Student ID: S3802846, S3858391

Student Name: Rahul Kumar, Denis Bharatbhai Vaghasia

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 honour code by typing "Yes": *Yes*.

Report on Heart Failure Prediction

Rahul Kumar
Denis Bharatbhai Vaghasia

Royal Melbourne Institute of Technology, Melbourne

s3802846@student.rmit.edu.au
s3858391@student.rmit.edu.au

May 23, 2021

Table of Contents

Abstract/Executive Summary

This report aims to investigate mortality rate due to heart failure. The dataset contains 12 different features like age of the patient, existing medical conditions and daily habits like smoking. Overall, the result indicates that people with pre-medical conditions like high blood pressure, diabetes and unhealthy lifestyle are more vulnerable to heart failure. Report concludes that age, gender, daily habits and underlying health problems contributes largely causing death of people suffering from cardiovascular disease. Looking at the pattern, it is recommended that heart diseases can be prevented by avoiding unhealthy behaviours like smoking, use of tobacco and liquor, obesity and physical inactivity. Additionally, early detection and proper treatment can help in recovery and avoiding deaths.

Introduction

Heart failure is a common disease in today's scenario and there has been a great increase in number of deaths by heart failure over the time. The heart fails when heart can't pump enough blood and oxygen to muscles and organs. It is important to know what are most possible reasons that can increase risk of heart failure and what can be done to lower the risk. Some of these factors cannot be controlled like gender or age, the risk of heart disease increases with age. On the other, there are other factors can be controlled to reduce to increase chances of survival. This includes controlling lifestyle changes - such as exercising, controlling blood pressure by reducing sodium in diet, managing diabetes, avoiding smoking and limiting alcohol consumption and managing stress.
This report will discuss various risk factors contributing to deaths due to cardiovascular diseases and how this analysis can be used to make predictions and informed decisions.

Methodology

The data is sourced from https://archive.ics.uci.edu/ml/datasets/ which contains medical reports of 299 heart failure patients collected during their follow-up period at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015.

The dataset contains 13 features which includes age, gender, lifestyle information and other medical conditions as described below.

S.No.	Feature	Description	Value
1	Age	Age of the patient	Years, 40-95
2	Creatinine Phosphokinase	Level of the CPK enzyme in the blood	(mcg/L), 23-7861
3	Ejection Fraction	Percentage of blood leaving the heart at each contraction	%, 14-80

4	Platelets	Platelets in the blood	kiloplatelets/mL, 25000-850000
5	serum_creatinine	Level of serum creatinine in the blood	(mg/dL), 0.5-9.4
6	serum_sodium	Level of serum sodium in the blood	(mEq/L), 113-148
7	Sex	Gender of the patient	Boolean, 0 = Female, 1 = Male
8	Diabetes	If patient is diabetic	Boolean 0 = No, 1 = Yes
9	Anaemia	Decrease of red blood cells or hemoglobin	Boolean, 0 = No, 1 = Yes
10	High blood pressure	Hypertension condition	Boolean, 0 = No, 1 = Yes
11	Smoking	Patient smokes or not	Boolean, 0 = No, 1 = Yes
12	Time	Follow up period	Days, 4-285
13	DEATH_EVENT (target)	If patient died during follow up period	Boolean, 0 = No, 1 = Yes

Data is analysed by first importing dataset using Pandas and then applying checks for null values and white spaces to make sure that data is clean. Continuous numerical variables are explored using histogram chart while categorical variables are explored using count plot. After that, we explored the relationship between various variables and did feature engineering to analyse important variables/features and their impact on the target. Supervised learning techniques such as K- Nearest Neighbour Classifier and Decision Tree Classifier have been applied for data modelling while Hyper Parametric method is used for improving the performance of the model. At last, we compared the accuracy of classification methods.

Results

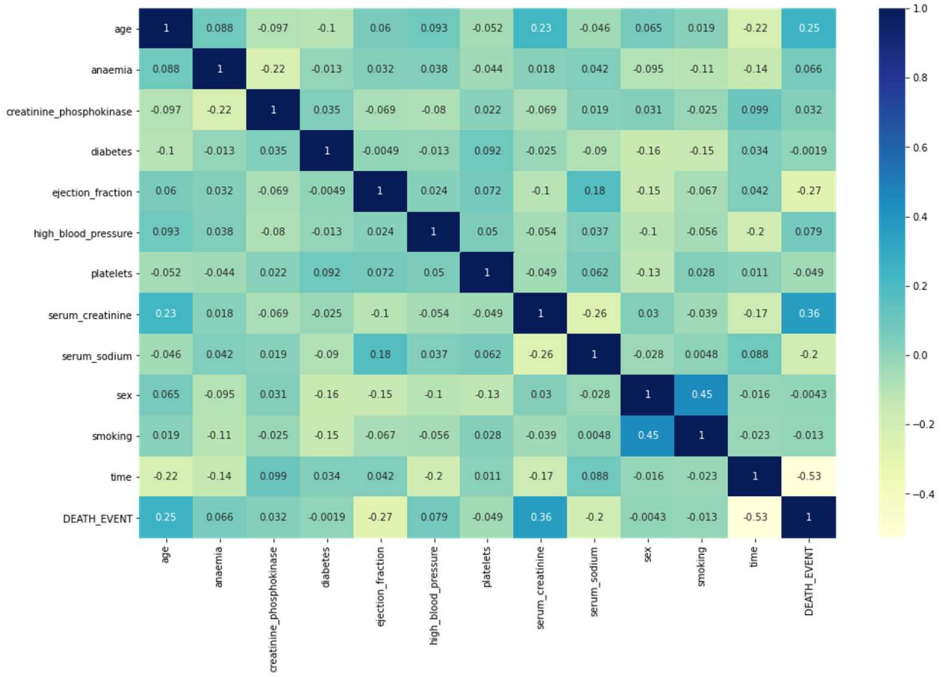
Data Cleaning

Data does not contain any missing values, special values and obvious errors (i.e., inconsistencies) which indicates that data is clean.

Data Exploration

Finding correlation between all columns

All columns are explored using heatmap to find correlation.

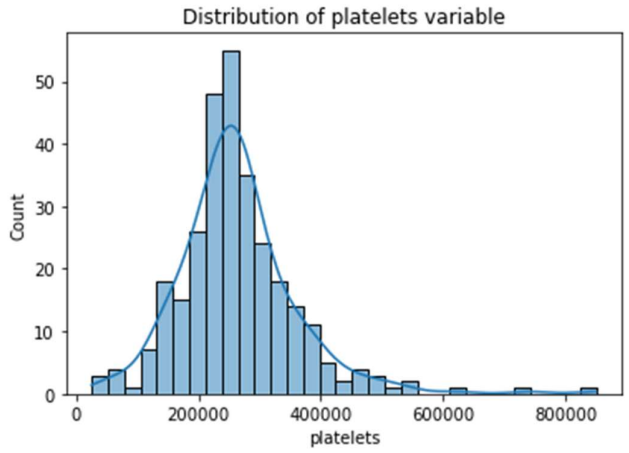


It can be seen that none of the features are highly correlated (not greater than 0.5), so we cannot remove any column.

Exploring continuous numerical variables

There are 7 numerical variables – Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, serum_creatinine, serum_sodium and time and these are analysed using histogram plot.

For e.g., below graph gives the distribution platelet count in all patients.

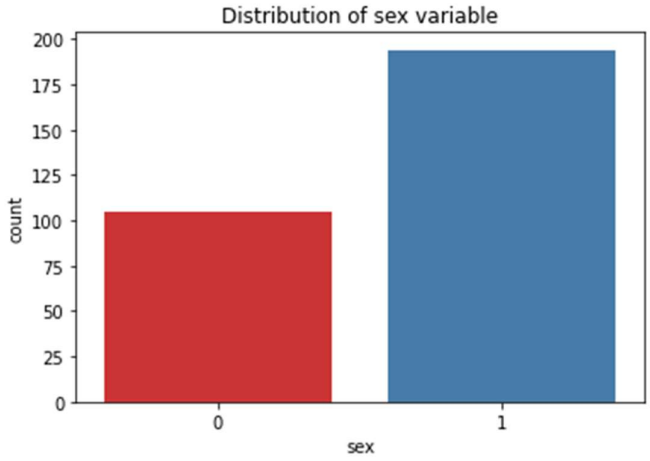


It can be seen from the graph that most of the people are having platelet count of around 30000 while only few patients have platelet count above 400000 which means people in this range are more prone to heart failure.

Exploring categorical variables

There are 6 categorical variables - Sex, Diabetes, Anaemia, High blood pressure, Smoking and DEATH_EVENT and these are analysed using count plot.

For e.g., below graph gives the distribution of gender/sex all patients.



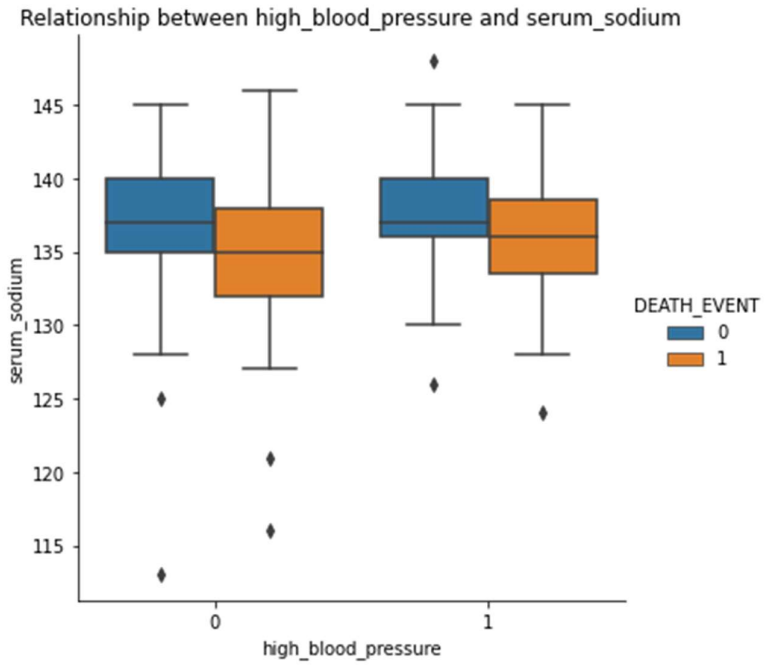
It can be seen from the graph that % of male patients (=1) is 60% while that of female patients (=0) is 40%. It can be inferred that men are more vulnerable to death due to heart failure as compared to women.

Relationship between variables

Explored different pair of attributes using different types of graphs to find out if there is any interesting relationship between different attributes.

Relationship between Death_Event and Age

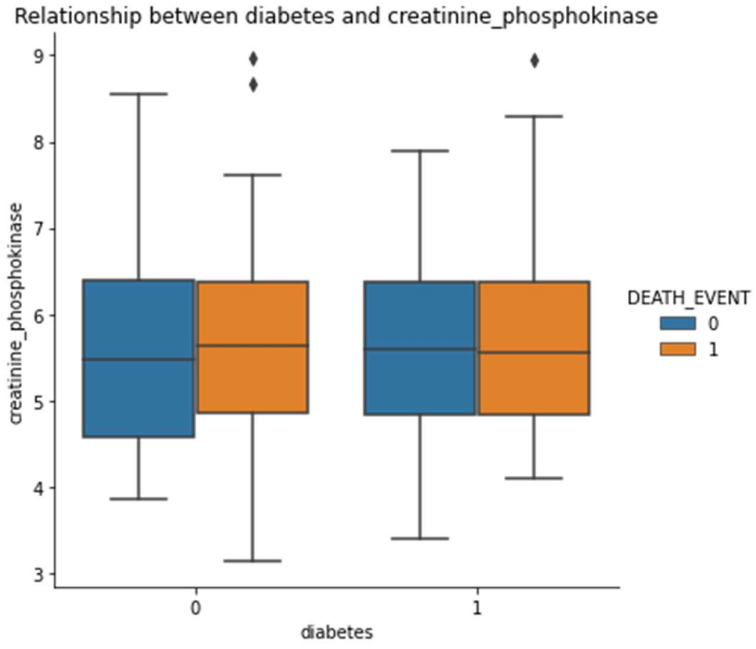
The below box plot shows relationship between high blood pressure and sodium level in all patients.



It can be seen from the graph that there is a strong relationship between high blood pressure and sodium level. Normal range of values of serum sodium is 135 - 145(mEq/L). Since the values of sodium level provided in dataset is within the range, in the plot above we see that death event is less in case of people not having high blood pressure. This concludes, low intake of sodium results into controlled blood pressure and ultimately low risk of heart failure.

Relationship between Diabetes and Creatinine_phophokinase

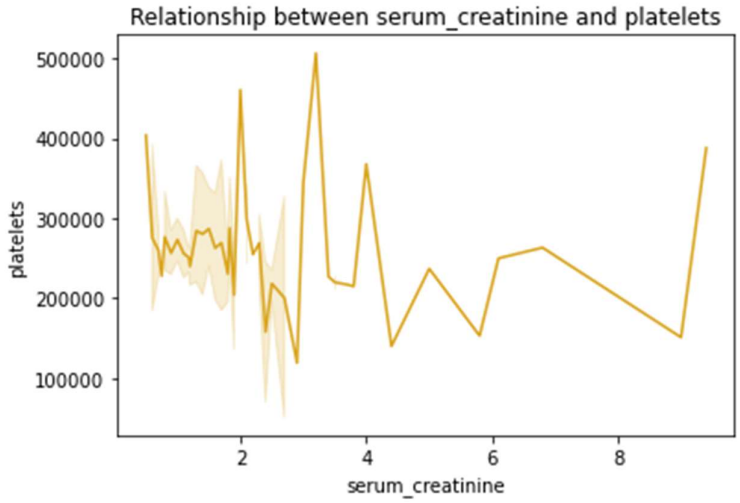
The below box plot shows relationship between diabetes and CPK level all patients.



It can be seen from the graph that there is no proper relationship between diabetes and CPK level in blood since the lines are not overlapping, thus it does not have much effect on target DEATH_EVENT.

Relationship between serum_creatinine and platelets

The below line graph shows relationship between creatinine level and platelets count in all patients.



It can be seen from the graph that as creatinine level increases, count of platelet decreases. Thus, it can be inferred that creatinine level should be low to maintain high platelet count.

Data Modelling

Used below Classification model for data modelling because classifier method works on both continuous and categorical data and “heart_failure_clinical_records_dataset” data set contains continuous as well as categorical numerical data.

K-Nearest Neighbor classifier

K-Nearest Neighbor classifier is a simple supervised classification algorithm. KNN does not make any assumptions on the data distribution, hence it is non-parametric.

- We have selected 12(['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets','serum_creatinine', 'serum_sodium', 'sex', 'smoking', 'time']) as data features into the **X** data-frame and target field ([‘DEATH_EVENT’] into a **y** data-frame.
- After splitting the data, we take 0.7% data for training and remaining for testing purposes.
- We import the classifier model from the sklearn library and fit the model and achieved an accuracy of 0.67%.
- We used confusion matrix to measure the performance of the model. Which is show below:

[[54 12]

[19 5]]

- From confusion matrix, we can state that the 54 times model predicted True positive and 5 times it predicted True Negative. We can also state that 12 times model predicted False Negative and 19 times it predicted False Positive.
- We used classification report to measure the quality of predictions from our Knn model.

	precision	recall	f1-score	support
0	0.74	0.82	0.78	66
1	0.29	0.21	0.24	24
accuracy			0.66	90
macro avg	0.52	0.51	0.51	90
weighted avg	0.62	0.66	0.63	90

- We optimized our model by changing the k value to get maximum accuracy.
- After optimizing the model, we achieved an accuracy of 0.70%.
- Again, used confusion matrix to measure the performance of the model which is shown below.

```
[[62  4]
 [23  1]]
```

- From confusion matrix, we can state that the 62 times model predicted True positive and 1 time it predicted True Negative. We can also state that 4 times model predicted False Negative and 23 times it predicted False Positive.
- We used classification report to measure the quality of predictions from our Knn model.

	precision	recall	f1-score	support
0	0.73	0.94	0.82	66
1	0.20	0.04	0.07	24
accuracy			0.70	90
macro avg	0.46	0.49	0.45	90
weighted avg	0.59	0.70	0.62	90

Decision Tree

A type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. It splits the population or sample into two or more homogeneous sets (or sub-populations), based on most significant splitter / differentiator in input variables.

- We have selected 12(['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets','serum_creatinine', 'serum_sodium', 'sex', 'smoking', 'time']) as data features into the **X** data-frame and target field (['DEATH_EVENT'] into a **y** data-frame.
- After splitting the data, we take 0.7% data for training and remaining for testing purposes.
- We import the decision tree classifier model from the sklearn library and fit the model and achieved an accuracy of 0.79%.
- We used confusion matrix to measure the performance of the model which is shown below:

```
[[55 11]
 [ 7 17]]
```

- From confusion matrix, we can state that the 55 times model predicted True positive and 17 times it predicted True Negative. We can also state that 11 times model predicted False Negative and 7 times it predicted False Positive.
- We used classification report to measure the quality of predictions from our decision tree model.

	precision	recall	f1-score	support
0	0.89	0.83	0.86	66
1	0.61	0.71	0.65	24
accuracy			0.80	90
macro avg	0.75	0.77	0.76	90
weighted avg	0.81	0.80	0.80	90

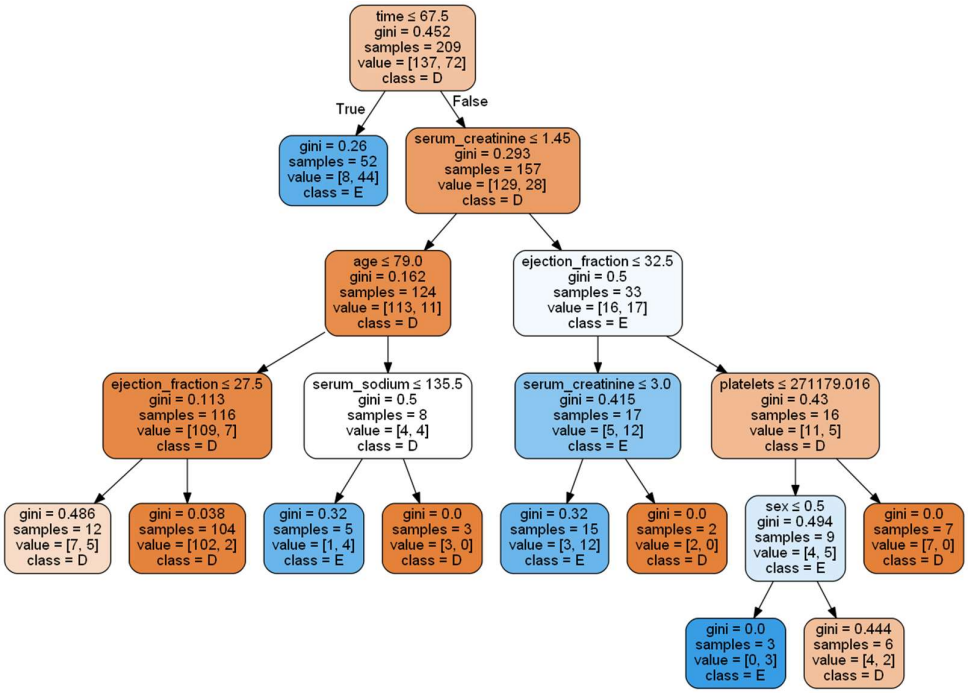
- We optimized your model by changing the max_leaf_nodes value to get maximum accuracy.
- After optimizing the model, we achieved an accuracy of 0.89%.
- Again, used confusion matrix to measure the performance of the model. Which is show below.

```
[[62  4]
 [6  18]]
```

- From confusion matrix, we can state that the 62 times model predicted True positive and 18 time it predicted True Negative. We can also state that 4 times model predicted False Negative and 6 times it predicted False Positive.
- We used classification report to measure the quality of predictions from our decision tree model.

	precision	recall	f1-score	support
0	0.91	0.94	0.93	66
1	0.82	0.75	0.78	24
accuracy			0.89	90
macro avg	0.86	0.84	0.85	90
weighted avg	0.89	0.89	0.89	90

- Tree Graph:



- Above we can see the tree built after training. During the training phase, the Decision Tree add nodes, split them into branches that lead to leaves.

By comparing the accuracy among two algorithms, K-nearest neighbour algorithm gives accuracy of 70% while Decision Tree classifier gives accuracy of 89%. Thus, Decision Tree classifier is more accurate.

Discussion

This study shows that there are multiple factors leading to heart failure which includes age, gender, pre-existing medical conditions and behavioural factors. As the data demonstrates, number of deaths due to cardiovascular disease is almost double in male patients as compared to female patients. Further, report indicates a smaller number of deaths in case of people having normal range of sodium in blood which means no high blood pressure. It is evident that people should be educated to maintain healthy lifestyle which can prevent health failures.

Conclusion

Heart Disease is one of the major concerns for society today as causes a significant number of deaths. There are multiple factors which increases the risk of heath failure some of which are controllable while some are not. Definitely, nothing can be done for unrollable factor but fortunately there are many things which can be done to reduces chances of getting heart disease. This includes, eating a healthy diet, physical activities, maintaining blood pressure by controlling sodium intake, maintaining diabetes level by controlling sugar intake and avoiding smoking. Different awareness programs through various channels can help in educating people on how to protect their heart.

References

- Boschettiand L. Massaron, Python Data Science Essentials, Chapters 4 and 6
- Pandas read_csv: http://pandas.pydata.org/pandas-docs/
- https://www.webmd.com/heart-disease/heart-failure/understanding-heart-failure-prevention
- https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5
- https://seaborn.pydata.org/generated/seaborn.heatmap.html
- https://seaborn.pydata.org/generated/seaborn.histplot.html
- https://seaborn.pydata.org/generated/seaborn.countplot.html
- https://seaborn.pydata.org/tutorial/categorical.html
- https://seaborn.pydata.org/generated/seaborn.lineplot.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html