Final Project Report
Introduction to Data Analytics

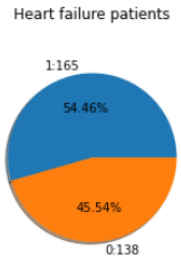
Project Title:
Prediction /Analysis of the factors causing heart failure

Prepared by:
Maryann Meilika

ITE 5201 – Summer 2022
Humber College

1. Problem Statement

Prediction /Analysis of the most factors causing heart failure

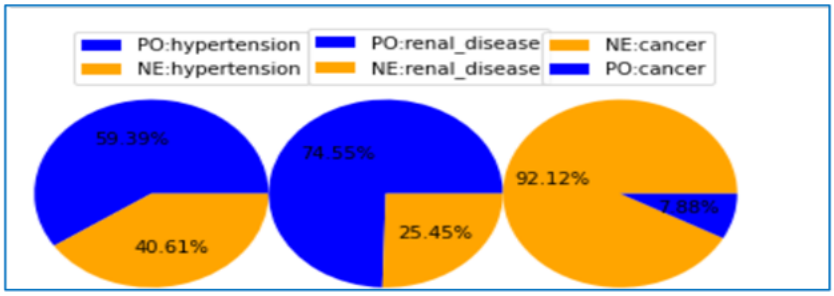


2. Dataset Description

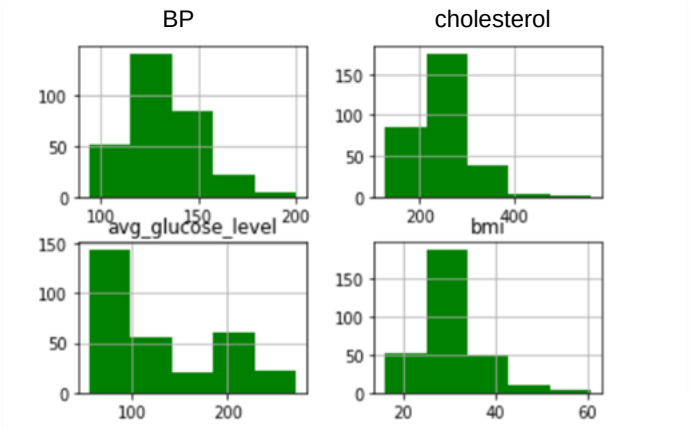
There are many factors causing the heart failure condition. Some of them are diseases like high blood pleasure, cholesterol, Thalassemia, hypertension, and the renal disease. On other hand some other external factors like age and gender. All this information from the dataset will help us to analyze and predict th e most factors that when exist increase the chance of having a heart failure.
My dataset has 14 columns and 303 rows.
I checked if the dataset has almost equal percentages of heart failure and non- heart failure patients to s ee if it is not a biased dataset, it turned out that
there is 55 % of heart failure patients and the rest don't have it. So, it is not biased as their percentages are almost equal and can be used.

3. Dataset Analysis and Observations

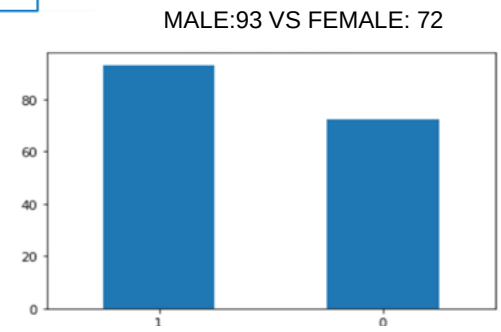
Depending on my dataset ,I used histograms , bar chart and a pie chart for single variate Data Analysis and the Heatmap for correlation.



According to the pie charts , many of those who had heart failure had the renal disease. Almost (75%) and less percentage had hypertension (60%) .But cancer isn't one of the factors causing heart failure as it a very low percentage got it .

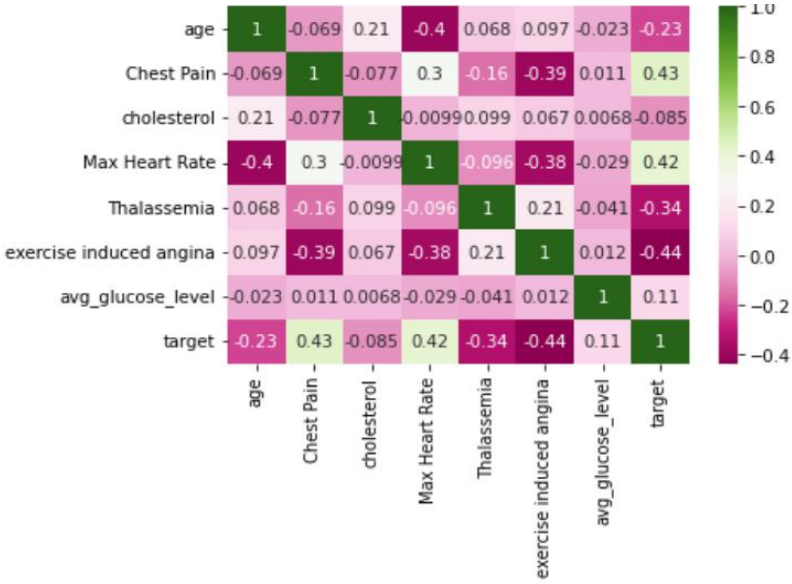


Those are Histograms for BP, cholesterol, BMI, avg_glucose_level, showing the average of each of them.



The difference between the percentages of males and females who had a heart failure, shows that males have a higher chance to get a heart failure.

In the Heatmap , Chest pain and Max Heart rate have a positive relation with the target showing that it is one of the factors that would increase the chance of having a heart failure. While exercising induced angina would decrease the chances of having a heart failure as it has a negative relation with the target.

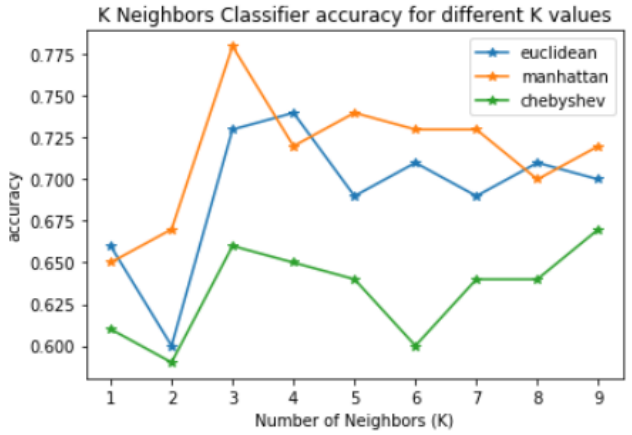


4. Proposed Analytical/Prediction Model

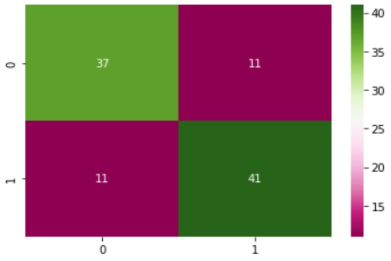
First, I chose to classify the data to check if the variables we got influence determining or predicting if someone might get a heart failure. I used two models which are KNN and the Random Forest classifier. So, this variety ensures that the data I am using is related to the target. I used the KNN model in my classification because it is an accurate model for smaller datasets, and I used the random forest model as it has high prediction accuracy and automated bootstrapping. My input is the dataset without the target while my output is the target column.

5. Results and Discussions

Modeling with k_NN

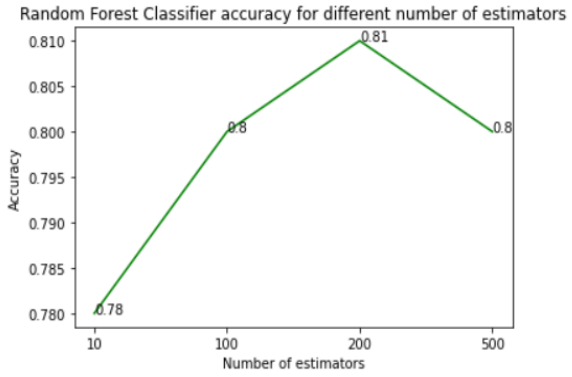


	precision	recall	f1-score	support
0	0.77	0.77	0.77	48
1	0.79	0.79	0.79	52
accuracy			0.78	100
macro avg	0.78	0.78	0.78	100
weighted avg	0.78	0.78	0.78	100

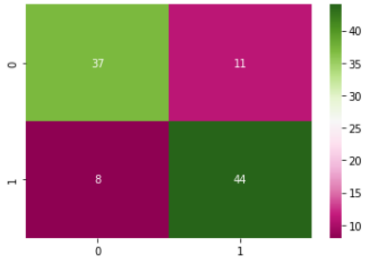


According to the classification report, plots above & the heatmap, the accuracy for the k_NN classifier with the arguments(matrix = 'manhattan' and k=3) has the highest classifying accuracy = 0.78 among the other metrics

Random Forest Classifier



	precision	recall	f1-score	support
0	0.82	0.77	0.80	48
1	0.80	0.85	0.82	52
accuracy			0.81	100
macro avg	0.81	0.81	0.81	100
weighted avg	0.81	0.81	0.81	100



According to the classification report, plots above & the heatmap, the highest accuracy was 0.81 when the total estimators were 200.

Conclusion:

The factors I chose were related to heart failure and the Random Forest classifier (the one better than K_NN classifier) got the highest accuracy(0.81), proving that they are highly causing factors. However, the heatmap and the data visualizations I did above showed that Chest pain and increased heart rate increases the chance of getting a heart failure.