PREDICTING THE HEART FAILURE

FINAL REPORT

AUGUST 30, 2020

1. INTRODUCTION

Heart failure is the inability of the heart to supply adequate blood flow and therefore oxygen delivery to peripheral tissues and organs. Under perfusion of organs leads to reduced exercise capacity, fatigue, and shortness of breath. It can also lead to organ dysfunction (e.g., renal failure) in some patients.

It is estimated that there are more than 15 million new cases of heart failure each year worldwide. There are more than 600,000 new cases of heart failure diagnosed each year in the USA, and ten times that number of Americans currently in heart failure. The numbers are rapidly increasing because of the aging population. Heart failure is the leading cause of hospitalization of patients over 65 years in age.

Despite many new advances in drug therapy and cardiac assist devices, the prognosis for chronic heart failure remains very poor. One year mortality figures are 50-60% for patients diagnosed with severe failure, 15-30% in mild to moderate failure, and about 10% in mild or asymptomatic failure.

2. PROBLEM

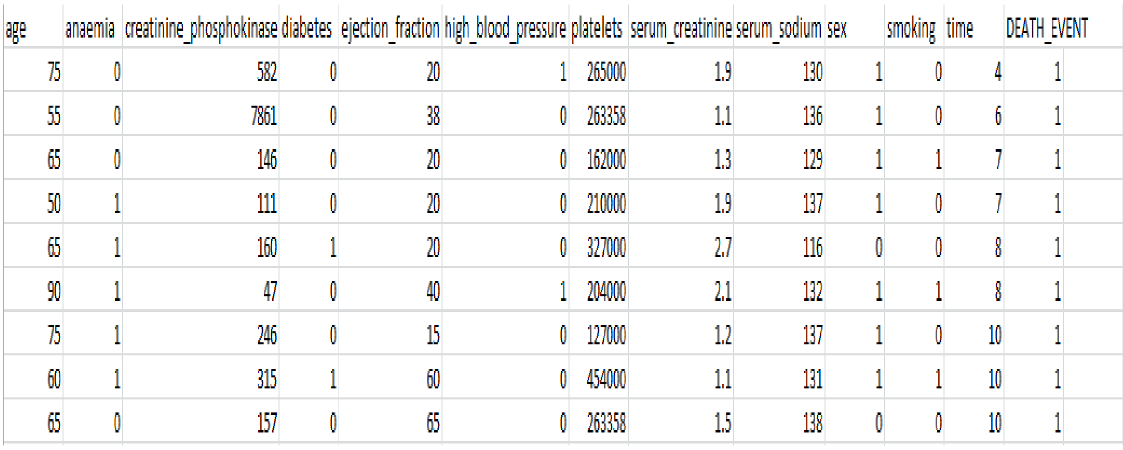
Data will contribute to find the cause of heart failure due to various factors. The project describes the significance of different feature problems that cause the death of the person due to heart failure. The aim of the project is to predict the most significant cause of the heart failure problem.

3. INTEREST

Obviously, doctors will be interested in this data analysis as this assists them by predicting the cause of the heart failure and take appropriate measures. People from other medical fields may also be interested.

4. DATA SECTION

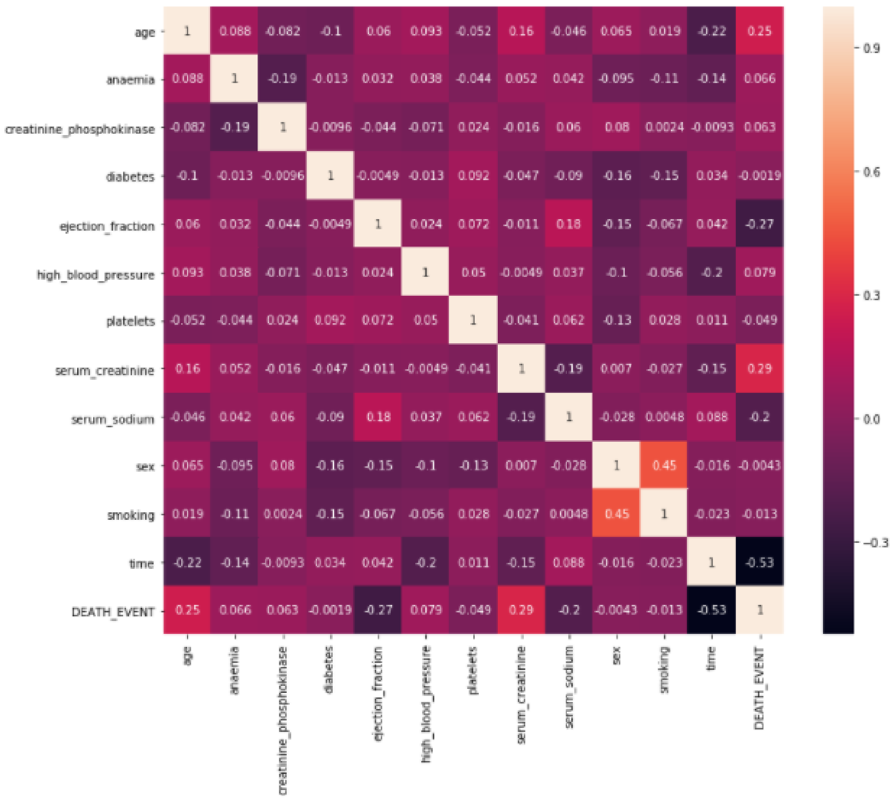
This section includes information about the data used for the problem.
The data was taken from **kaggle dataset** (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data/download). The data has 299 entries of different patients from age range of 40 to 95. Other features like anaemia, creatinine phosphokinase level, diabetes, ejection fraction, high blood pressure, platelet count, serum creatinine, serum sodium, sex, the person is smoking or not and follow up period (time) were used to predict the model based on the death event. Figure below shows the first 10 samples of the data used.



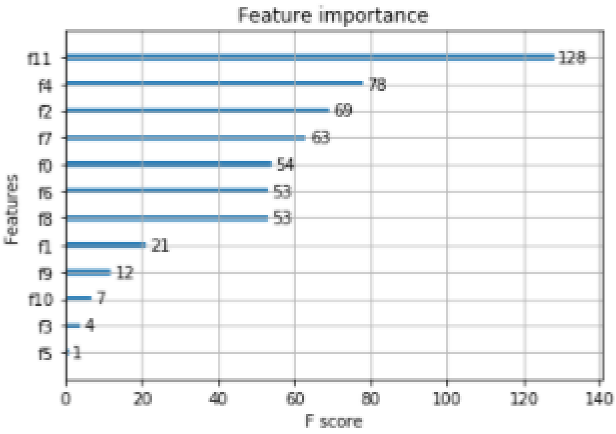
age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358	1.5	138	0	0	10	1

5. METHODOLOGY SECTION

- Libraries like pandas, numpy, matplotlib were imported.
- The dataset was imported as a csv file and the head was printed for a reference.
- All the warnings were set to be ignored.
- Features like anaemia, diabetes, blood pressure, smoking and death event were changed to type bool.
- Check for missing values was made, and there were no missing values.
- The correlation between all the features were plotted as a heat map as shown in figure below.



- Then histograms were plotted for all the non bool features to find the frequency of the corresponding data.
- Description of the death event was taken.
- Numbers of deaths due to Boolean based features like anaemia, smoking, high blood pressure, sex, diabetes were found individually and in combinations.
- The data was split into train and test data using train_test_split with test size 0.3 and random state 4.
- Pre-processing was done using min-max scalar.
- Logistic Regression was used for the classification problem.
- F1-score, precision and recall were calculated to find the significance of the features.
- XGBoost classifier was used to plot a comparative horizontal bar plot of importance of different features as shown in figure below.



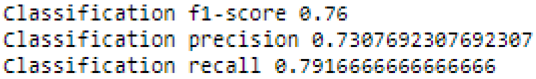
6. RESULT SECTION

- Death value based on Boolean features on a relative basis.

	Anaemia	Diabetes	High blood pressure	Sex	Smoking
Anaemia	#	18	19	26	12
Diabetes	*	#	17	20	12
High blood pressure	*	*	#	22	14
Sex	*	*	*	#	27
Smoking	*	*	*	*	#

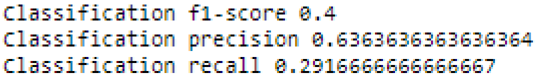
- Death value based on Boolean features on an individual basis.

  Anaemia – 46
  Diabetes – 40
  High blood pressure – 39
  Male – 62
  Smoking – 30
- Total number of deaths due to heart failure is 96.
- It was found that age is not a great determining factor.
- The scores were calculated without age factor.



```
Classification f1-score 0.76
Classification precision 0.7307692307692307
Classification recall 0.7916666666666666
```

- Time was the most significant factor, so when the time feature was removed from the feature list the scores dropped down to a much lower value.



```
Classification f1-score 0.4
Classification precision 0.6363636363636364
Classification recall 0.2916666666666667
```

- Similarly other scores were tried for additional information by adding and removing features to find the feature significance.

7. DISCUSSION SECTION

- The most significant feature was time. It was verified using boxplot and regplot.
- Two third of the death population were male.
- Platelet count did not give significant information about the death event.
- Nearly half of the death population were anaemic.

8. CONCLUSION

This model was used for a comparative study for the significant feature which was found to be time. The study can be further extended by comparing the accuracy scores of different algorithms or by predicting the threshold level of different enzymes in the blood. If new factors other than those given in the data are more significant than those in the given dataset, then the model will be modified based on the new features.