Data Science Techniques and Applications Coursework 02

Student: Rohan Khanolkar Student ID: 13199041

Phase 1

As concluded in coursework one, we can see that the aim is to analyze the " Heart Failure Prediction " dataset. This data set is available on both Kaggle [1] and UCI websites [2] and this has 13 features and 299 observations of patients .

In this second coursework we will use the three selected dimensions from the previous coursework:

- **'serum_sodium'**

  The serum sodium actually relates to sodium (mineral) in blood that helps in functioning of muscles in body. When this mineral is abnormally low in the blood, it can lead to heart failure event [3]. This variable has no observation which are negative values and it ranges from 100 to 140, thus we can say that this variable do not comply with Gaussian distribution.

- **'serum_creatinine'**

  The serum creatinine is a waste product of creatinine which can lead to misfunctioning of a person's kidney [3]. This variable has a range of 0 to 9 with a median of 1.

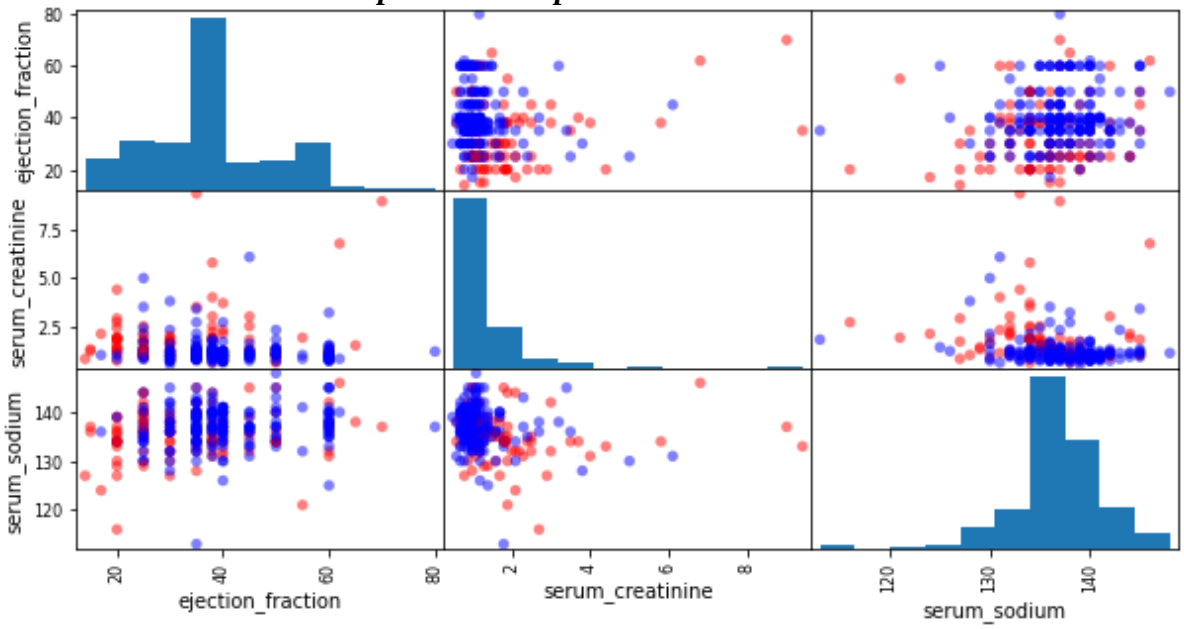
- **'ejection_fraction'**

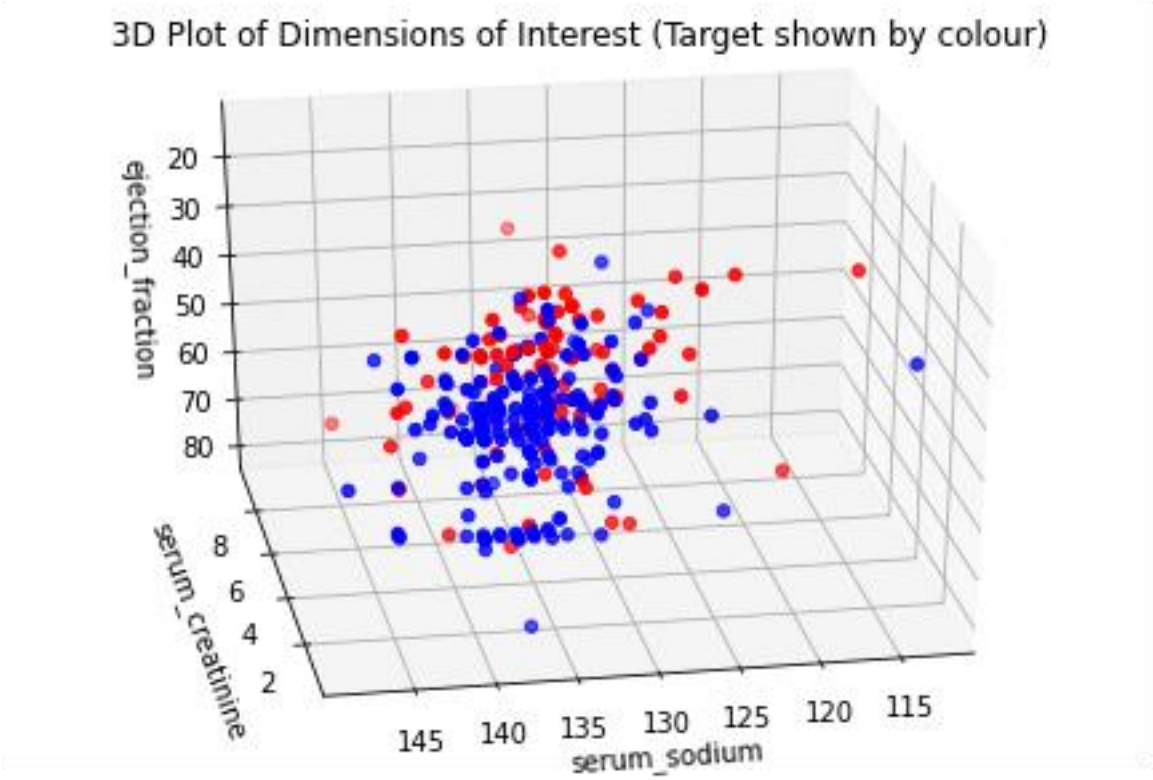
  The ejection fraction is the amount of blood pumped (in percentage) per heart contraction from the left ventricle of heart. The normal range of heart is between 52% - 73% and if this reduces to 45% or lower, then there is a high chance of heart failure issue.

These three continuous variables have been chosen as it seems to give a separation of class compared to the target class. The Pandas scatter matrix plot can be seen in Appendix 1 which consist of all continuous variables in the dataset. These three continuous variables also showed how their increase or decrease can affect a persons heart and it can lead to a death event. The Appendix 2 shows the box and whisker plots which were compared to the Death_Event target variable.

The three chosen continuous variables are classified into a simple three dimension scatter plot matrix and its coloured target class observations. You can note that the there is no clear separation of target class in the row data.

Graph 1: Scatter plot matrix





Phase 2

As per the coursework guidelines, the three chosen continuous variable (dimensions) ejection_fraction, serum_creatinine and serum_sodium are to be carried out with principal component analysis (PCA). The coursework guidelines also mention to use python program (attached as separate .py file) to reduce the three continuous variable along with target

variable from the dataset using the Scikit-Learn functions [4] and using the Gaël Varoquaux model [5] specified for this coursework.

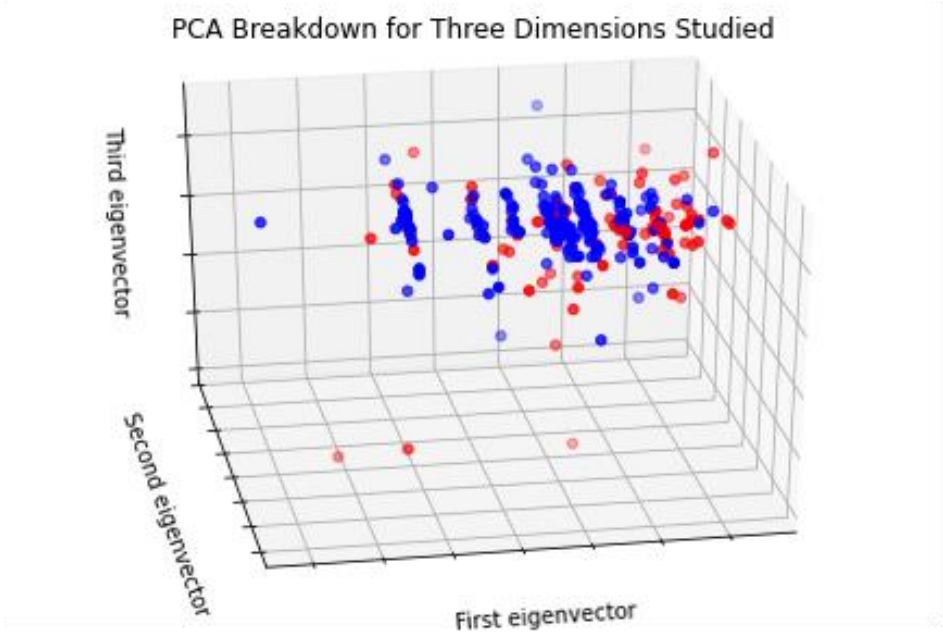
In this coursework we have not chosen to scale the data, as the dataset do not contain high number of datapoints which could be reduced to sample every fractions of dataset points. Also note that, scaling (normalisation) is applied to scale the data to the standard deviation of 1 and the mean of 0 which avoids.

	ejection_fraction	serum_creatinine	serum_sodium
count	299.000000	299.00000	299.000000
mean	38.083612	1.39388	136.625418
std	11.834841	1.03451	4.412477
min	14.000000	0.50000	113.000000
25%	30.000000	0.90000	134.000000
50%	38.000000	1.10000	137.000000
75%	45.000000	1.40000	140.000000
max	80.000000	9.40000	148.000000

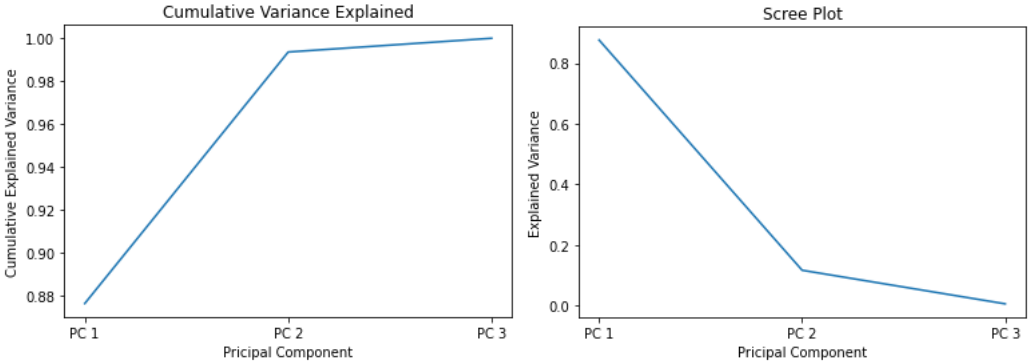
We can note from the above picture that the serum_creatinine is not at normally distributed variable. Thus we have not chosen the scale the data.

Phase 3

As we can see from the below image of PCA Breakdown for Three Dimensions, the 3D scatter plot seems to generate a clearer separation of class than the 3D scatter plot generated from the raw data. You can also note that there remains some intermixed relativity between the classes.



Please find below the plot for cumulative variance explained and Scree Plot which shows that the first principal component has variance of significant proportion in the dataset and the remaining variance of the dataset are explained by initial two principal components.



Apart from that kindly find the below the weightings of each dimension principal components as shown in the scikit learn [4].

As we can see in the below table in PC 1 ‘serum_sodium’ is the most significance from the other two which could be likely due to high standard deviation. In PC 2 ‘ejection_fraction’ has the higher negative and it can be relatively said that the risk of heart failure is more due to this factor and it will also show a negative correlation with the serum_sodium and serum_creatinine. In PC 3 ‘serum_creatinine’ has highest weights as compared to PC 1 and PC 2, this can be due to the lower standard deviation of the variable.

	‘serum_sodium’	**‘serum_creatinine’**	**‘ejection_fraction’**
PC 1	0.99714259	-0.0014545	0.07552836

PC 2	0.0755113	0.04786005	-0.99599571
PC 3	-0.00216612	0.99885299	0.04783313

Thus we can note that all the three dimensions have a relative significance with each other and it can we clearly seen by the principal component weights. We can note that sodium is essential for muscle control and lower sodium in body can weaken the muscle tissue of the heart, whereas, creatinine controls the kidney function which helps to flush out the toxins from the body and higher creatinine can cause the toxins to be released in the blood leading to heart failure. On the other hand, 'ejection_fraction' is a percent of blood pumped per contraction and lower percent of blood pumped mean more likely chances of heart failure.

Thus, I had chosen these three dimensions for this coursework. We could have also compared 'Age' instead of 'serum_sodium' as age also plays an important factor to failure of heart, but I felt age can only give a limited information for heart failure as most of the heart failure cases are between 60 – 80 years range.

D. Chicco and G. Jurman [3] paper also ranked the dimensions by using the Mann-Whitney U test which was a univariate application between each feature and the target variable. The 'serum_creatinine' as first with a p-vale of 0, 'ejection_fraction' was ranked second with a p-value of 0.000001, 'Age' was ranked as third with a p-value of 0.000167 and 'serum_sodium' was ranked as fourth with a p-value of 0.000293. Thus we can say the dimensions selected were correct.

Conclusion:

From the dimension reduction of the Heart Failure Prediction dataset it is difficult to draw conclusion because the failure of heart is a complicated classification problem and the subsets were limited to focus relevant medical features in the Kaggle dataset. If additional features were added to this analysis it would have been interesting to further work on it and predict the pattern of Heart Failure.

References :

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

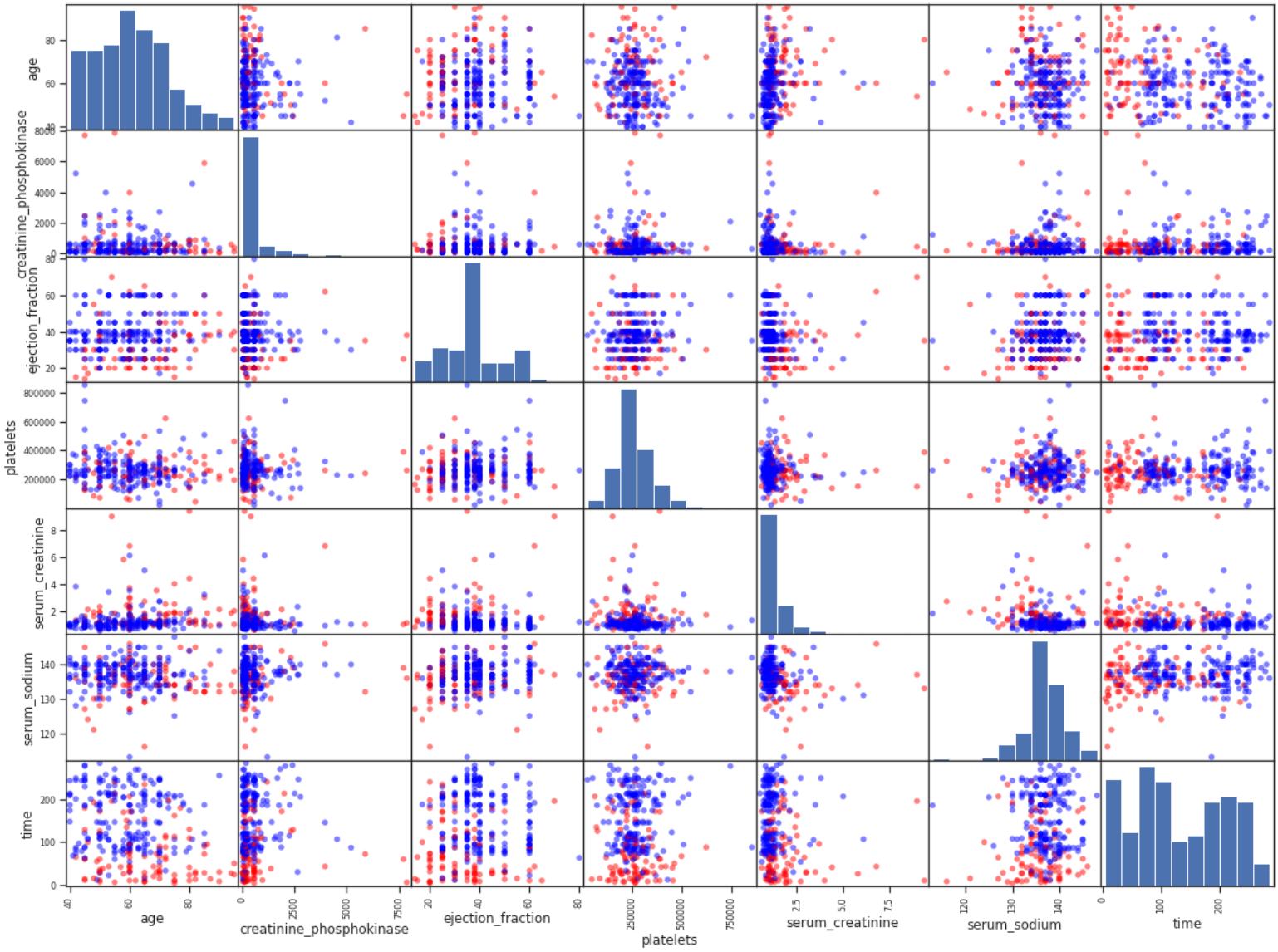
[2] “UCI Machine Learning Repository: Heart failure clinical records Data Set,” *Uci.edu*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. [Accessed: 1- Mar-2021].

[3] 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.

[4] “sklearn.decomposition.PCA — scikit-learn 0.24.1 documentation,” *Scikit-learn.org*. [Online]. Available: http://scikitlearn.org/stable/modules/generated/sklearn.decomposition.PCA.html . [Accessed: 12-Mar-2021].

Appendex 1

Graph 1 : Scatter Matrix Plot (Numeric features)



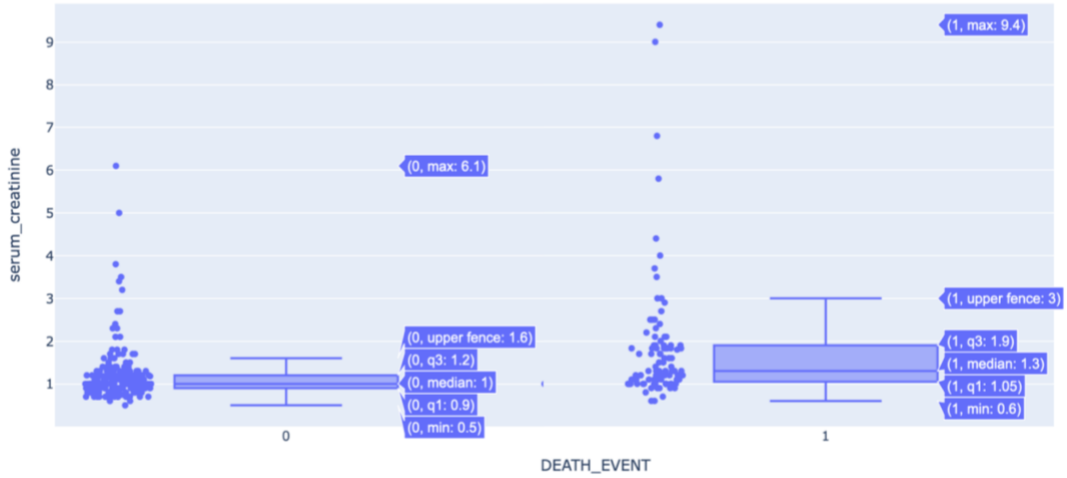
Appendix 2

serum_sodium in mEq/L

Graph 9 : Box plot for 'serum_sodium'

serum_creatinine in mg/dL

Graph 10 : Box plot for 'serum_creatinine'



ejection_fraction in Percentage %

Graph 11 : Box plot for 'ejection_fraction'