**Table of Contents**

|  |  |  |
| --- | --- | --- |
| [**Introduction, Problem Statement and Research Questions.**](#page3) | | [**2**](#page3) |
| [Introduction:](#page3) | | [2](#page3) |
| [Problem Statement:](#page3) | | [2](#page3) |
| [Potential Research Questions:](#page3) | | [2](#page3) |
| [**Literature Review.**](#page4) | | [**3**](#page4) |
| [**Problem Approach.**](#page6) | | [**7**](#page6) |
| [**The Data.**](#page7) | | [**8**](#page7) |
| [1.](#page7) | [Source and Dataset Description:](#page7) | [8](#page7) |
| [2.](#page9) | [Cleaning the Dataset:](#page9) | [11](#page9) |
| [3.](#page9) | [Data Exploration and Analysis:](#page9) | [11](#page9) |
| [**Appendix**](#page15) | | [**17**](#page15) |
| [**References**](#page16) | | [**18**](#page16) |

1

Introduction, Problem Statement and Research Questions.

**Introduction:**

According to the World Health Organization (WHO), the two most fatal diseases around the world are heart related, making the jobs of healthcare specialists critical and the accuracy of precautionary testing extremely important. The Healthcare industry sits on a pool of untapped potential for predicting and preventing such fatal diseases through their collection of data recorded in clinical notes and Electronic Health Records (EHRs). This potential can be utilised through Machine Learning (ML) techniques used for diagnosis and outcome prediction, including the identification of high risk for medical emergencies such as relapse or transition into another disease state.

During this project, we shall aim to study datasets related to Heart Disease and underlying attributes that would possibly indicate the development of the disease while also testing ML algorithms to predict the prevalence of the disease.

**Problem Statement:**

“We shall analyse the various risk factors for Heart Disease and apply various statistical and computational models to predict the probability of Heart Disease in individuals.”

**Potential Research Questions:**

* What are the most important metrics for consideration?
* What trends are common with respect to risk factors? Does any particular risk factor influence another?
* Which machine learning models provide the best results in terms of predictions?
* What interesting correlations exist between the dataset variables?

2

Literature Review.

1. International Application of a New Probability Algorithm for the Diagnosis of Coronary Heart Disease:

The study aimed to test the reliability and clinical utility of a probability algorithm (Discriminant Function) derived from the clinical and test characteristics of a group sample of 303, could accurately predict the probability coronary artery disease (CAD) in other different sample groups, while also comparing the derived algorithm to an existing Bayesian method.(Detrano et al., 1989) The study found that over/under-estimation were prevalent in both methods applied, though significantly lower for the DF method. Regarding clinical utility, the percentage of correctly classified patients was modestly superior in the DF method compared to the Bayesian method. One of the things we have picked up across the literature reviews is that no one single algorithm is the superior predictor across the health domain; for example the Bayesian model was superior in the Diabetes study by Sisodia (Appendix 1)(Sisodia & Sisodia, 2018).

1. Multivariate Analysis of Risk Factors for Coronary Heart Disease:

The study samples men from Sweden, who did not exhibit signs of CHD. Using a multiple logistic model, the researchers analysed nine probable risk factors and found that six of the nine probable factors increased the risk of CHD significantly. Amongst the six factors, the researchers determined that three of them including cholesterol, smoking and systolic blood pressure are closely related to CHD. (WILHELMSEN et al., 1973) The predictive power of the logistic model applied in another randomly selected population also found that the model accurately defined the groups according to risk level and manifestations of CHD. Given our study aims to predict and classify Heart Disease, we will use this study as a basis of selecting and explaining attributes that correlate with Heart Disease.

1. Further research on multivariate analysis:

The study utilised a multivariate Cox-proportional-hazards regression model to identify the risk factors associated with the development of CHD. The results of the study showed that several factors were strongly associated with an increased risk of developing CHD including age, male sex, smoking, diabetes and family history of premature CHD. On the other hand, factors that lowered the risk included moderate alcohol consumption and regular physical activity. The analysis also found that the presence of multiple risk factors, such as hypertension, diabetes, and high total cholesterol, increased the risk of CHD in a synergistic manner as well. Overall, the study demonstrated the importance of multivariate analysis in identifying the complex interplay between different risk factors for CHD. The findings of the study have important implications to make sure our model looks at multiple risk factors simultaneously (Preis et al., 2009).

1. Artificial intelligence in coronary computed tomography angiography: Demands and solutions from a clinical perspective:

The article reviews several studies that have evaluated the performance of AI algorithms in cardiovascular imaging. For example, one study evaluated its use in analysing cardiac MRI images for the detection of myocardial infarction (MI). The study found that the deep learning algorithms had a high accuracy in detecting MI, with a sensitivity and specificity of 91.5% and 92.5%, respectively. Another study evaluated the use of AI algorithms to analyse coronary computed tomography angiography (CCTA) images for the detection of significant stenosis. The study found that the AI algorithm had a higher accuracy than traditional methods.

3

The article also discusses the potential clinical applications of AI in cardiovascular imaging reducing the need for manual interpretation by radiologists or cardiologists. AI algorithms could also be used to predict patient outcomes, such as the risk of heart failure or mortality, based on imaging data and other clinical information. However, the article also highlights several challenges associated with the implementation of AI algorithms in cardiovascular imaging. These include the need for large datasets for training and validation, the need for standardised imaging protocols, and the potential for biases in the algorithms based on the demographics of the training data. Overall, the article provides an alternate view to the structured data we will be using in our analysis and shows the future directions of AI in cardiovascular imaging, highlighting both the potential benefits and challenges associated with its implementation (Baeßler et al., 2023).

1. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure:

The study recruited both CHD cases and controls, from two large-scale population-based cohorts in Sweden. The participants underwent a clinical examination, including blood tests, electrocardiogram (ECG), and echocardiography. Additionally, genetic data were obtained from each participant using genotyping arrays. The study utilised a machine learning algorithm known as the Elastic Net, which is a type of logistic regression that is commonly used for feature selection and classification in high-dimensional datasets. The results of the study showed the machine learning algorithm was able to accurately identify CHD cases using clinical and genetic data. The algorithm achieved an area under the receiver operating characteristic (ROC) curve of 0.72, which is considered to be a moderate-to-good performance for a diagnostic test. The study also evaluated the contribution of different clinical and genetic variables to the performance of the algorithm. The analysis showed that the strongest predictors of CHD were traditional clinical risk factors similar to the other studies above. However, the inclusion of genetic variables in the algorithm improved its performance, suggesting that genetic information may provide additional predictive value for CHD risk. The study has several strengths, including its large sample size and the use of a machine learning algorithm that is well-suited for high-dimensional datasets. However, the limitations include the fact that the algorithm was tested in a single population-based cohort. Overall, the study provides evidence for the potential utility of machine learning algorithms in identifying CHD cases using clinical and genetic data (Luo et al., 2022).

4

Problem Approach.

The overall approach to this problem will follow the CRISP-DM methodology:

Stage 1, Business Understanding and Data Understanding:

As part of this stage, we have begun to understand the data in both a technical manner as well as from a business perspective. Going forward, we shall better apply our understanding to determine additional goals and outcomes for this project, aside from our initial research questions and problem statements.

Stage 2, Data Preparation:

Having understood the state of the dataset and the requirements for completion of our project, we shall embark on further cleaning and pruning of the dataset to remove errors that would hinder our understanding of the data and our outcome. We will also visualise our data to understand the important variables and use it to discover any insights that might be useful for our research question.

Stage 3, Modelling:

Applying both descriptive and inferential statistical approaches, we shall decide on which machine learning models would represent the best predictions and give us the most accurate results. Researching the various models that have already been used and novel new approaches will be a part of this stage as well, helping analyse various trends in the data. Training, predicting and tuning of the dataset and its parameters will also be a part of this stage.

Stage 4, Evaluation:

The final stage of this project will involve evaluating the models we’ve chosen as well as the parameters we have used to finalise on the deployment we wish to move forward with.

5

The Data.

1. **Source and Dataset Description:**

We have utilised UCI Machine Learning Repository to source our Heart Disease Data Set. While most research work focuses on a subset of 14 fields, for the purposes of this project, the entire dataset was used which has 76 distinct attributes including the target variable.

Furthermore, the dataset is a combination of 4 different sub-datasets, each of which contains data from a different geographical group. Most research is focused on just the Cleveland dataset. So by combining the datasets, we hope to extend the body of work and gain a more comprehensive dataset to aid our understanding of the overall effect and predictions that the various parameters have. Each individual dataset group contains between 124 and 293 rows of data each, for a total of 903 entries.

|  |  |
| --- | --- |
|  | **Data Dictionary** |
|  |  |
| **Variables** | **Description** |
|  |  |
| age | age in years |
|  |  |
| sex | sex (1 = male; 0 = female) |
|  |  |
| cp | chest pain type |
|  | -- Value 1: typical angina |
|  | -- Value 2: atypical angina |
|  | -- Value 3: non-anginal pain |
|  | -- Value 4: asymptomatic |
|  |  |
| trestbps | resting blood pressure (in mm Hg on admission to the hospital) |
|  |  |
| chol | serum cholesterol in mg/dl |
|  |  |
| cigs | cigarettes per day |
|  |  |
| years | number of years as a smoker |
|  |  |
| fbs | fasting blood sugar > 120 mg/dl (1 = true; 0 = false) |
|  |  |
| dm | 1 = history of diabetes; 0 = no such history |
|  |  |

6

|  |  |
| --- | --- |
| famhist | Family history of coronary artery disease (1 = yes; 0 = no) |
|  |  |
| restecg | Resting electrocardiographic results |
|  | -- Value 0: normal |
|  | -- Value 1: having ST-T wave abnormality |
|  | -- Value 2: showing probable left ventricular hypertrophy by Estes' criteria |
|  |  |
| prop | Beta blocker used during exercise ECG: 1 = yes; 0 = no |
|  |  |
| nitr | Nitrates used during exercise ECG: 1 = yes; 0 = no |
|  |  |
| pro | Calcium channel blocker used during exercise ECG: 1 = yes; 0 = no |
|  |  |
| diuretic | Diuretic used used during exercise ECG: 1 = yes; 0 = no |
|  |  |
| thalach | Maximum heart rate achieved |
|  |  |
| thalrest | Resting heart rate |
|  |  |
| tpeakbps | Peak exercise blood pressure (first of 2 parts) |
|  |  |
| tpeakbpd | Peak exercise blood pressure (second of 2 parts) |
|  |  |
| trestbpd | Resting blood pressure |
|  |  |
| exang | Exercise induced angina (1 = yes; 0 = no) |
|  |  |
| oldpeak | ST depression induced by exercise relative to rest |
|  |  |
| slope | The slope of the peak exercise ST segment |
|  | -- Value 1: upsloping, Value 2: flat , Value 3: downsloping |
|  |  |
| ca | Number of major vessels (0-3) coloured by fluoroscopy |
|  |  |
| thal | 3 = normal; 6 = fixed defect; 7 = reversible defect |
|  |  |
| num | Diagnosis of Heart Disease (angiographic disease status) |
|  | -- Value 0: < 50% diameter narrowing |
|  | -- Value 1: > 50% diameter narrowing |
|  | (in any major vessel: attributes 59 through 68 are vessels) |
|  |  |

7

**2. Cleaning the Dataset:**

In order to cleanse the dataset we explored the data and in conjunction with the requisite documentation on the UCI Machine Learning Repository to identify any missing or unclean values.

The first stage in cleaning the dataset was to eliminate any missing values. We eliminated these rows from the data because according to the documentation null values are denoted as “-9”, therefore any missing values indicated incomplete data. Additionally, while loading in the dataset, we noticed an encoding error with specific entries in each dataset. Studying the dataset revealed a pattern in the end of file values that were causing dummy inputs, and so these rows were dropped as well.

In the second stage we examined the fields that contain a single value or a dummy variable. Some of these were explained in the documentation as being populated with dummy values for privacy reasons such as the name and social security number of the patient. SInce it will add no value to the models we will build we have eliminated them which reduced the number of variables from 76 to 26 which we have listed as “important\_titles” and provided in the data dictionary above. If we find these are important variables or a way to clean them we can add them back in a later stage.

The list of dropped columns is as follows:



***'rcadist', 'ramus', 'cday', 'htn', 'diag', 'thalsev', 'cmo', 'om1', 'cathef', 'lvx4', 'cxmain', 'thaldur', 'restckm', 'junk', 'thaltime', 'earlobe', 'exerckm', 'ekgday', 'ekgyr', 'painloc', 'lvf', 'smoke', 'laddist', 'rcaprox', 'rldv5', 'cyr', 'lvx2', 'exerwm', 'xhypo', 'relrest', 'proto', 'lvx1', 'ekgmo', 'dig', 'thalpul', 'id', 'lmt', 'rldv5e', 'painexer', 'lvx3', 'pncaden', 'exeref', 'ladprox', 'restef', 'om2', 'restwm', 'dummy', 'ccf', 'met'***

Lastly, we use PandasProfiling to identify null and alternative junk/dummy values in the dataset (denoted by ‘-9’) and eliminate or replace these values with 0 so as to not impact the machine learning models we will go on to train.

**3. Data Exploration and Analysis:**

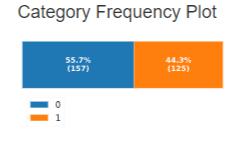
The next stage involves running various statistical and visual tools to understand the correlation, distribution and variations in the data.

Tools that were used for this include the generation of histograms, bar charts, distribution graphs and correlation matrices, however, for ease of access and neater representation, we have used PandasProfiling and PlotNine, which are popular libraries for such tasks.

8

After cleaning the dataset thoroughly, we applied the PandasProfiling libraries to build statistical models and figures to gauge the correlation between the various elements and prove our data cleaning processes.

The Cleveland dataset appeared to be the cleanest in values and form, and so we have focused our efforts on exploring it first, applying what we learn to the other datasets to streamline the cleaning process.

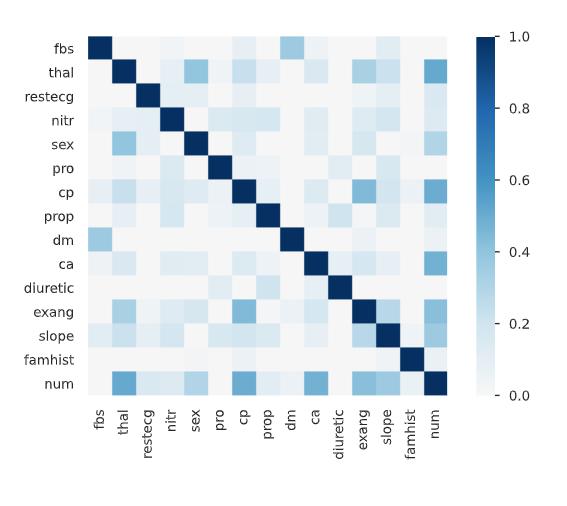


*(Number of patients with heart condition)*

The Cleveland dataset is fairly balanced in expressing the target class (the diagnosis of Heart Disease). Out of 282 patients, 55.7% did not have Heart Disease whilst 44.3% do have Heart Disease. This will assist in training the model accurately without the need for statistical adjustments such as oversampling that is required for unbalanced datasets.

Something observed through our research was that the prevalence of heart, stroke or vascular disease in the Australian population was only 6.25% (*Heart, stroke and vascular disease: Australian facts, coronary heart disease* 2023) which indicates the sample might not bereflective of the wider population. Depending on how the sample was created, we will need to be careful before translating any insights or predictions to the general population.

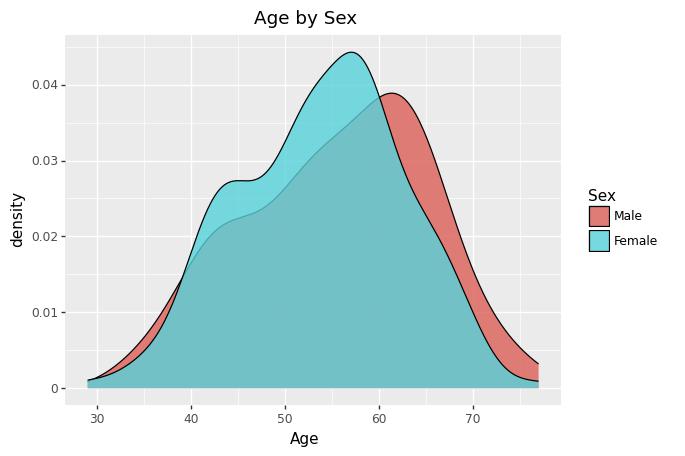
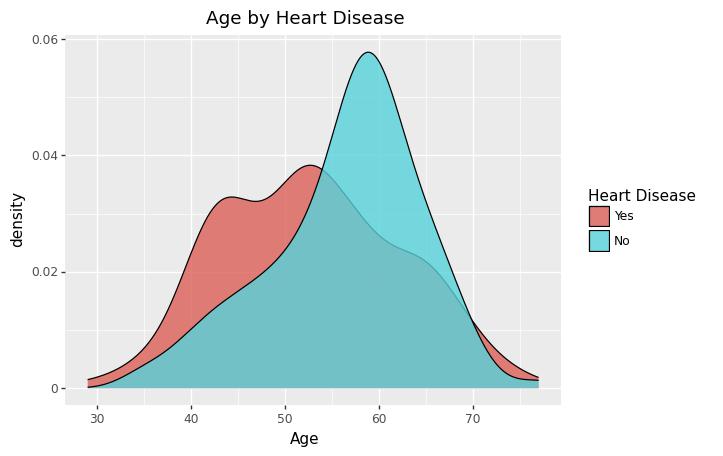
9



*(Correlation Matrix)*

The correlation matrix shows the strength of the correlation between variables within data. For example, *num* which is the target class of whether a patient has Heart Disease, shows some correlation with *thal* (Max Heart Rate), *cp* (Chest Pain) and *ca* (Number of major vessels). We will explore these interesting variables in more depth below.

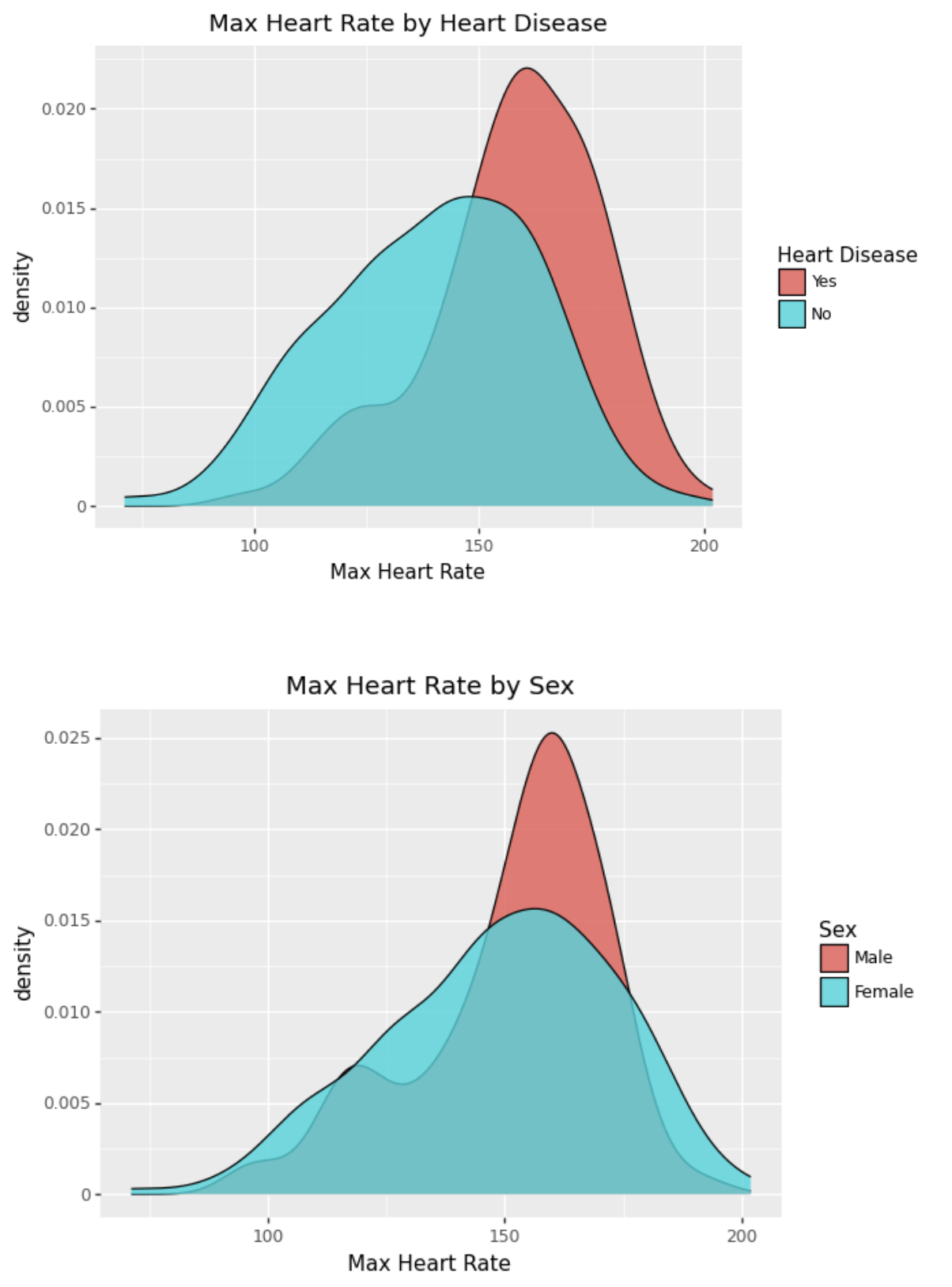
10



*(Numeric Variable Distribution Graphs)*

When observing the age distribution it appears that the male and females are distributed fairly similarly in the data with the peak at a slightly younger age for women. However, when it is displayed by whether the patient had Heart Disease or not, the distribution was very different suggesting this might be an important variable when predicting Heart Disease.

11

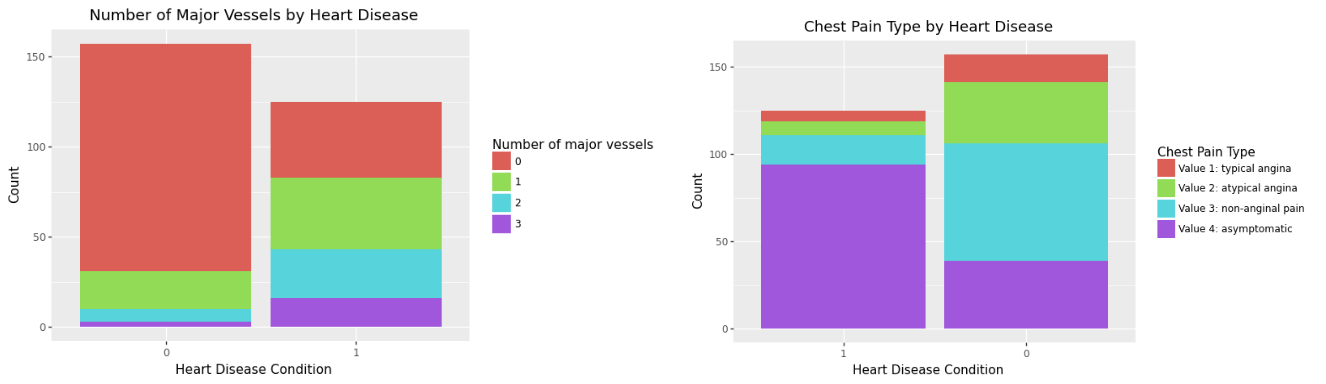


*(Max Heart Rate)*

When observing the distribution by max heart rate, segmented by Heart Disease condition, it appears those with Heart Disease have higher mean heart rate and a more

12

Leptokurtic distribution. Similarly when segmenting by gender, males tend to have a taller and skinnier distribution.



*(Categorical Bar Charts)*

To explore the categorical values, bar charts were created to observe the difference in the target class. We can observe that patients with Heart Disease appear to show a higher frequency of a higher number of major blood vessels. When looking at chest pain type, results that seemed counterintuitive appeared; patients with Heart Disease mostly displayed asymptomatic chest pain. More exploration is required to confirm whether these observations are significant.

13

**Appendix**

**Literature Reviewed**

**Appendix 1 -Prediction of Diabetes using Classification Algorithms:**

In this study, the researchers focused on classifying pregnant women suffering from

diabetes. Naive Bayes, SVM, and Decision Tree machine learning classification algorithms are used and evaluated on the dataset to find the prediction of diabetes in a patient. The researchers evaluated eight attributes, measuring the accuracy, precision, recall, F-measure of each algorithm’s performance. (Sisodia & Sisodia, 2018) The study found that the Naive Bayes outperforms the other algorithms, and this is verified using the ROC plot. The key aspects that pertain to our study are the performance parameters, ROC evaluation and the use of WEKA tool. We can use similar parameters used in this study to evaluate its effectiveness with predicting CAD, moreover, we can verify the tests using the ROC plot as used by the researchers in this study. Finally, as mentioned, the WEKA tool was used in this study for performing the experiments and it was found useful for personalising according to requirements, hence adaptable for our experiments. (Sisodia & Sisodia, 2018)

**Appendix 2 - Machine learning in Medicine: a practical introduction:**

This paper demonstrates the use of machine learning techniques by developing three predictive models for cancer diagnosis using descriptions of nuclei sampled from breast masses. (Sidey-Gibbons & Sidey-Gibbons, 2019) The study uses General Linear Regression, Support vector machines and artificial neural networks on the datasets, testing for benign or malignant tumours and evaluating based on sensitivity, specificity, and accuracy. The results showcased that all three algorithms proved effective in classifying the nuclei samples, SVMs being marginally better than the other algorithms. (Sidey-Gibbons & Sidey-Gibbons, 2019) Like our study, this research uses an open data set, applying the algorithms to nine features, a relatively simple amount to efficiently compute. Therefore, the principles used in this study can be applied to our datasets and with the successful performance demonstrated by the algorithms in classifying the disease, we may use the algorithms’ open-source software in our study to test its effectiveness in classifying Heart Disease.

14

**REFERENCES**

1. Preis, S. R., Pencina, M. J., Hwang, S.-J., D'Agostino, R. B., Savage, P. J., Levy, D., & Fox, C. S. (2009). Trends in cardiovascular disease risk factors in individuals with and without diabetes mellitus in the Framingham Heart Study. *Circulation*, *120*(3), 212–220. https://doi.org/10.1161/circulationaha.108.846519
2. Baeßler, B., Götz, M., Antoniades, C., Heidenreich, J. F., Leiner, T., & Beer, M. (2023). Artificial Intelligence in coronary computed tomography angiography: Demands and solutions from a clinical perspective. *Frontiers in Cardiovascular Medicine*, *10*. https://doi.org/10.3389/fcvm.2023.1120361
3. Luo, C., Zhu, Y., Zhu, Z., Li, R., Chen, G., & Wang, Z. (2022). A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. *Journal of Translational Medicine*, *20*(1). https://doi.org/10.1186/s12967-022-03340-8
4. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K., Lee, S., &amp; Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310. <https://doi.org/10.1016/0002-9149(89)90524-9>
5. Sidey-Gibbons, J. A., &amp; Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. BMC Medical Research Methodology, 19(1). <https://doi.org/10.1186/s12874-019-0681-4>
6. Sisodia, D., &amp; Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585. [https://doi.org/10.1016/j.procs.2018.05.12](https://doi.org/10.1016/j.procs.2018.05.122)2
7. WILHELMSEN, L. A. R. S., WEDEL, H. A. N. S., &amp; TIBBLIN GOSTA. (1973). Multivariate analysis of risk factors for coronary heart disease. Circulation, 48(5), 950–958. <https://doi.org/10.1161/01.cir.48.5.950>
8. Australian Institute of Health and Welfare. (n.d.). Heart, stroke and vascular disease: Australian facts, coronary heart disease. AIHW, Australian Government. Retrieved March 25, 2023, from https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/sum mary-of-coronary-heart-disease-and-stroke/coronary-heart-disease

15