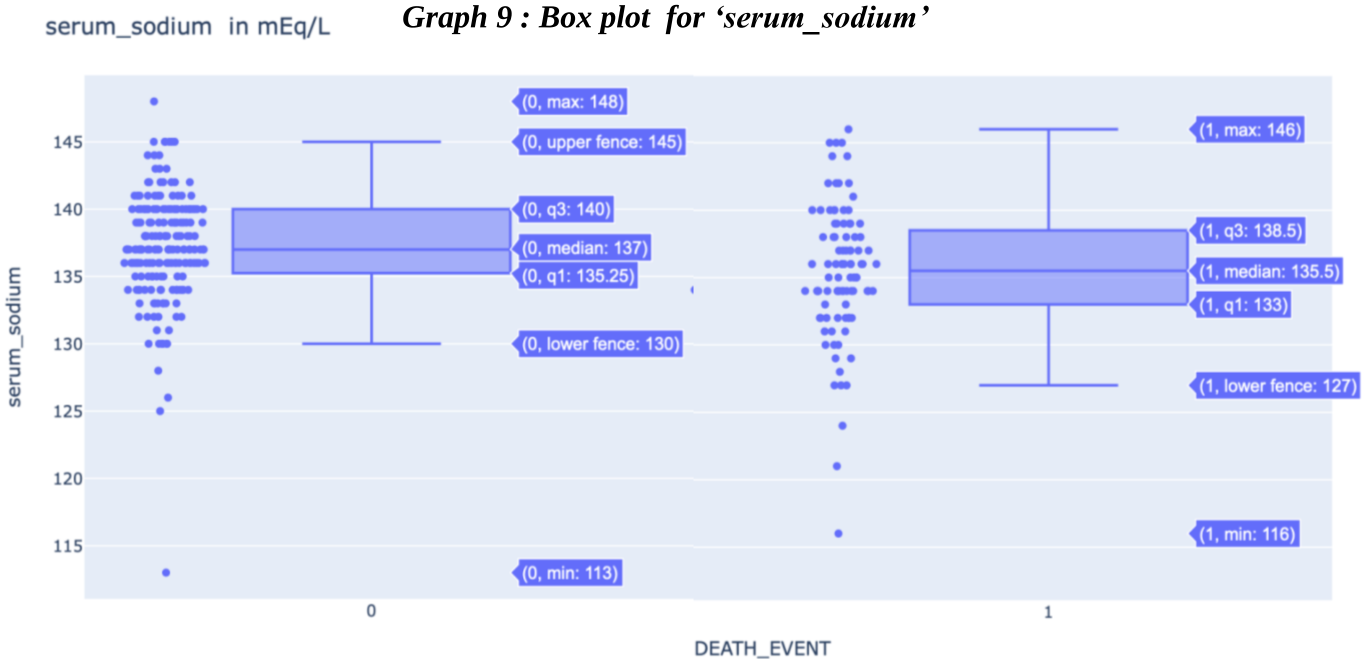
In our case we will be selecting the numeric (continuous) variables for further analysis as these features produced a most target separation in the scatter matrix plot and in the heatmap. The features are:

- ‘serum_sodium’
- ‘serum_creatinine’
- ‘ejection_fraction’

## ‘serum_sodium’

The D. Chicco and G. Jurman [5] discusses about this feature related to appropriate functioning of muscles, the mineral sodium can cause heart failure when found abnormally low in the blood.

*Graph 9 : Box plot for ‘serum_sodium’*



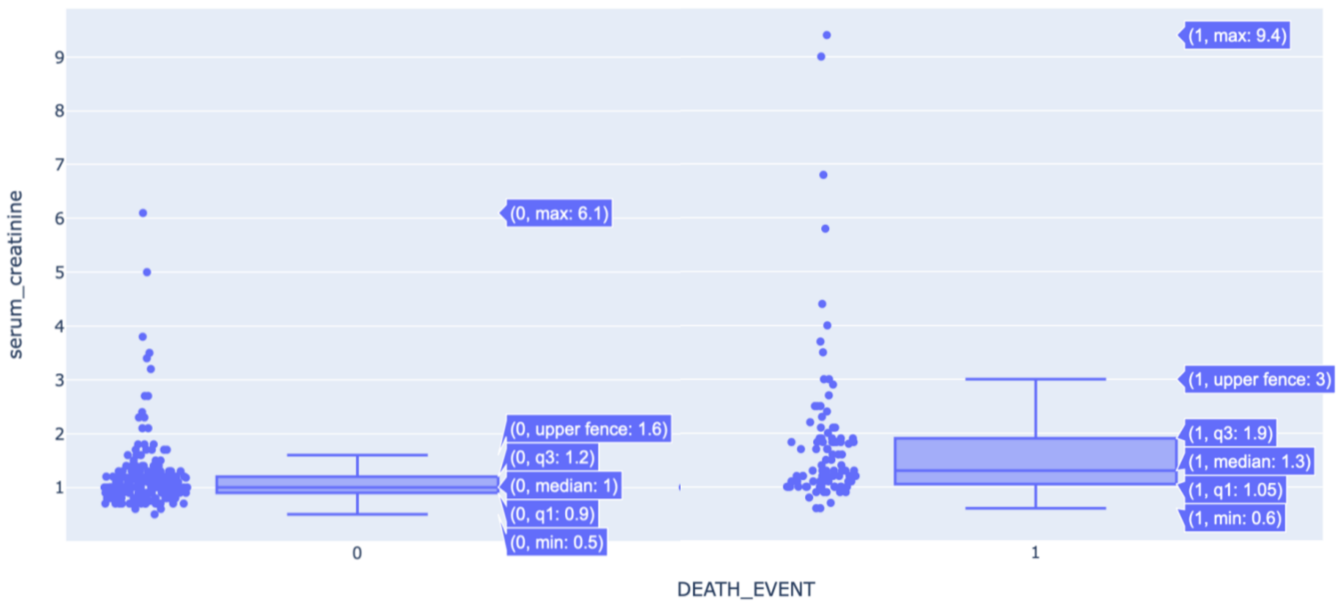
According to the [10], if the sodium level is upto 135 mEq/L or lower the body goes into hyponatremia which causes body swelling (with water content) or causes swelling (with water content) near the brain that can damages the cells and nerves. [10] Further mentions this swelling as an often problem within the older adult group.

Looking at the above box plot (graph 9) shows the range of ‘serum_sodium’ is between 113 – 148 mEq/L. The median of  survived patients (target 0) is 137 mEq/L which is marginally above to the required baseline of 135 mEq/L. On the other hand, the patient who did not survive (target 1) has a lower median along with larger ‘lower interquartile range’. This proves that the patients who died have a high possibility of having lower levels of ‘serum_sodium’ which could had led to heart failure. Thus this feature proves to be a choice of further investigation.

## ‘serum_creatinine’

The D. Chicco and G. Jurman [5] discusses about this feature as a waste product of creatinine and high level of serum creatinine can lead to misfunctioning of kidney. They also mentioned that this feature was one of the two features which they has selected in making the ML algorithm.

*Graph 10 : Box plot for ‘serum_creatinine’*

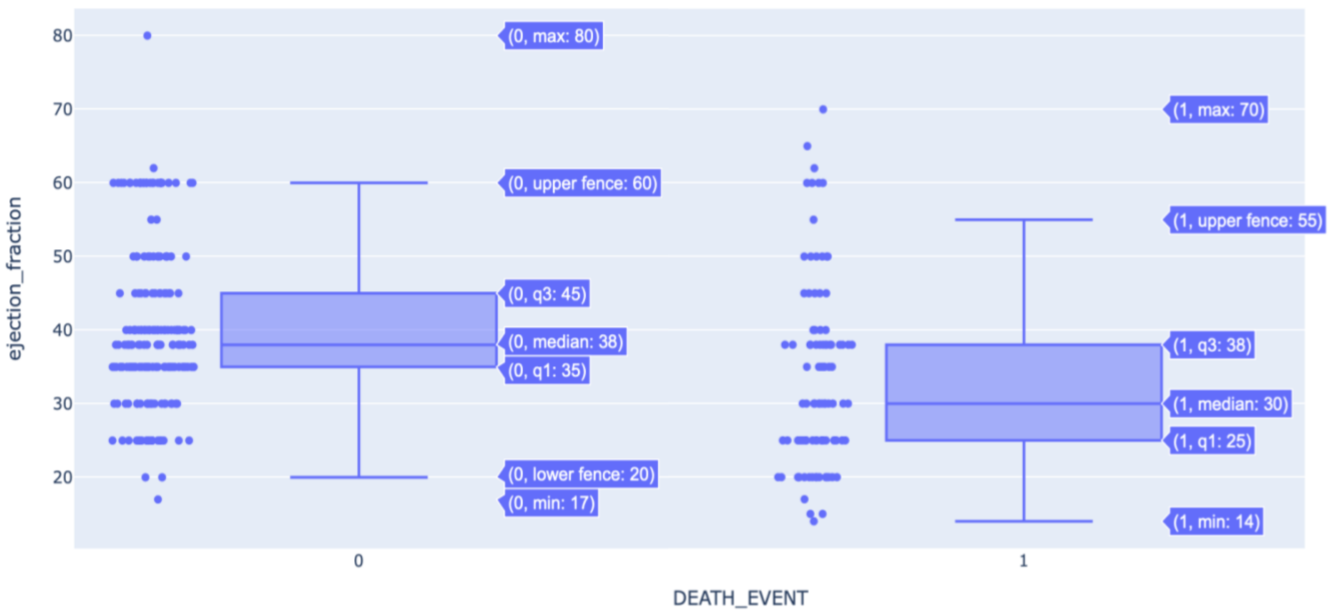


The box plot (graph 10) clearly shows that the patient who survived (target 0) had low level of serum creatinine in their blood. Whereas, the patients who died (target 1) had high interquartile range. This proves that the patients could have died due to high serum creatinine in their blood. Thus this feature proves to be a choice of further investigation.

## ‘ejection_fraction’

The D. Chicco and G. Jurman [5] discusses about this feature as a percent of blood pumped out from the left ventricle of heart per contraction. [5]They also mentioned that this feature was the second selected feature for developing ML algorithm. [8] States that the if the percent of ejection function is less than 45% it indicates heart failure issue, the normal range is between 52 – 73 % .

*Graph 11 : Box plot for ‘ejection_fraction’*



The box plot (graph 11) above show that the median of both the Death Event (target 0 & 1) is much lower than the indicated heart failure range. If we compare it with target 0 to target 1 we can note that target 1 is much lower than the target 0, thus we can say that the target 1 had lower ejection fraction which could have led to heart failure. Thus this feature proves to be a choice of further investigation.

### Quality of Dataset and Integrity Issues:

I preferred this Kaggle dataset over other because this dataset on because it was not having any missing values, the dataset in the UCI website is the same in Kaggle and seems to have no changes sone on it. Another reason for choosing this dataset was that it already had two published papers on it.

**Ambiguity of features:**

The features like ‘ Anaemia ‘, ‘ High Blood Pressure ‘ and ‘ Diabetes ‘ are mentioned as Boolean values. if these features were given a precise range the prediction on heart fail can be much better. There is no explanation in the manuscript of the dataset, why these values were recorded as Boolean.  Apart from this, this dataset do not provide any additional information about any primary kidney diseases and type of follow-up which was carried out on the patients.

**Discrepancies is data:**

The original dataset published In 2017 had a different names of a feature  ‘death event’ which was known as ‘event’ and in feature called ‘age’  few age were mentioned with a decimal value. Apart from that, in the paper of D. Chicco and G. Jurman [5] the feature ‘Sex’ is mentioned as binary (in its measurement column) whereas it is a Boolean (i.e., 0 or 1). These minor differences do seem to undermine the integrity of the overall dataset.

**Conclusion:**

This dataset and the selected features are showing no signs of discrepancies and prove to be a high quality dataset to investigate further. D. Chicco and G. Jurman [5] only selected ‘ejection_fraction’ and  ‘serum_creatinine’ features for building ML models, where as we can see that  ‘serum_sodium’ feature also looks promising to investigate further. Thus we selected three features along with the target variable (‘DEATH_EVENT’ feature) . I believe completing the second course work, which involve the use of Principle Component Analysis will give me an answer to the choice of the feature selection done in the paper by D. Chicco and G. Jurman. As we proceed towards coursework two I believe that this dataset will be within the range to for the dimensionality reduction and predict the Death Event in a patient suffering from heart failure.

# References:

[1] Larxel, “Heart Failure prediction.” . Available: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data/tasks?taskId=3070 . [Accessed: 15-Feb-2021].

[2] “UCI Machine Learning Repository: Heart failure clinical records Data Set,” *Uci.edu*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. [Accessed: 15-Feb-2021].

[3] “WHO | world heart day,” 2018. Avainalble: https://www.who.int/cardiovascular_diseases/world-heart-day/en/ . [Accessed: 15-Feb-2021].

[4] “Heart Failure,” Nih.gov. [Online]. Available: https://www.nhlbi.nih.gov/health-topics/heart-failure. [Accessed: 23-Feb-2021].

[5] D. Chicco and G. Jurman, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 16, 2020.

[6] T. Ahmad et al., “Survival analysis of heart failure patients: A case study,” no. 1, p. 8, Jul. 2017.

[7] “Creatine phosphokinase (CPK),” *Hopkinslupus.org*, 06-Jul-2009. [Online]. Available: https://www.hopkinslupus.org/lupus-tests/clinical-tests/creatine-phosphokinase-cpk/. [Accessed: 23-Feb-2021].

[8] M. Carroll, “Ejection fraction: Normal range, low, and treatment,” *Healthline.com*, 07-May-2018. [Online]. Available: https://www.healthline.com/health/ejection-fraction. [Accessed: 23-Feb-2021].

[9] E. Roth, “Creatinine Blood Test,” *Healthline.com*, 20-Aug-2012. [Online]. Available: https://www.healthline.com/health/creatinine-blood. [Accessed: 23-Feb-2021].

[10] C. Case-Lo, “Blood sodium level test: Purpose, procedure, and results,” *Healthline.com*, 29-Sep-2018. [Online]. Available: https://www.healthline.com/health/sodium-blood. [Accessed: 23-Feb-2021].

[11] M. Cantarovich, N. Giannetti, and R. Cecere, “Correlation between serum creatinine, creatinine clearance, the calculated creatinine clearance and the glomerular filtration rate in heart transplant patients,” *J. Heart Lung Transplant.*, vol. 21, no. 7, pp. 815–817, 2002.

[12] H. H. Billett, “Hemoglobin and Hematocrit,” in *Clinical Methods: The History, Physical, and Laboratory Examinations*, H. K. Walker, W. D. Hall, and J. W. Hurst, Eds. Chatswood, NSW, Australia: Butterworths, 2011.

[13] “How high blood pressure can lead to a heart attack,” *Heart.org*. [Online]. Available: https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attack. [Accessed: 23-Feb-2021].

[14] S. Watson, “Diabetes: Symptoms, causes, treatment, prevention, and more,” *Healthline.com*, 04-Oct-2018. [Online]. Available: https://www.healthline.com/health/diabetes. [Accessed: 23-Feb-2021].

## **Codes :**

```
from pandas import read_csv
import matplotlib.pyplot as plt; plt.rcdefaults()
import numpy as np
from matplotlib import pyplot
import plotly.express as px
from pandas import plotting
import pandas as pd
from pandas.plotting import scatter_matrix
import seaborn as sns
%matplotlib inline

# Loading the dataset and shape
filename = 'heart_failure_clinical_records_dataset.csv'
data = read_csv(filename)
print(data.shape)

# To find datatypes
types = data.dtypes
print(types)

# Dataset Discribtion
data.describe()

# To find if Target is balanced
class_counts = data.groupby('DEATH_EVENT').size()
print(class_counts)

#Scatter Plot Matrix for Numeric Variables
X =
data[['age','creatinine_phosphokinase','ejection_fraction','pl
atelets','serum_creatinine','serum_sodium','time']]
Y = data['DEATH_EVENT']
```

```
pd.plotting.scatter_matrix(X, c = Y, cmap = plt.cm.bwr,
figsize = [20,15],s=30, marker = '0')
plt.show()


# Numerical Variable Correlation Matrix
df = pd.DataFrame(data= data,
columns=['age','creatinine_phosphokinase','ejection_fraction',
'platelets','serum_creatinine','serum_sodium','time'])
df.corr()

# Seaborn Heatmap Correlation Matrix
mask = np.zeros_like(df.corr())
tringle_indices = np.triu_indices_from(mask)
mask[tringle_indices] = True
mask

plt.figure(figsize = (16,10))
sns.heatmap(df.corr(),mask= mask, annot=True,
annot_kws={"size": 14})
sns.set_style('white')
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.show()

# Plotting for  categorical variables features
g = sns.countplot(data['DEATH_EVENT'])#checking for class
imbalance
g.set_xticklabels(['0','1'])
plt.show()
print(data.DEATH_EVENT.value_counts())

g = sns.countplot(data['anaemia'])#checking for class
imbalance
g.set_xticklabels(['0','1'])
plt.show()
print(data.anaemia.value_counts())

g = sns.countplot(data['high_blood_pressure'])#checking for
class imbalance
g.set_xticklabels(['0','1'])
plt.show()
print(data.high_blood_pressure.value_counts())

g = sns.countplot(data['diabetes'])#checking for class
imbalance
g.set_xticklabels(['0','1'])
plt.show()
print(data.diabetes.value_counts())
```

```
g = sns.countplot(data['sex'])#checking for class imbalance
g.set_xticklabels(['0','1'])
plt.show()
print(data.sex.value_counts())
g = sns.countplot(data['smoking'])#checking for class
imbalance
g.set_xticklabels(['0','1'])
plt.show()
print(data.smoking.value_counts())


# Plotting Box Plot
fig = px.box(data, x='DEATH_EVENT', y='serum_sodium',
points="all")
fig.update_layout( title_text="serum_sodium  in mEq/L")
fig.show()

fig = px.box(data, x='DEATH_EVENT', y='serum_creatinine',
points="all")
fig.update_layout(title_text="serum_creatinine  in mg/dL")
fig.show()

fig = px.box(data, x='DEATH_EVENT', y='ejection_fraction',
points="all")
fig.update_layout(title_text="ejection_fraction  in Percentage
% ")
fig.show()
```

**The Academic Declaration :**

**“I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software.”**