Heart Disease Classification

Machine Learning 2021-2022

Table of Contents

[1 Introduction 5](#_Toc92316875)

[2 Exploratory Data Analysis (30%) 5](#_Toc92316876)

[2.1 Dataset Information 5](#_Toc92316877)

[2.2 Class Balance 6](#_Toc92316878)

[2.3 Feature Analysis 7](#_Toc92316879)

[2.4 Missing Values 8](#_Toc92316880)

[2.5 Duplicate Rows 8](#_Toc92316881)

[2.6 Noise 9](#_Toc92316882)

[2.7 Correlation Analysis 11](#_Toc92316883)

[3 Method (25%) 12](#_Toc92316884)

[3.1 Aims and Objectives 12](#_Toc92316885)

[3.2 Data Processing 12](#_Toc92316886)

[3.3 Classification Algorithms 12](#_Toc92316887)

[3.4 Configuration and Tuning 12](#_Toc92316888)

[3.5 Validation Method 12](#_Toc92316889)

[3.6 Evaluation Metrics 12](#_Toc92316890)

[4 Results and Discussion (25%) 13](#_Toc92316891)

[4.1 Narrative and Depth 13](#_Toc92316892)

[4.2 Objective Discussion 13](#_Toc92316893)

[4.3 Statistical Significance 13](#_Toc92316894)

[4.4 Real-world Connections 13](#_Toc92316895)

[5 Conclusion (5%) 14](#_Toc92316896)

[5.1 Further Work 14](#_Toc92316897)

Table of Figures

**No table of figures entries found.**

Table of Tables

# 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.

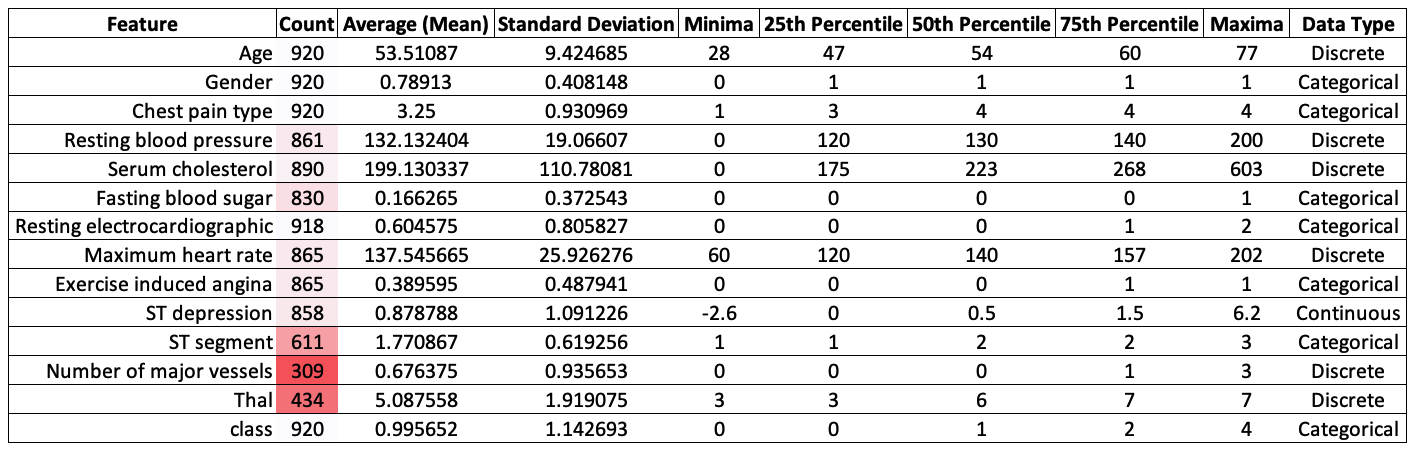
In this investigation, I will implement various machine learning and classification algorithms on existing patient records so the models can be used on unseen data to classify, detect, and quantify the degree of present heart disease.

# Exploratory Data Analysis (30%)

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.

## Dataset Information

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.



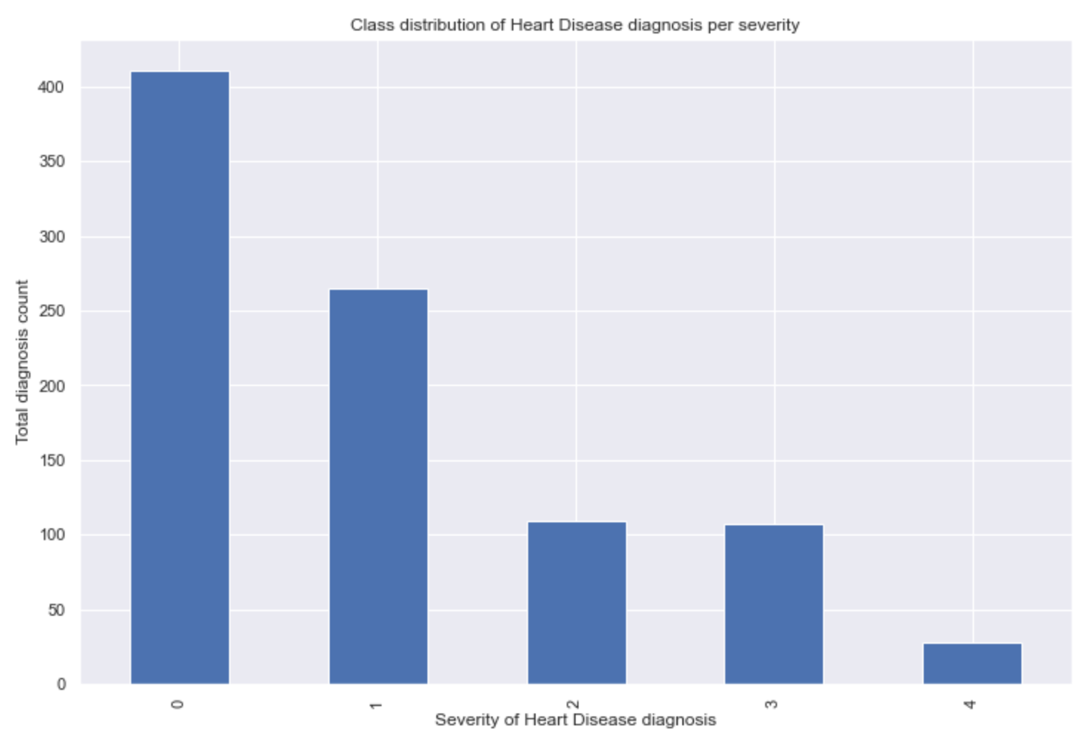
Table

Description automatically generated

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).

## Class Balance

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.



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.

## Chart Description automatically generatedFeature Analysis

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.

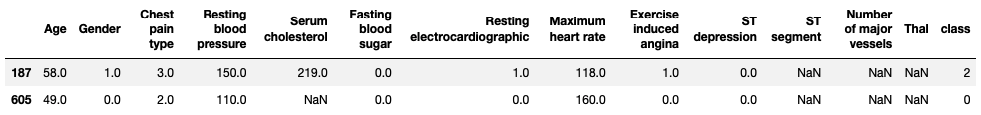
The diagonal histograms give insights on the class distribution, positioning, and skew of each feature. For example, the peaks of each class’ curve for resting blood pressure appear further up the scale with each degree of heart disease presence, signifying those with higher resting heart rates have 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.

## Missing Values

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.

## Duplicate Rows



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.

## Noise

Chart, box and whisker chart

Description automatically generated 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.

Chart, box and whisker chart

Description automatically generated 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.

Chart, box and whisker chart

Description automatically generated 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.

## Correlation Analysis

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. If ‘Number of major blood vessels’ and ‘Thal’ didn’t have so many missing values, they would be good indicators of the classification of heart disease. The best correlation considering missing values is that of Chest pain type and Class. There are no missing values and displays a 49% correlation, meaning Chest pain type is an important feature to consider.

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.

# Method (25%)

## Aims and Objectives

The primary objective of this investigation is to successfully apply machine learning algorithms to accurately detect the presence and classify the severity of heart disease in patients based on body examination results such that successful classification constitutes 90% of total classification predictions.

The secondary objective of this investigation is to explore the effect of scrutiny on features which consist of over 50% missing values, and how it affects the accuracy metrics of the classifiers used.

I am going to investigate the secondary objective by performing the classification on the original dataset and an edited version of the data where features with above 50% of the data is missing are removed.

## Feature Selection

Content

## Data Processing

Content

### Zero Values

Content

### Missing Values in Rows

Content

### Missing Values in Columns

Based on the amount of missing data from these features, they were dropped on the edited dataset. Instances with 5 or more missing values were dropped in the full dataset. To keep the data consistent, these instances were also dropped from the edited dataset. This was done because the features removed in the edited dataset may or may not have contributed towards the thresholds of removing instances based on their missing values.

Text

Description automatically generated K-Nearest Neighbour was selected for data imputation due to its ability to process both continuous and discrete data (Batista & Monard, 2002). It works by estimating what the missing value should be based on similar instances’ data. Other methods such as mean/median imputation and zero/constant value imputation were not considered for use as correlation between features is especially important within medical datasets that concern real-life patients.

Since a patient cannot realistically have cholesterol level of zero, the decision was made to nullify these readings. However, this meant that the ‘Serum cholesterol’ feature now consists of 21.96% null values, which is a significant increase from 3.26%.

## Feature Selection

God help me

## Classification Algorithms

The data in this dataset is labelled with a classification diagnosis of degree of heart disease. As such, the supervised models and algorithms K-Nearest Neighbour, Random Forest, and Naïve Bayes were selected to classify the instances in this dataset.

K-Nearest Neighbour was selected

Neural Net

Random Forest

## Configuration and Tuning

The grid search method was used to select hyperparameters and run the models with them. Grid search uses an iterative loop to compare the results of models using different parameters each time they are run, resulting in a near-optimum set of hyperparameters to be used by the final model.

## Validation Method

Content

## Evaluation Metrics

Content

# Results and Discussion (25%)

Content

## Narrative and Depth

Content

## Objective Discussion

Content

## Statistical Significance

Content

## Real-world Connections

Content

# Conclusion (5%)

Content

## Further Work

Content

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.