Heart Disease Classification

Machine Learning 2021-2022

Table of Contents

Table of Figures

Table of Tables

1 Introduction

Heart disease is the number one cause of death among men and women, and most racial and ethnic groups in the United States (Centers for Disease Control and Prevention, 2021). In Europe, cardiovascular disease (CVD) accounts for 45 percent of all deaths and 37 percent of deaths in the European Union (Wilkins, et al., 2017). There are around 7.6 million people living in the United Kingdom with a heart or circulatory disease (British Heart Foundation, 2021). In 2016 to 2017, heart disease cost the United States 363 billion United States Dollars (USD) (Centers for Disease Control and Prevention, 2021). These statistics show that cardiovascular diseases not only cause a substantial portion of deaths all over the world, but that it also costs a significant amount of money for diagnosis, treatment, and productivity lost due to such diseases. Angiography, a method of visualising blood flow to major blood vessels, is regularly used to diagnose cardiovascular and circulatory diseases, but it is associated with high costs and side effects (Arabasadi, et al., 2017). Due to this fact, it is critical to implement digital and technological solutions for the costly problem of not only diagnosing if cardiovascular disease is present in any given patient, but also to what extent it is present. In this report, I will implement various machine learning algorithms including classifiers and clustering algorithms on existing patient records so the models can be used on unseen data to classify, detect, and quantify the degree of present heart disease.

2 Exploratory Data Analysis

Data analysis is the cornerstone of any Data Science project; the Data Scientist must examine their dataset prior to analysis so that they know what steps to take when cleaning and pre-processing the dataset.

2.1 Dataset Information

```
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         920 non-null    float64
 1   Gender                      920 non-null    float64
 2   Chest pain type             920 non-null    float64
 3   Resting blood pressure      861 non-null    float64
 4   Serum cholesterol           890 non-null    float64
 5   Fasting blood sugar         830 non-null    float64
 6   Resting electrocardiographic 918 non-null   float64
 7   Maximum heart rate          865 non-null    float64
 8   Exercise induced angina     865 non-null    float64
 9   ST depression               858 non-null    float64
 10  ST segment                  611 non-null    float64
 11  Number of major vessels     309 non-null    float64
 12  Thal                        434 non-null    float64
 13  class                       920 non-null    int64
```

Figure 1: Dataset description

The dataset used in this exploratory analysis contains 920 labelled instances defined by 13 features and their class labels. Each record feature contains numerical continuous data in the form of floating-point numbers. Each record's class contains a numerical value corresponding to the developed degree of heart disease in the form of incrementing integers.

	Age	Gender	Chest pain type	Resting blood pressure	Serum cholesterol	Fasting blood sugar	Resting electrocardiographic	Maximum heart rate	Exercise induced angina	ST depression	ST segment	Number of major vessels	Thal	class
0	63.0	1.0	4.0	140.0	260.0	0.0	1.0	112.0	1.0	3.0	2.0	NaN	NaN	2
1	44.0	1.0	4.0	130.0	209.0	0.0	1.0	127.0	0.0	0.0	NaN	NaN	NaN	0
2	60.0	1.0	4.0	132.0	218.0	0.0	1.0	140.0	1.0	1.5	3.0	NaN	NaN	2
3	55.0	1.0	4.0	142.0	228.0	0.0	1.0	149.0	1.0	2.5	1.0	NaN	NaN	1
4	66.0	1.0	3.0	110.0	213.0	1.0	2.0	99.0	1.0	1.3	2.0	NaN	NaN	0

Figure 2: First five instances of the dataset used in this study

The features appear to represent multiple areas of documentation including individual factors (Age, Gender), symptoms (Chest pain type, Exercise induced angina), blood and circulation conditions (Resting blood pressure, Serum cholesterol, Fasting blood sugar, Maximum heart rate, Thalassemia), and medical tests (Resting ECG, ST depression, ST segment, Number of major vessels).

2.2 Class Balance

```
0    411
1    265
2    109
3    107
4     28
Name: class, dtype: int64
```

Figure 3: Numerical representation of class distribution

All 920 instances in this dataset are labelled with classes, ranging from 0 – representing absence of heart disease – to 4 – representing severe presence of heart disease. Classes 0, 1, 2, 3, and 4 contribute ~44.7%, ~28.8%, ~11.8%, ~11.6%, and ~3.1% of the distribution respectively.

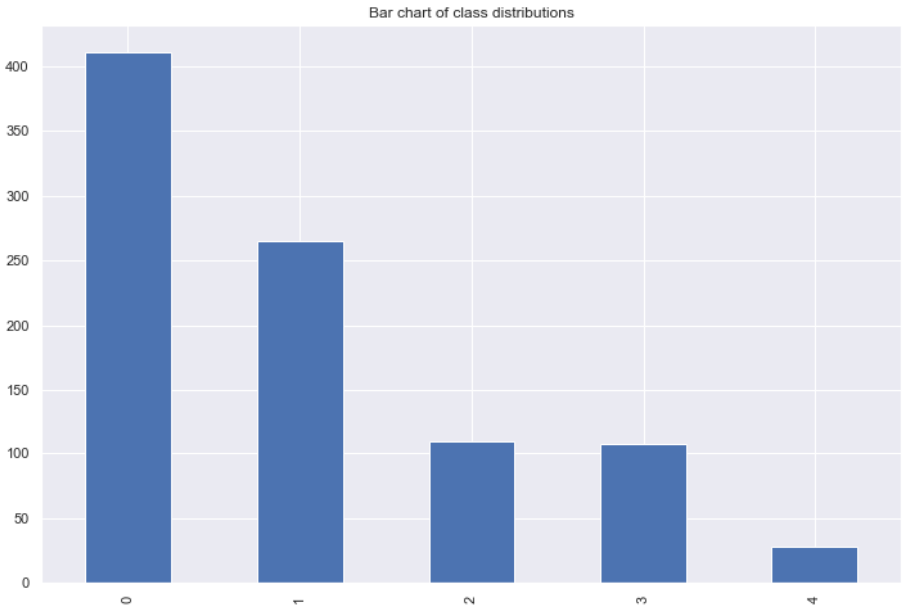

Figure 4: Visual representation of class distribution using bar chart

The class distribution of this dataset loosely resembles a negative exponential distribution as the classes decrease in frequency substantially, followed by decreasing by smaller increments each time. Almost 50% of the dataset is labelled as having an absence of a heart disease diagnosis. Due to the discrepancy in class representation, I may have to oversample the instances with higher class values, or undersample the instances with lower class values.

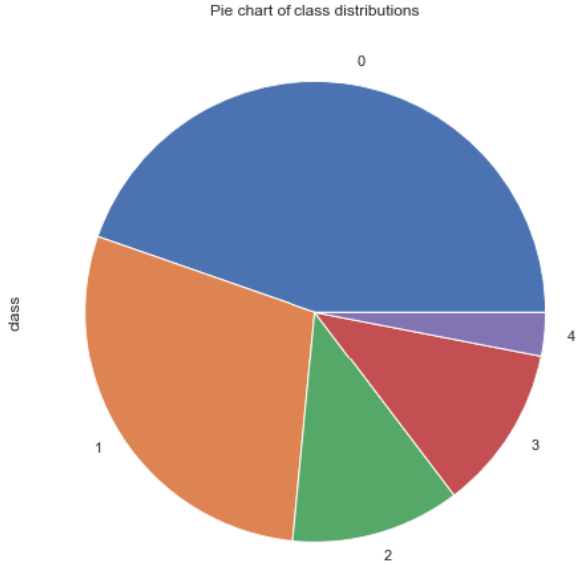

Figure 5: Visual representation of class distribution using pie chart

2.3 Feature Analysis

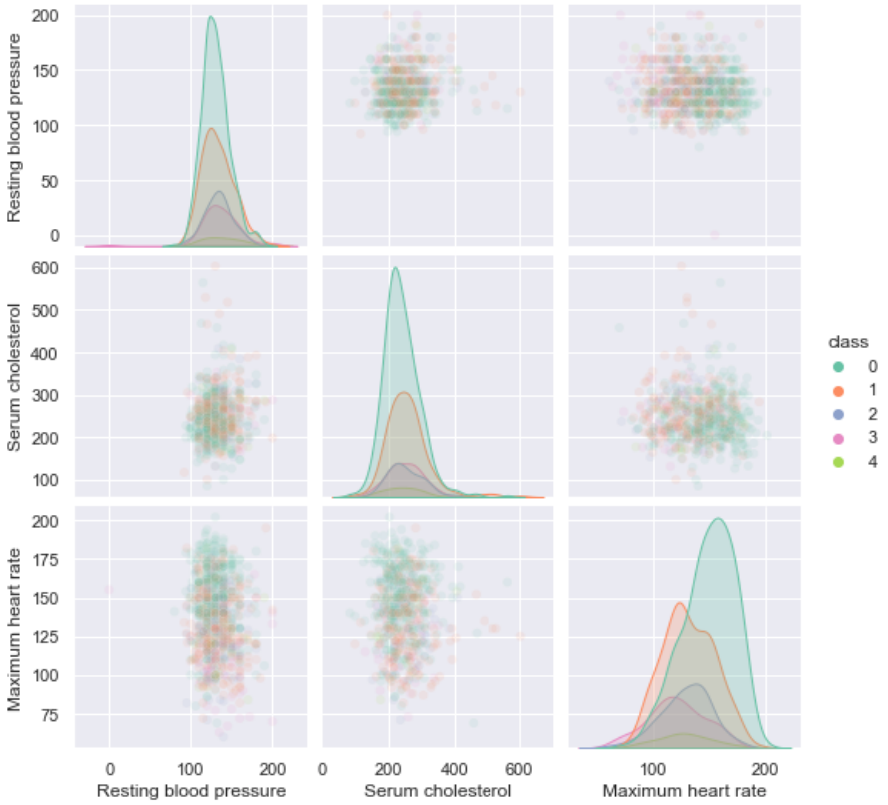

Figure 6: Pairplot comparing Resting blood pressure, Serum cholesterol, and Maximum heart rate, representing classes

I chose to focus on subplots comparing blood and circulation metrics due to their likeliness to give insight on how they affect the application of class labels. In the subplots comparing maximum heart rate to both resting blood pressure and serum cholesterol, lower class values are present. One insight I made from this is that instances of lower classes can achieve higher maximum heart rates. Likewise, higher valued classes appear to occupy the lower ends of these spectrums. However, serum cholesterol doesn't appear to have much of an effect on the determination of class labels due to the appearance of the subplot class points being grouped together with no degree of congregation.

The diagonal histograms give insights on the class distribution, positioning, and skew of each feature. For example, the peaks of each class' curve for resting blood pressure appear further up the scale with each degree of heart disease presence, signifying those with higher resting heart rates have a higher probability of having a higher advancement of heart disease. This is particularly prominent in the curve for the worst degree of heart disease, with the peak rising higher than the sloping curve of the other degrees of presence.

2.4 Missing Values

```
Age                              0
Gender                           0
Chest pain type                  0
Resting blood pressure          59
Serum cholesterol               30
Fasting blood sugar             90
Resting electrocardiographic     2
Maximum heart rate              55
Exercise induced angina         55
ST depression                   62
ST segment                     309
Number of major vessels        611
Thal                           486
class                            0
dtype: int64
```

Figure 7: Quantity of feature values which are null

Most features in this dataset contain missing values, and these are spread across most instances. The only three features which do not contain null values are Age, Gender, and Chest pain type. The worst offender for missing values is the Number of major vessels. This is to be expected as only the minority of patients would have had fluoroscopy undertaken when their feature values were recorded for the dataset. Thalassemia is also a feature which contains over half of its values as null. This is likely since a test was not undertaken to classify whether a patient had Thalassemia during every single consultation, and as such, these metrics were unable to be recorded as a definitive value was not declared.

These missing values can either result in the instances they're contained in to be discarded from the dataset or those values must be imputed considering the existing values for that feature. Missing values can mislead classification and clustering algorithms into taking the presence or absence of the feature as a feature in and of itself and consider it a feature which dictates the class of the instance that contains it.

2.5 Duplicate Rows

	Age	Gender	Chest pain type	Resting blood pressure	Serum cholesterol	Fasting blood sugar	Resting electrocardiographic	Maximum heart rate	Exercise induced angina	ST depression	ST segment	Number of major vessels	Thal	class
187	58.0	1.0	3.0	150.0	219.0	0.0	1.0	118.0	1.0	0.0	NaN	NaN	NaN	2
605	49.0	0.0	2.0	110.0	NaN	0.0	0.0	160.0	0.0	0.0	NaN	NaN	NaN	0

Figure 8: Two instances of rows containing duplicate values

There are only two occurrences of instances being duplicated in this dataset. These rows will have to be discarded as the probability that a patient had identical feature values to another patient is near zero.

2.6 Noise

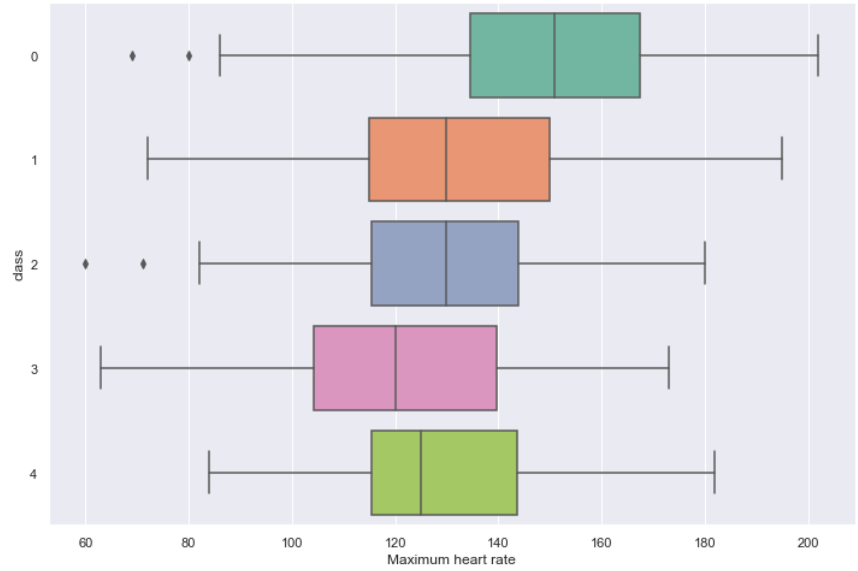

Figure 9: Box plot comparing Maximum heart rate to class

When displaying maximum heart rates against their classes, it's clear that those who are classified as having an absence of heart disease can achieve a higher heart rate. This trend continues as the increments of presence increase. However, a maximum heart rate of 60 beats per minute (BPM) is unlikely to occur during exercise. Therefore, the instances with this feature value could be flagged as potential noise. Also, due to those with an absence of heart disease being able to achieve higher heart rates, instances with recorded heart rates outside of the lower whisker could be flagged as potential noise as they don't fit within the trending bounds of this feature.

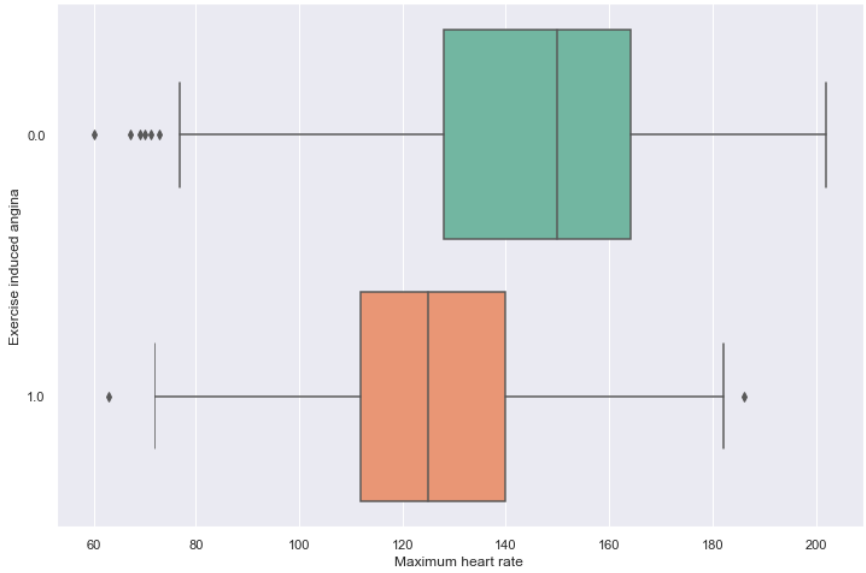

Figure 10: Box plot comparing Maximum heart rate to Exercise induced angina

When comparing maximum heart rate against exercise induced angina, on average, those that didn't experience exercise induced angina achieve higher heart rates. This is likely because those patients that experience heart pain stop exercising, preventing their heart rates from increasing further. Due to this fact, there are several instances that could be flagged as potential noise because they have recorded heart rates outside of the lower whisker. Similarly, there is one point of potential noise in those who experienced exercise induced angina as they achieved a heart rate higher than the upper whisker.

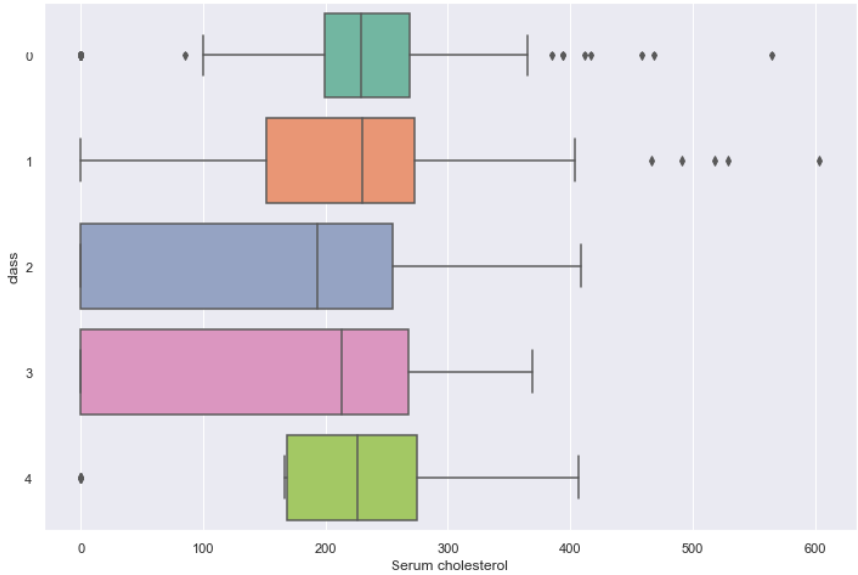

Figure 12: Box plot comparing Serum cholesterol to class

When comparing Serum cholesterol levels against their classes, there is a considerable amount of noise outside of the upper whiskers for those classified as having an absence of heart disease, and those classified as having a low presence of heart disease. However, not all cholesterol is detrimental to human health. One conclusion that could be drawn from such noise due to the presence of low-density lipoprotein (LDL), which is good cholesterol (Ma & Shieh, 2006). Therefore, cholesterol level is likely a bad feature to use for classifying the severity of heart disease due to the polarity in types of cholesterol.

For classes 2 and 3, the boxes extend down to zero. This is likely due to the quartiles being forced to drop to zero because of the number of zero values in this feature. This creates a substantial amount of noise as patients from these classes likely weren't subject to serum cholesterol tests when their feature data was recorded. These zero values can either result in the containing instances being omitted or those values imputed.

2.7 Correlation Analysis

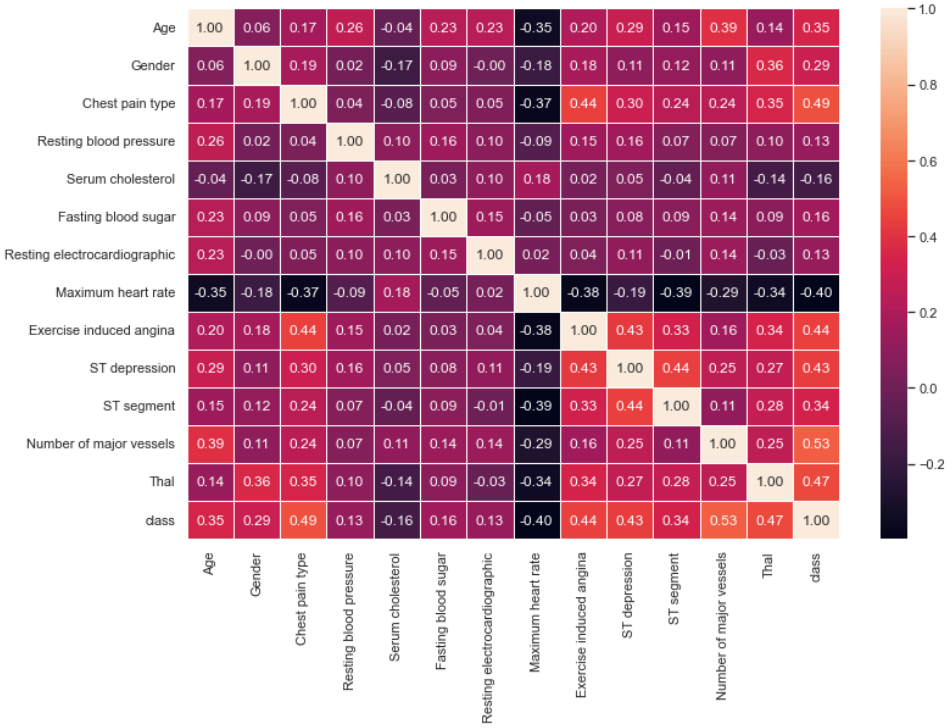

Figure 13: Correlation matrix visualising correlation values whilst comparing each feature

The correlation matrix generated using Seaborn did not display any strong correlation. However, the strongest correlations are between number of major vessels and class, chest pain type and class, and Thalassemia and class. These feature pairs barely have 50% correlation, meaning there are no metrics which define a clear correlation between them. There are no strong negative correlations either, the strongest being between maximum heart rate and class. Whilst not strong, these correlations make sense as the more advanced the heart disease classification, the lower the recorded maximum heart rate is likely to be. Likewise, the more advanced the heart disease classification, the more severe and the more frequent exercise induced angina is likely to be.

Works Cited

Arabasadi, Z. et al., 2017. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Computer Methods and Programs in Biomedicine,* Volume 141, pp. 16-26.

British Heart Foundation, 2021. *Facts and figures - Information for journalists | BHF*. [Online]
Available at: https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures
[Accessed 18 November 2021].

Centers for Disease Control and Prevention, 2021. *Heart Disease Facts | cdc.gov*. [Online]
Available at: https://www.cdc.gov/heartdisease/facts.htm
[Accessed 15 November 2021].

Ma, H. & Shieh, K.-J., 2006. Cholesterol and human health. *The Journal of American Science,* 2(1), pp. 46-50.

Wilkins, E. et al., 2017. *European Cardiovascular Disease Statistics,* Brussels: European Heart Network.