Investigating Heart Disease Datasets and Building Predictive Models

by

Brandon Simmons II

A Thesis submitted to the Graduate Faculty of
Elizabeth City State University
in partial fulfillment of the
requirements for the Degree of
Master of Science in
Applied Mathematics.

May

2021

APPROVED BY

Julian A. D. Allagan, Ph.D.
Committee Chair

Kenneth L. Jones, Ph.D.
Committee Member

Mohammed Talukder, Ph.D.
Committee Member

Gabriela H. Del Villar, Ph.D.
Committee member

ABSTRACT

We investigate several heart disease datasets commonly found on popular data sites such as Kaggle, Dataport and UCI machine learning repository. We discovered many issues in our attempts to authenticate these medical datasets as they relate to human errors (encoding) and sometimes negligence (duplicates); these underlying issues have undoubtedly weakened many inferences or predictive models built on some of the datasets that are already published. We addressed these issues through features analysis. Further, using Random forest and logistic regressions, we determine the best dataset for machine learning and statistical analysis: the Cleveland data on a reduced set of six features. Three of which are statistically significant at explaining or classifying patients as 'Heart Disease'. They are thalach (maximmum heart rate), oldpeak and cp (chest pain).

DEDICATION

This thesis work is dedicated to those past, present and future for the opportunity to keep learning.

ACKNOWLEDGEMENT

My humblest thanks to my professors, for their guidance and tutelage throughout my time in the program.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Problem Description

Everything from organic to inorganic has been designed with a heart structure to aid in fulfilling daily functions. Humanity has utilized this same format with advancing technology today. In fact, the Center for Disease Control and Prevention (CDC) in 2018 classifies heart disease as the leading cause of mortality in the United States and remains the leading cause of mortality to date [18]. The previous research invested in developing information surrounding the heart has sought to improve the methods of managing our heart's condition [2]. With all the data being produced, machine learning aids with patient-level observations, where algorithms sift through vast numbers of variables, looking for combinations that reliably predict outcomes [22]. Since the conclusion of a 2018 research study, the American Heart Association website posted that there has been a 15.1 % decrease across the United States [2].Further, according to CDC, in the United States, someone has a heart attack every 40 seconds [1] and every year, about 805,000 Americans have a heart attack[3]. Further, about half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease namely high blood pressure, high cholesterol, and diabetes [2].

Due to the complexity and the variations of the increasing number of risk factors, modern researchers are relying on Data Mining and Machine Learning techniques. Because of privacy concerns and other issues related to accessing public data, there are very few heart disease data sets available for the public to analyze. One these sites is Kaggle. Kaggle is a very popular site which allows users to find and publish data sets, explore and build models in a web-based data-science environment, and collaborate with other data scientists and machine learning engineers, and even enter competitions to solve data science challenges. In our quest for health related data, we found on the site [16] that, heart.csv data (11.06 Kb) has been one of the most analyzed by avid data analysts/scientists around the world. Currently, [as of 03/17/2021, 3:00pm U.S.A E.T] this data has 1,321,009 views, 198,520 downloads, 1,567 notebooks.—This number continues to grow each day. This demonstrates both the importance and interest in this topic and hence the need of insuring that the data is correct.

We explored some of the reported analysis results on the data. We found that, most of the analysis were not done on a well-processed data. For instance, the variables 'sex' and 'cp' (Chest Pain) were kept as numeric instead of converted to categorical. Further, these analysis often involve machine learning codes which return a classifier output without enough information for the reader to accurately gauge the features' importance or the classification's criteria.

As we attempt to interpret some of the results further, we were getting the opposite of what is expected. For instance, we found out more younger individuals were getting classified as having heart disease compared to older individuals which was a major red flag. Upon investigating the data source from the University of California, Irvine, Machine Learning Repository[8], as reported in Kaggle, we could not tell which of the four datasets were used. The most obvious of the four (since the other three were grossly incomplete and require extensive data processing) which we believed was used in Kaggle is the Cleveland data. Still, this data target variable is nominal with 5 levels, and contains some missing values, compared to the Kaggle data which target variable is binary and contains no missing value.

For a medical record to be analyzed and yield an accurate result, it needs to be authentic and verifiable. This is necessary to inspire confident in the population if we wish to inform individuals and stakeholders or recommend certain behavioral changes. This work investigates the Kaggle data content to determine its authenticity, any issue related to its target variable as we compare its features to those of the Cleveland data. Further we found two other heart disease data sources. One (Statlog) is very similar to the Cleveland data although a bit smaller in size while the other (Orig) is supposed to be a combination of five datasets of heart disease; two of which are the Cleveland, the Statlog and the remaining (3) cannot be found from the source UCI repository [8, 10].

Our work shows how to fix the Kaggle data as it becomes essentially the Cleveland data. We determine if there is any relation between the Statlog data and the Cleveland data. Then, we identify both of these data from the most recently published (combined) data on heart disease. Finally, we analyze these data using logistic regression along with random forest feature selections to determine the best models with the least AI criterion. We conclude that, the Cleveland data on a reduced (features) model is the best statistical model.

1.2 Research Questions and Approaches

This research attempts to:

1. Investigate one of the most popular heart data that has been studied by many data scientists around the world in Kaggle. We found that, although the data is supposed to have originated form the University of California Irvine Machine Learning Repository archive, the predictor's encoding is backward. This lead to misleading interpretations and incorrect conclusions about the features relation with the target variable. However, this data is very similar to Cleveland data which is also found in the UCI repository. Except, the Cleveland data has 6 missing values and the target variable has 5 levels. With the right encoding, we conclude that Kaggle (with target 1=No disease; 0=Disease) is the Cleveland. Still, we found another complete dataset in another archive of the UCI repository, called Stalog which is similar to Cleveland except in size.

2. Compare the Cleveland data (303 obs.) to the Statlog data (270 obs.). This ensures one is not derived from the other. We use "Pandas Profiling" to access overall data content such as size and variable names, types and their distributions. Then, we compare pairwise, the distribution of each feature, and each feature against the target variable. We see very similar associations between

each feature and the target and upon further investigations, we found that the Statlog data is a proper subset of the Cleveland data. Thus, the Cleveland data= Statlog +33 new records. We drop the Statlog data and kept the Cleveland data for further analysis.

3. Find the remaining three datasets that were corrupted in the UCI dataset repository. To do so, we needed to first identify the Cleveland (and thus the Statlog datasets) from the most recently published data, ORIG—We merge the Cleveland data with ORIG data and remove all duplicated records. In which case, any record that appears in both Cleveland, and ORIG is removed. The remaining dataset is called MISS and perhaps represents the (3) missing datasets that were deemed corrupt in the UCI repository. Unfortunately, we are unable to authenticate the origin of the ORIG data. Despite our attempt to reach out to the originator, we received no response. So, we proceed to further investigate the content of the MISS data. We found that about 25% (303 observations) are duplicated records.

4. Investigate the MISS data by exploring the distribution of its features, and the features associations with the target variable. We found that the data has skewed significantly the gender, and unually high records of '0' values for cholesterol level.

5. Analyze each of the three datasets: CLEVELAND, MISS, and STATLOG. We build a logistic regression on each of the data, and using Random Forest algorithm we select the most important features from each data against the target. The process led to building some reduced models and determine the best model for statistical analysis.

1.3 Literature Review

The data sets conducted on heart disease may show different results when applying the machine learning techniques to sift through through the records [22].

This topic is prevalent to the current leading mortality rate. When reviewing the data recorded on the CDC for the country wide statistics, the records reflect a 2016-2018 for all diseases currently affecting the nation in 2020-2021 rise of the covid crisis on the USA Facts report [18].

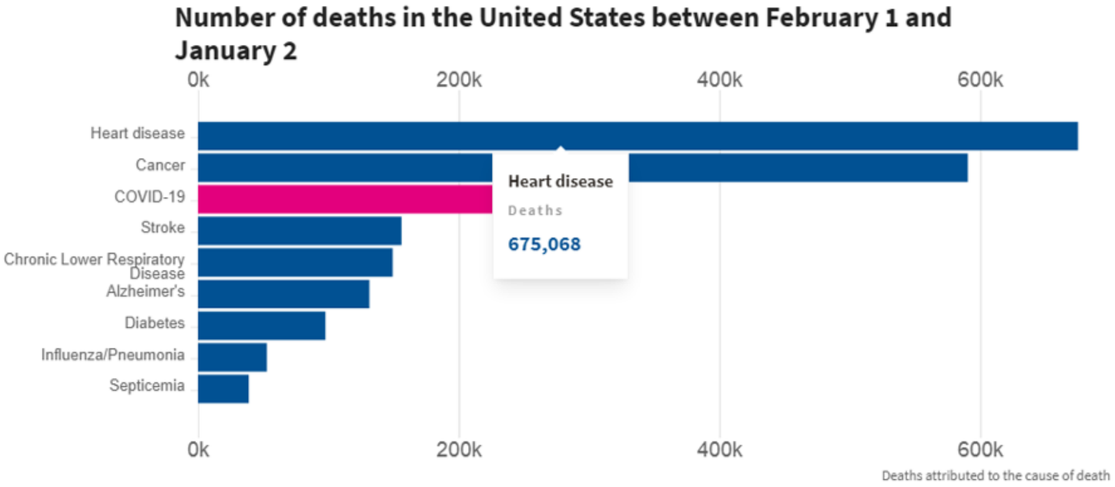


Figure 1.1: USA Facts Published Statistics for the Current 2020-2021

I have reviewed a handful of research articles dealing with a particular heart disease data set that predicts the diagnosis of the patients. Many researchers delved into the a plethora of ways to analyze that data. Some have conducted exploration through principal component analysis and clustering methods to improve the performance of clustering methods resulting in the gaining insight from the sensitivity, specificity, and accuracy of heart disease data [21]. In another case from a particular dataset of 303 patients and 54 attributes, the approach of applying the utilization of data mining and feature creation algorithms with the intent to achieve methods of high accuracy to enrich the dataset [1]. Other data mining techniques techniques to process of sizeable data to provide organizational display [23]. The research where classification algorithms and associative methods are incorporated to check the accuracy of the data inputs.

1.4 Background

Since the primary population that was sampled in all the data used in this research comes from the United States, we research the heart disease trends and statistics as reported by the CDC. The CDC the United States Heart Disease mortality rate records span between 2016-2018.

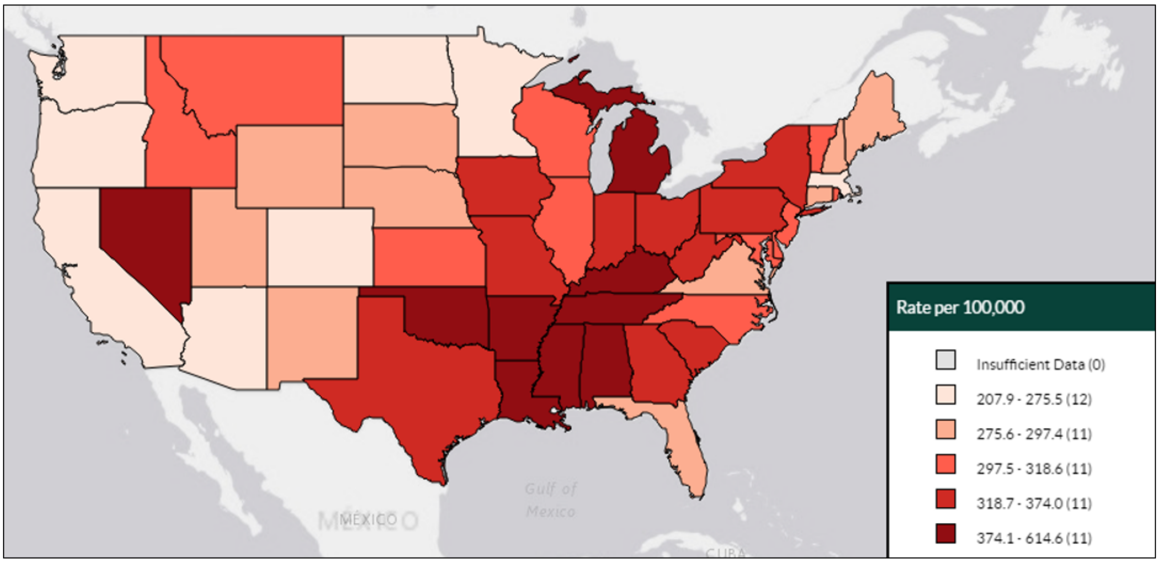


Figure 1.2: CDC United States Statistics for Heart Disease Deaths by State 2016-2018

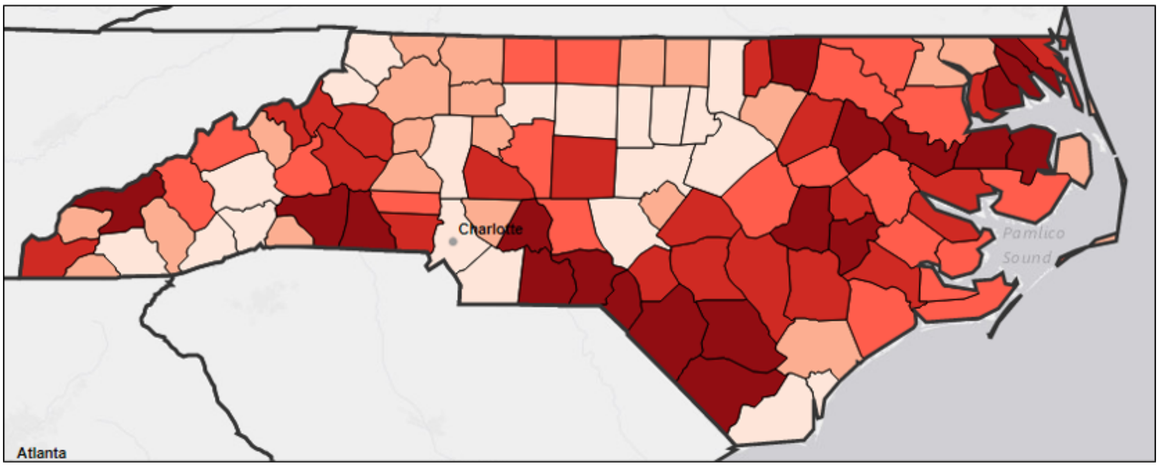


Figure 1.3: CDC North Carolina Statistics for Heart Disease Deaths by County 2016-2018

To begin there are many sub-classes of heart disease which complicates the directed influence of my search with regards to American Heart Associations claim.

Interestingly enough the general term of "heart disease" is categorized into four major subdivisions which are as follows: Coronary Artery Disease, Arrhythmia, Heart Valve Disease, and Heart failure[12]. The research invested in developing information surrounding the heart has circulated to improve methods of managing the heart's conditions[19]. I will briefly cover these major subdivisions. A summary of existing studies have grouped them based on three criteria.

1.4.1 Criteria 1: Blood Flow

Starting the most common of the heart disease types is Coronary heart disease (CHD) also known as Coronary Artery disease (CAD) (National Heart, Lung, and Blood Institute, 2020). For the sake of consistency I will utilize CAD as the appropriate description of this disease. CAD is defined as the obstruction of the blood flow to the heart muscles through the coronary arteries by plaque accumulation on the walls which indefinitely leads to heart attacks[12]. The CAD mortality rate visually presents the stats for 35 and older for both sexes and all ethnicity [3]. North Carolina falls in the range 148.6 - 167.9 with the estimated value of 165.2.

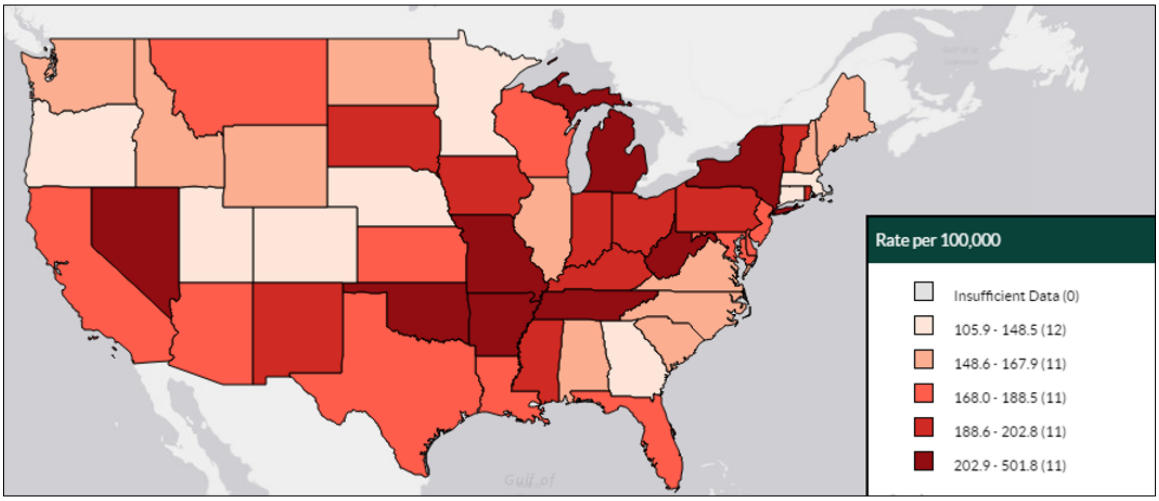


Figure 1.4: CDC United States Statistics for CHD Deaths by State 2016-2018

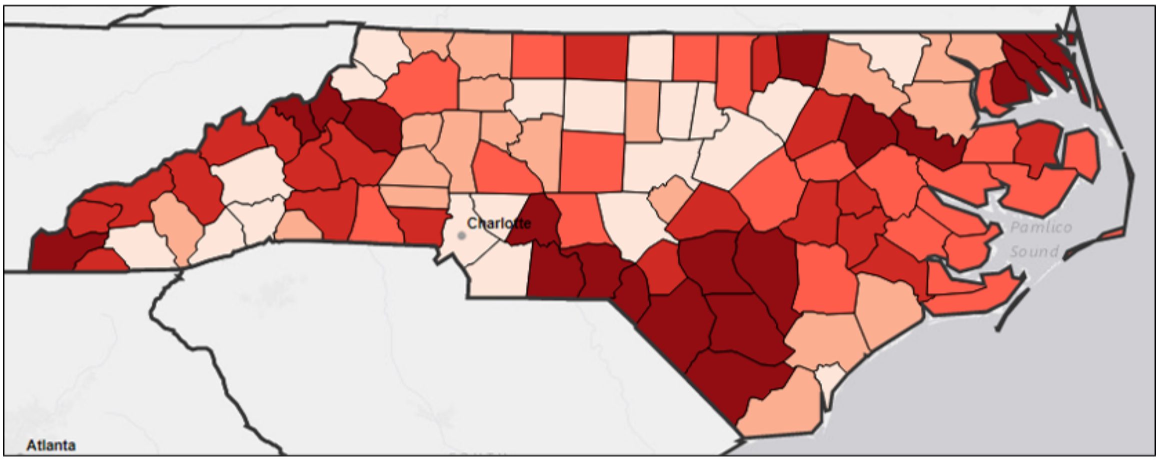


Figure 1.5: CDC North Carolina Statistics for CHD Deaths by County 2016-2018

1.4.2 Criteria 2: Structure

Heart Valve Disease (HVD) involves the issues occurring with the four values that renew and direct blood flow for a smooth transition within the circulatory system[12]. In essence, when one or more of the valves is improperly functions then the patient's condition is classified as HVD. Symptoms include stenosis, which is the limiting of muscle mobility due to valves fusing; Valve Leakage which causes blood to flow back into the heart; and lastly Atresia, which absence of valve opening[12].

1.4.3 Criteria 3: Overall Function

The second common heart disease is Arrhythmia. Arrhythmia is the electrical impulses that initiate the pumping function of the heart muscles. The issue of heart palpitation irregularities which results in broad array of symptoms[12].

Heart Failure which bears a misconception due to the naming connotations not specifically dealing with the heart's inability to function[12]. It is the ramifications stemming from the combinations of the aforementioned diseases limiting the capacity to facilitate function for rest of the body. The primary symptoms resulting from weakness developing is duly to the effects of a patient's first encounter with CAD. A person born with a heart with weak/faulty valves, walls or blood vessels, is said to have a congenital heart defect.

1.4.4 Contributing Factors

Upon review of further literature, the main criteria that is represent in the data sets are blood flow. It is apparent these following factors contribute to the number of heart disease cases within the United States in general. The factors of age, sex, and geographic location increase the risk of a heart attack or stroke[19]. It is also interesting that our heart age may be higher than our actual age due to the influences of diet, stress, activity and heritage. Until the age of 45 years old and older, medical personnel prescribe the public to have a routine heart health check at least every two years. The condition of participants within the data sets pertains to the data retrieved from testing. At the time of data collection traditional risk factors for CAD are cholesterol levels, blood pressure, blood sugar, and pain developed from exercise recorded during the thallium test [13].

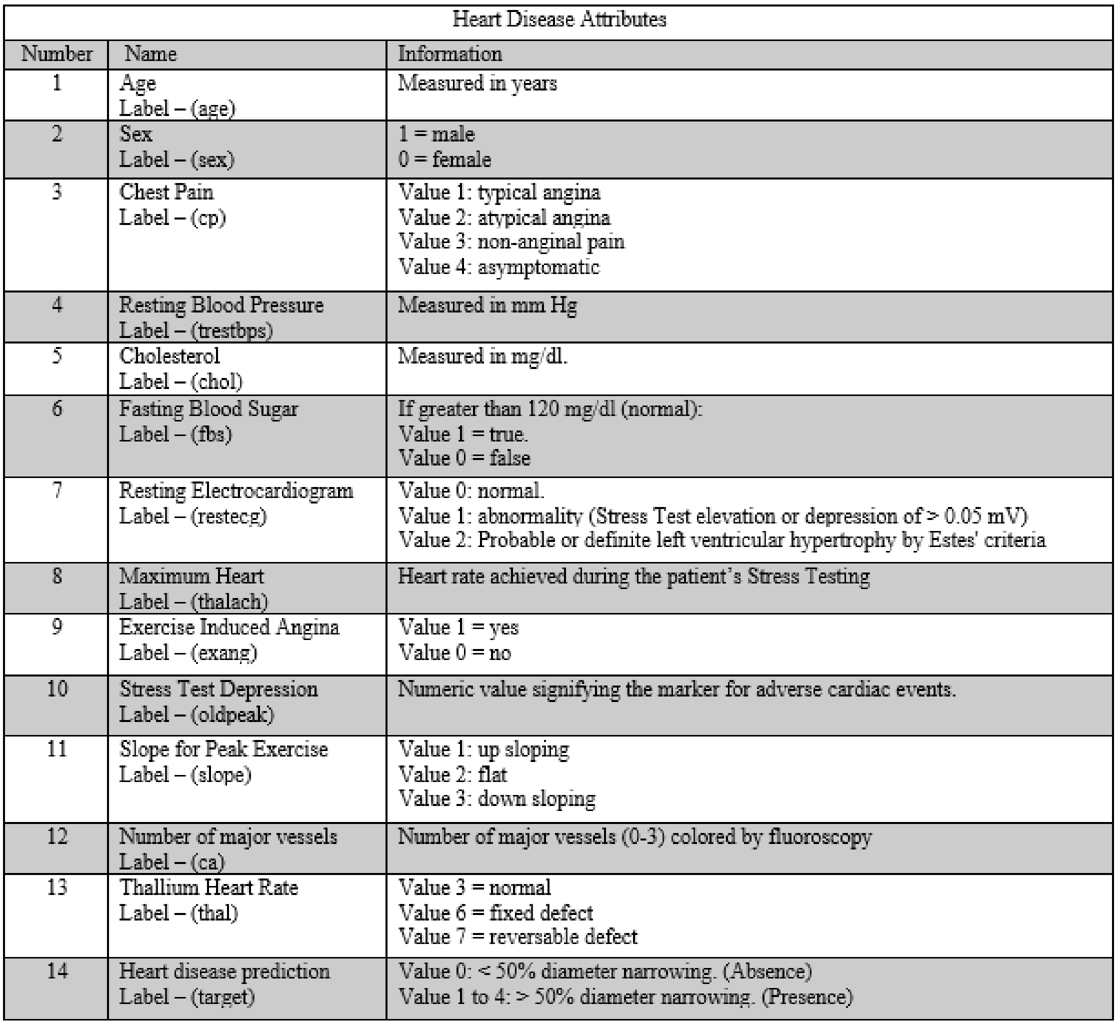
1.5 Conclusion

CAD is ranked as the most common among the other heart problems which the research will center on as the meaning of heart disease since it has a higher propensity to occur and is reflected in the data set. Because of the current paradigm of accessing solutions for the country's health has caused many to research the data for possible provision of a solution to mitigate contracting disease.Experts from the Mayo Clinic organization have further defined heart disease as: "generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease [11]." The datasets collected on the University of California Irvine has been shared to different data bases. However after reviewing the data, it suggests that the information was to geared toward a participant's blood flow as the determining factor for presence of heart disease symptoms.

Chapter 2 Data Exploration

2.1 Data Descriptions

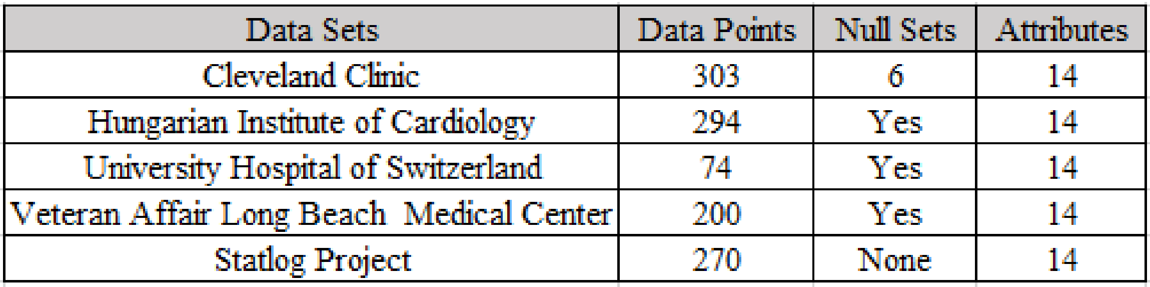
The initial dataset came from Kaggle, a popular repository for data, which referenced a data file on Heart disease. The referenced dataset is from University of California Irvine (UCI) Machine Learning Repository [8]. Each originally collected 76 variable attributes, but condensed to only 14 attributes due to the number missing data points from the 62 variables. See Table 2.1 for a list of the 14 attributes. This process is the beginning stages of data cleaning. For a number of unknown reasons to this researcher, the original raw source data became corrupted and was never re-uploaded to the repository.



Heart Disease Attributes		
Number	Name	Information
1	Age Label – (age)	Measured in years
2	Sex Label – (sex)	1 = male 0 = female
3	Chest Pain Label – (cp)	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4	Resting Blood Pressure Label – (trestbps)	Measured in mm Hg
5	Cholesterol Label – (chol)	Measured in mg/dl.
6	Fasting Blood Sugar Label – (fbs)	If greater than 120 mg/dl (normal): Value 1 = true. Value 0 = false
7	Resting Electrocardiogram Label – (restecg)	Value 0: normal. Value 1: abnormality (Stress Test elevation or depression of > 0.05 mV) Value 2: Probable or definite left ventricular hypertrophy by Estes' criteria
8	Maximum Heart Label – (thalach)	Heart rate achieved during the patient's Stress Testing
9	Exercise Induced Angina Label – (exang)	Value 1 = yes Value 0 = no
10	Stress Test Depression Label – (oldpeak)	Numeric value signifying the marker for adverse cardiac events.
11	Slope for Peak Exercise Label – (slope)	Value 1: up sloping Value 2: flat Value 3: down sloping
12	Number of major vessels Label – (ca)	Number of major vessels (0-3) colored by fluoroscopy
13	Thallium Heart Rate Label – (thal)	Value 3 = normal Value 6 = fixed defect Value 7 = reversable defect
14	Heart disease prediction Label – (target)	Value 0: < 50% diameter narrowing. (Absence) Value 1 to 4: > 50% diameter narrowing. (Presence)

Table 2.1: Condensed 14 Attributes Retrieved from UCI Repository

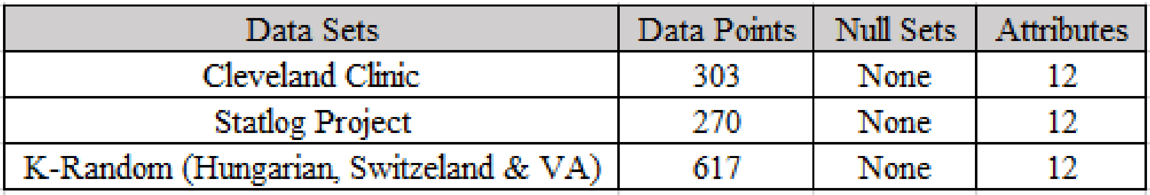
The collective data bases found in the repository were the Cleveland Clinic, Hungarian Institute of Cardiology, Switzerland University Hospital, and Veteran Affairs Long Beach Medical Center. Another heart disease data file with identical labeled attributes was found in the UCI repository entitled Statlog [10]. The Statlog dataset contained a completed 270 data points with no empty sets. This can all be seen in the table provided below.



Data Sets	Data Points	Null Sets	Attributes
Cleveland Clinic	303	6	14
Hungarian Institute of Cardiology	294	Yes	14
University Hospital of Switzerland	74	Yes	14
Veteran Affair Long Beach Medical Center	200	Yes	14
Statlog Project	270	None	14

Table 2.2: Data Files Information Retrieved from UCI Repository

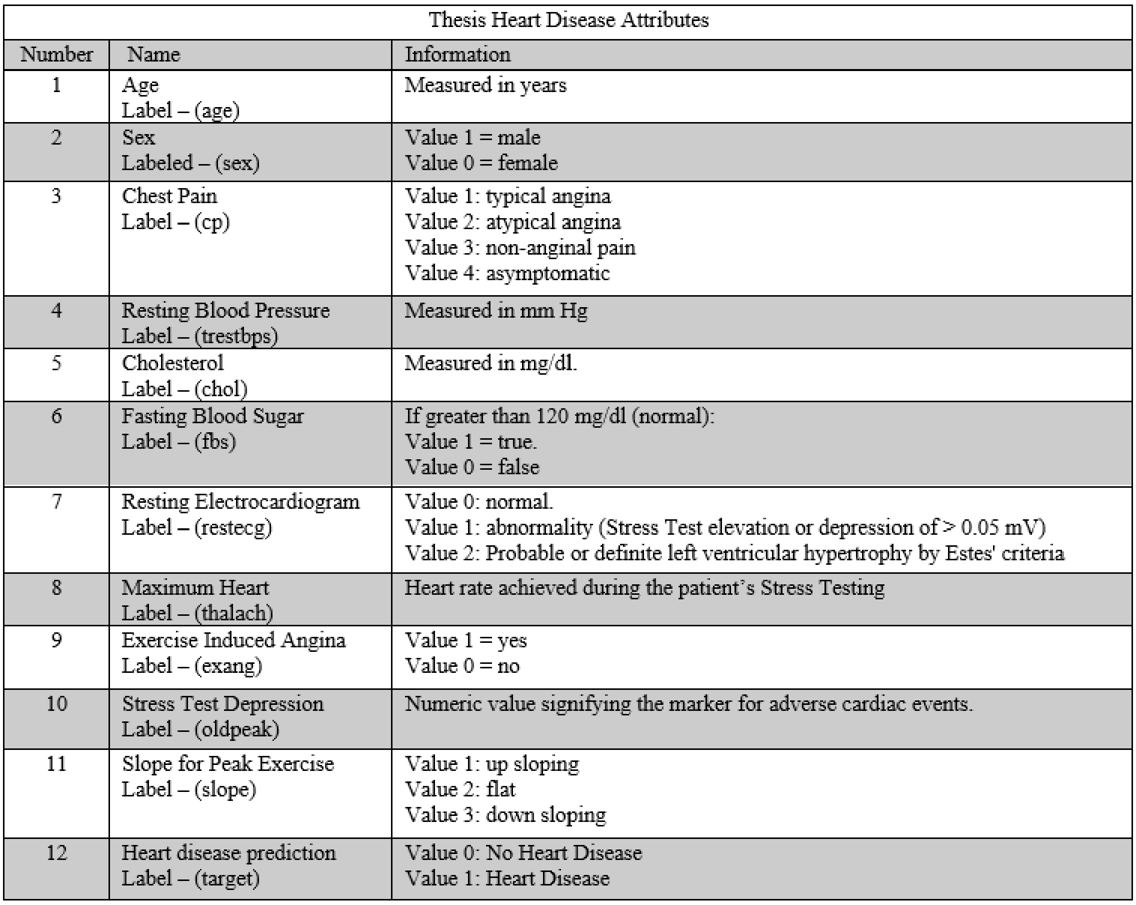
For Cleveland data set to be considered complete for the analysis for this thesis attributes "Number of Major Vessels" and "Thallium Heart Rate" were removed for the purpose to easing the data pre-processing. In an attempt to keep sizable amount of data points involved in the analysis, it was optimal to select the two most complete data sets which are Cleveland and Statlog respectively. In addition, several research were done with the decision to create a composite data set of all remaining incomplete data sets.



Data Sets	Data Points	Null Sets	Attributes
Cleveland Clinic	303	None	12
Statlog Project	270	None	12
K-Random (Hungarian, Switzeland & VA)	617	None	12

Table 2.3: Pre-Processed Data Sets

There are many factors that contribute to the prevalence of Heart Disease which have stemmed from age, inactivity, obesity, smoking, diabetes, family history, high blood pressure, excess levels Density Lipoproteins, and stress. For reasons that we explain later, we have relied on only 12 attributes as shown in Table 2.4.



Thesis Heart Disease Attributes

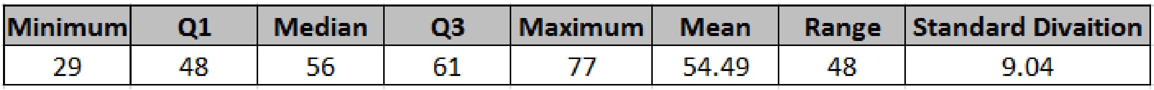
Number	Name	Information
1	Age Label – (age)	Measured in years
2	Sex Labeled – (sex)	Value 1 = male Value 0 = female
3	Chest Pain Label – (cp)	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4	Resting Blood Pressure Label – (trestbps)	Measured in mm Hg
5	Cholesterol Label – (chol)	Measured in mg/dl.
6	Fasting Blood Sugar Label – (fbs)	If greater than 120 mg/dl (normal): Value 1 = true. Value 0 = false
7	Resting Electrocardiogram Label – (restecg)	Value 0: normal. Value 1: abnormality (Stress Test elevation or depression of > 0.05 mV) Value 2: Probable or definite left ventricular hypertrophy by Estes' criteria
8	Maximum Heart Label – (thalach)	Heart rate achieved during the patient's Stress Testing
9	Exercise Induced Angina Label – (exang)	Value 1 = yes Value 0 = no
10	Stress Test Depression Label – (oldpeak)	Numeric value signifying the marker for adverse cardiac events.
11	Slope for Peak Exercise Label – (slope)	Value 1: up sloping Value 2: flat Value 3: down sloping
12	Heart disease prediction Label – (target)	Value 0: No Heart Disease Value 1: Heart Disease

Table 2.4: 12 Attributes Being Addressed in the Thesis

2.2 Attributes Descriptions

2.2.1 Age

Because heart health risk increases with age, the screening age range for participants is greater than of equal to 35. Although heart risk is not typical for younger age groups, Heart Foundation(HF) doesn't detour younger participants especially family's that have a history with heart health risks [7].



Minimum	Q1	Median	Q3	Maximum	Mean	Range	Standard Divaition
29	48	56	61	77	54.49	48	9.04

Table 2.5: Kaggle Age Summary

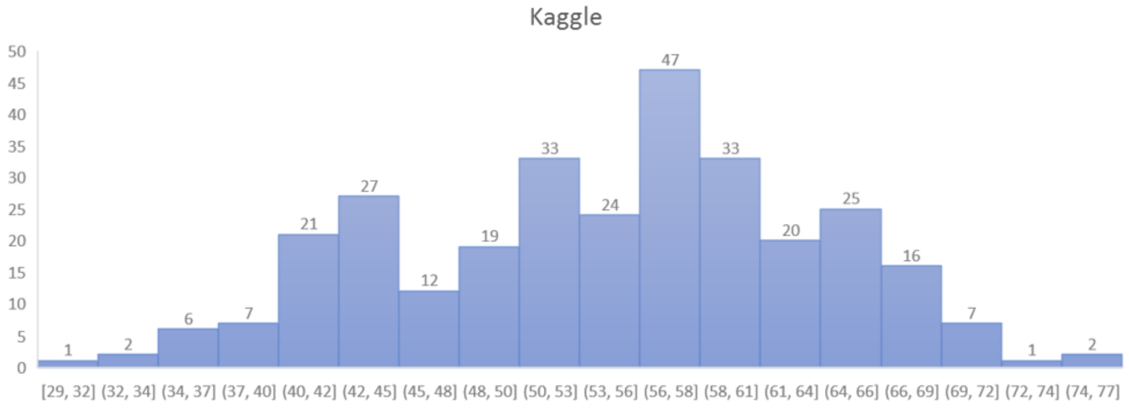


Figure 2.1: Bin Width Kaggle Age Summary

2.2.2 Sex

The standing statistic for heart disease its the primary cause of death for both men and women in the U.S.[12]. The nominal values 0 and 1 are assigned to female and male, respectively, to facilitate concise binary evaluation. Assigning these values is necessary for certain types of machine learning which will be used in the evaluation later after more critical attributes have been considered.

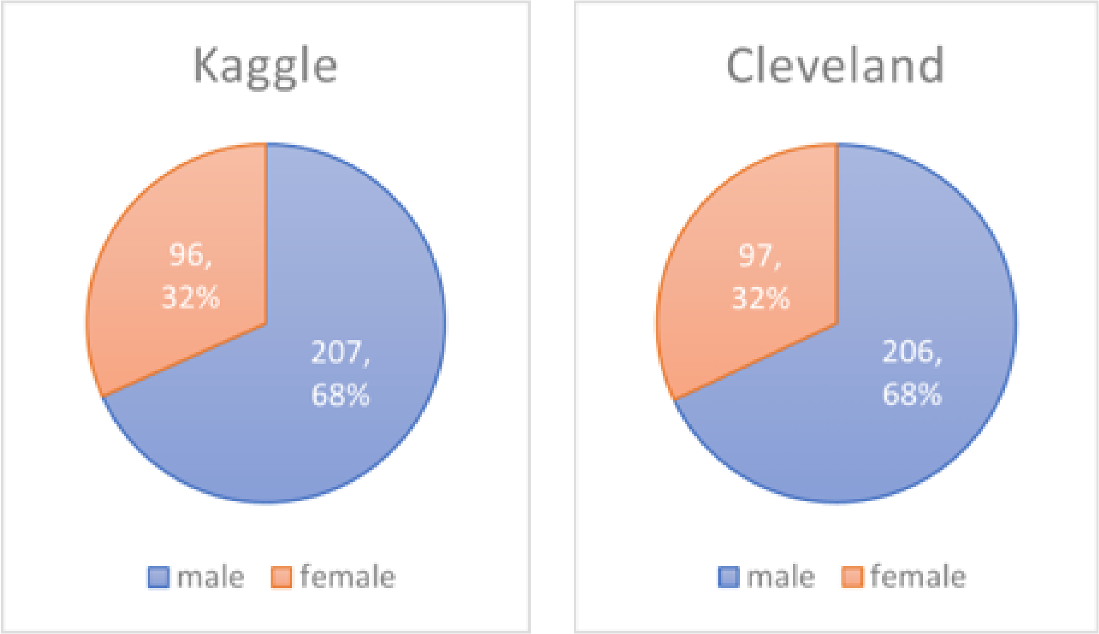
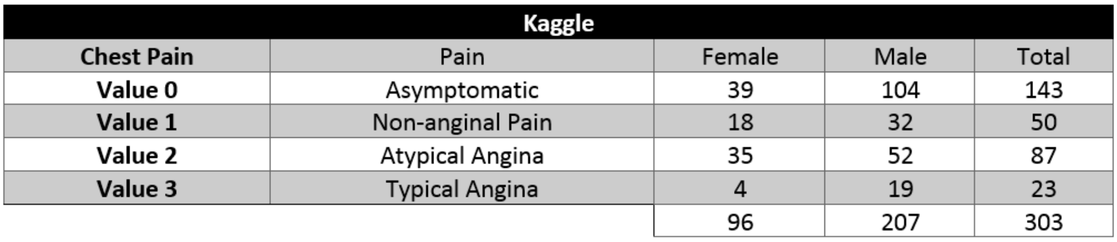


Figure 2.2: Similarities Represented between Kaggle and Cleveland

2.2.3 Chest Pain

The most important indicator of heart disease is the chest discomfort or chest angina [12]. Because discomfort provides the first indication of beginning symptoms for many disease, chest angina establishes there is a problem present. The data tables below angina levels may be case by case due to pain tolerance [12]. Although this is one of the most dangerous symptoms of many diseases, this attribute can act as a control for initial diagnoses verse actual diagnoses.

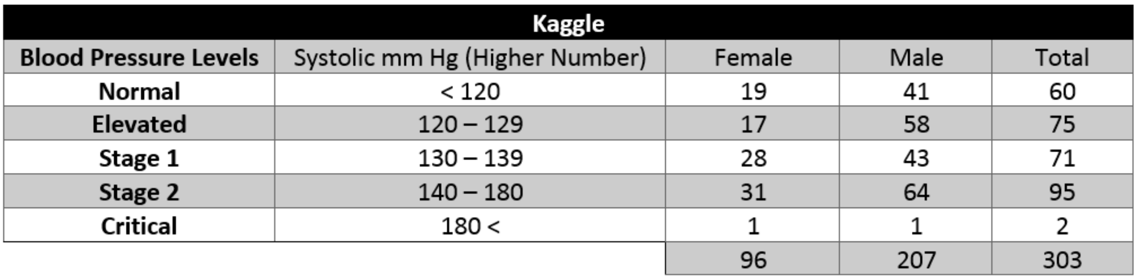


Kaggle				
Chest Pain	Pain	Female	Male	Total
Value 0	Asymptomatic	39	104	143
Value 1	Non-anginal Pain	18	32	50
Value 2	Atypical Angina	35	52	87
Value 3	Typical Angina	4	19	23
		96	207	303

Table 2.6: Range of Chest Angina

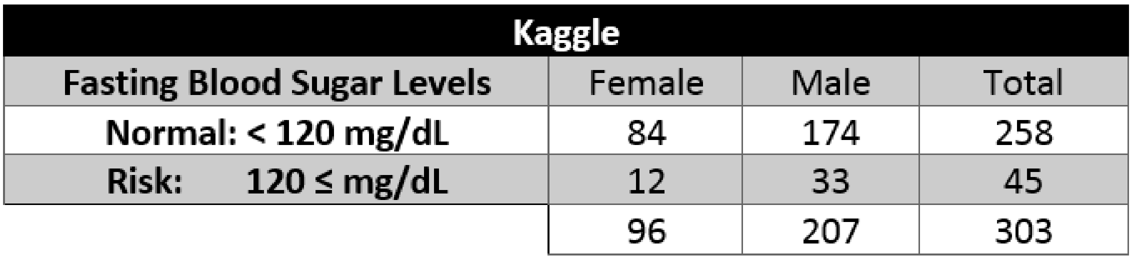
2.2.4 Blood Pressure

The blood pressure contributes the to systematic structure for indicating the hearts proper functioning. As with other fields of study the,pressure measures the stress of the blood by the contraction of the surrounding dimensions. As blood flows through blood vessels, the Blood pressure rises and falls naturally throughout the day [9]. By analyzing this innate movement, research data collected may provide an more insight to medical professionals about the patient's heart condition. When the pressure remains too high for a extensive time period, the repercussions may results in high risk for CAD, heart attack stroke, and a series of other indicators for heart failure [12]. From the CDC has recorded that High Blood Pressure also known as Hypertension affects about 1 in 4 adults (24%) with hypertension have their condition under control in the US [3]. Which infers American adults may not even be aware of they have it because the symptoms aren't prevalent as diseases [20]. Resting blood pressure in millimeters of mercury (mm Hg) when the patient was admitted to the hospital.



Kaggle				
Blood Pressure Levels	Systolic mm Hg (Higher Number)	Female	Male	Total
Normal	< 120	19	41	60
Elevated	120 – 129	17	58	75
Stage 1	130 – 139	28	43	71
Stage 2	140 – 180	31	64	95
Critical	180 <	1	1	2
		96	207	303

Table 2.7: Resting blood pressure gender comparison

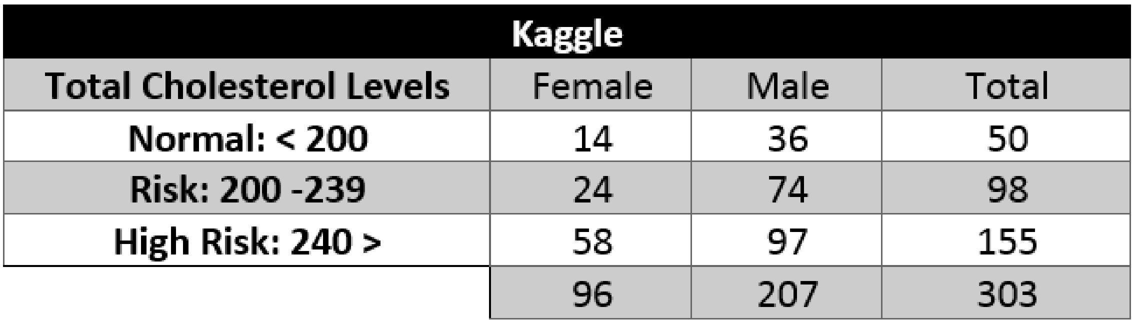


Kaggle			
Fasting Blood Sugar Levels	Female	Male	Total
Normal: < 120 mg/dL	84	174	258
Risk: 120 ≤ mg/dL	12	33	45
	96	207	303

Figure 2.3: Resting Blood Pressure Comparison

2.2.5 Cholesterol

Cholesterol is a fat-like substance called a lipid that's found naturally in the blood. Lipids is vital for the normal functioning of the body. The human body manufactures all the cholesterol it needs from diet. Cholesterol can be measured with a simple blood test [6].

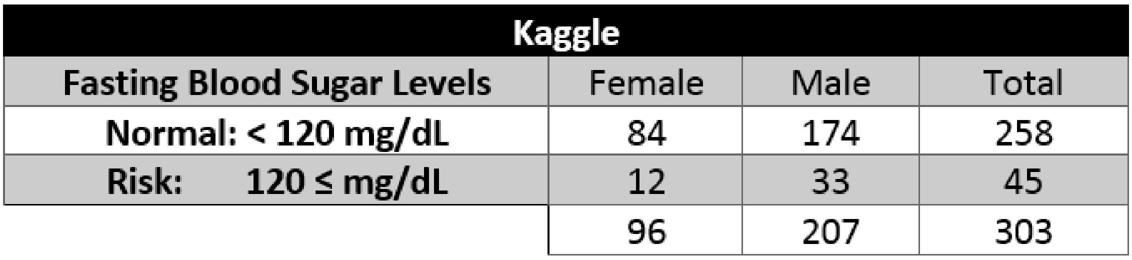


Kaggle			
Total Cholesterol Levels	Female	Male	Total
Normal: < 200	14	36	50
Risk: 200 -239	24	74	98
High Risk: 240 >	58	97	155
	96	207	303

Table 2.8: Total cholesterol levels for gender comparison

2.2.6 Fasting Blood Sugar

Blood Glucose or commonly recognized as Blood Sugar Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high [15]. Milligrams per deciliter, is a measurement that indicates the amount of glucose in a specific amount of blood. According to the CDC, about 1 in 4 people with diabetes don't know they have the disease. For the data sets, blood sugar is distinguished whether the patient's blood sugar is higher than 120 mg/dl or not.

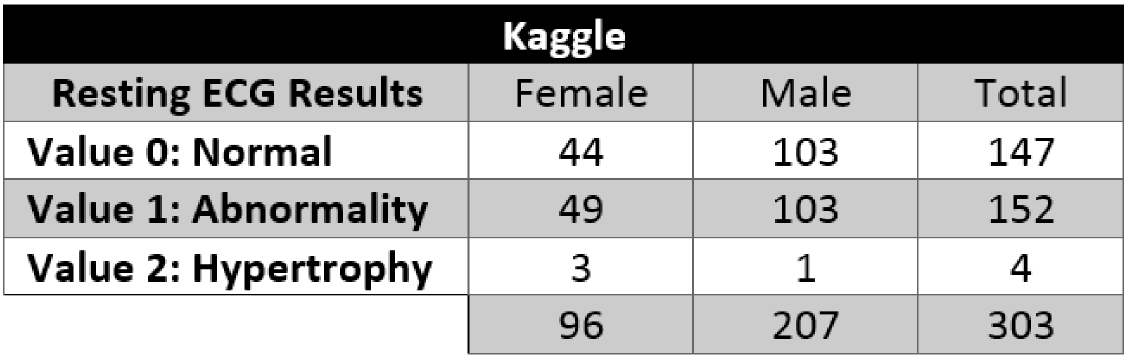


Kaggle			
Fasting Blood Sugar Levels	Female	Male	Total
Normal: < 120 mg/dL	84	174	258
Risk: 120 ≤ mg/dL	12	33	45
	96	207	303

Table 2.9: Fasting blood sugar levels for gender comparison

2.2.7 Electrocardiogram

The electrocardiogram results are accepted as the current standard for evaluation of patients. Results cater to the patients body during the exercise. A patient within stable angina occurring through exercise which happens even on rest the disease got worse. This has to be why there are so few patients that show an abnormality on the heart rate on rest, and it is also why seeing this abnormality is very indicative of a presence of a heart disease. On the other hand, the value 0, probable presence of a hypertrophy, doesn't seem to be very indicative of the presence of a heart disease by itself.



Kaggle			
Resting ECG Results	Female	Male	Total
Value 0: Normal	44	103	147
Value 1: Abnormality	49	103	152
Value 2: Hypertrophy	3	1	4
	96	207	303

Table 2.10: Kaggle resting EKG results

"Thallium Stress Test" also known as "Nuclear Stress Test" or "Cardiac Test" benefits heart disease research by analyzing the condition of blood flow [14]. The gamma camera through nuclear imaging tracks the participant's blood flow which carries a sample amount of Thallium, radioactive isotope [14].

2.2.8 Heart Rate

The maximum heart rate was recorded during Thallium stress test. The data set showcase the optimal maximum healthy heart rate depends on the age (220 - age). Thus, higher rates tend to be from younger patients.

2.2.9 Exercise Pain

This attribute is the patient's level of angina or induced pain during exercise which is necessary input for the presence of heart disease.

2.2.10 Old Peak

The resting Stress Test segment depression is the marker for adverse cardiac events. The old peak monitors a certain level in a normal heart beat which indicates a displacement for the presence of a heart disease.

2.2.11 Slope

The part of the Stress Test for indicators of exercise. The slope by itself can help determine whether there is heart disease or not if it is flat or ascending. Adding a third variable where we can see if the slope is descending, the depression of the ST segment can help to determine if the patient has a heart disease.

2.2.12 Target

The target is designated as the condition of the patient for heart disease after the conducting stress testing indicators. The results of the testing had an extensive range that determined for the disease presence. For the sake of simplicity and design of approach, the patients which had any indication of disease present then in this analysis it was considered diagnosed as heart disease.

Chapter 3 Heart Disease Data Investigation

3.1 Kaggle

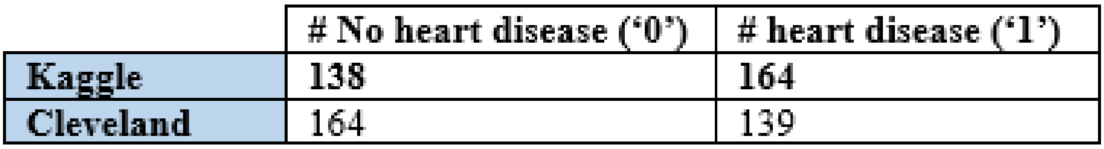
The heat.csv data found in Kaggle supposedly originated from the University of California, Irvine, Machine Learning Repository [http://archive.ics.uci.edu/ml]. This database contains 76 attributes, but all published results rely on using a subset of 10-14 of them. The archive contains four datasets for coronary artery heart disease: namely, Cleveland Clinic Foundation Heart disease (303 records and available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease). Hungarian heart disease, Long-Beach-V.A. heart disease and the Switzerland data from the University Hospital, Zurich. We noticed that the same database has been uploaded on other popular data sites such as data world (https://data.world) A closer look at the archives shows that all except the Cleveland data is in a 'workable' condition, and any other data has either too many missing (often shown as 'NA' or '?' or 'blank') or inconsistent values (duplicated rows, or '0' where inappropriate); apparently the original datasets have been corrupted. This leads to our first question. Which of the four original UCI ML Repository datasets has been used in Kaggle? Looking simply at the dataset sizes, the Cleveland data appears to be the most obvious one since both data have the same size. We hypothesize that the Kaggle data is likely the Cleveland data and opt to verify their features records, distributions and perhaps authenticate the Kaggle data records.

3.2 Cleveland vs Kaggle data

The original Cleveland data is called processed.cleveland.data, which can be downloaded from the UCI Machine Learning repository site (in “Data Folder”) and exported as a txt file. It was donated in July 1988 and currently has 1,471,444 web hits. This data has no header, and yet the headers information were provided on the website (under Attribute Information)—There are 303 records and 14 features (5 numerical, 9 categorical) information were provided which helped determine each column header in the order listed in the Table B, below. Although the Cleveland data features share similar distributions to those of Kaggle (see Appendix I), there are a couple of differences that we list below:

(a) Contrary to the Kaggle data, the original Cleveland data has 6 missing values (recorded as ‘?’): 4 missing values are found for the number of blood vessels (ca) feature and 2 missing values for the thalassemia (thal) feature. Because each feature is categorical, it appears that, each of their missing values was replaced in the Kaggle data with a closest value generated by the Nearest Neighbor Algorithm. However, since these two features do not appear in a larger dataset (ORIG) which we explore later, for practical comparative analysis, decided to drop them. Granted that this is a minor difference given the data size.

(b) The target variable in the Kaggle data is binary, with ‘0’ for ‘Absence’, and ‘1’ for ‘Presence’ of a disease. However, the target variable in the Cleveland is nominal and has five levels—0 for Absence of disease and ‘1’, ‘2’, ‘3’, ‘4’ for different levels of heart disease’s conditions. Once again, given that a larger dataset (ORIG) which we explore later has only a binary target, for practical comparative analysis purpose, we decided to recode the target as ‘0’ for ‘Absence’ and ‘1’ (for ‘1’, ‘2’, ‘3’, ‘4’) as ‘Presence’. Note that, ‘0’ was already declared as ‘Absence’ of disease in the original list of attributes’ information about the Cleveland data (see Table B). When we compare the number of records associated with each class of the target variable in both data, we found a strong contrast between the records as shown in Table 3.1



	# No heart disease (‘0’)	# heart disease (‘1’)
Kaggle	**138**	**164**
Cleveland	164	139

Figure 3.1: Recorded target values for Kaggle vs Cleveland datasets

The profiles of these two data are almost identical when comparing all other features except Target, the predictor variable.—we also note few other minor differences. For instance, there are 206 Males, 97 Females in the Cleveland data while there are 207 Males, 96 Females in the Kaggle data. Any reasonable data analysis of these two datasets will likely result into divergent conclusions. Yet, it seems that, there are no articles or studies that have brought these conflicting observations. One reason we suspect is that, most of the machine learning data analysis presented focus on the results outputs but little on the meaning or interpretation of the results. Because of the “reverse “coding of the target variable in these two files, we have every reason to further investigate and clear any doubt regarding these two medical records. Before we begin a comparative analysis of some of the features of the two data, we present a second most original heart disease data, found in the UCI Machine Learning repository.

3.3 Statlog Data

The original Statlog data is called heart.dat and it can be accessed in the UCI Machine Learning repository site under the “Data Folder”. It is also available in the UCI repository (at:[10] `https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29` although in a different archive or data folder. This file can also be downloaded/exported as a txt file and it consists of 270 completed records and 14 attributes including a target variable. Its source and donation date were recorded as ‘N/A’, i.e., as unknown. Further, the site stated that, it is a heart dataset that is similarly to a dataset already present in the repository (Heart Disease databases) but in a slightly different form. The data’s columns have no heading yet, its attributes are the same as those of the Cleveland data which are shown in Table B.

We also note that the target variable of Statlog data has values '1' (for 'Absence' of disease) and '2' (for 'Presence' of disease) instead. We re-encoded these values as '0' (for Absence) and '1' (for Presence) to match those of the Cleveland and Kaggle data.

Further, to each of these two datasets, we added these headers and the predicted attribute was (re)named "target". Given its size, we concluded that Statlog is not Cleveland data. Therefore it could be one of the (3) corrupt datasets. We use this data to help verify any unusual or abnormal distribution that may occur as we compare Cleveland and Kaggle datasets to determine which has the correct 'target' coding.

As stated earlier and shown in [2] age, high blood pressure (trestpbs) and chest pains (exang) are three of the leading key risk factors which are often associated with patients with heart disease. To determine the validity of the target variable encoding, we plot each of the corresponding variable against the target variable. Naturally, we expect Cleveland data and the Kaggle data to be telling opposite stories. Meanwhile, to help support these risk factors, we relied on the Statlog data to validate or confirm the expected trend. For instance, we expect a patient with higher blood pressure to be more likely classified as 'Disease' patient or '1'. Since both Cleveland and Kaggle data will likely classify such patient in opposite classes (i.e., one of the datasets may assign the patient to Class '1', while the other dataset may assign the patient to Class '2'), Statlog's trend/classification of such patient will help break any tie.

3.4 Comparative Analysis of 3 leading risk features in three datasets

Age is one of the most important risk factors in developing cardiovascular or heart diseases. According to the NIH[https://www.nia.nih.gov/health/heart-health-and-aging], people age 65 and older are more likely than younger people to suffer a heart attack, a stroke, or to develop coronary heart disease (aka called heart disease) and heart failure. As you get older, our heart can no longer beat fast enough during physical activity and one of the major concerns is the building up of fatty deposits in the arteries over the years.—such build-ups are often felt as pain or **angina pain**. Further, as we age, the arteries stiffen, raising our **blood pressure** or hypertension which is now common among younger individuals. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55. So, we examine each of these risk features.

3.4.1 Age Factor

From the Kaggle data, we found that the older you get the less likely you will have a heart disease as shown in the plot. Certainly, the opposite is expected.

The Cleveland data and the Statlog are certainly in agreement that, at earlier age, we see more individuals being classified as ‘No heart disease’ even though at later age, we note some older individuals fall within the same category. See Figure 3.2 for details, and the Appendix for larger images.

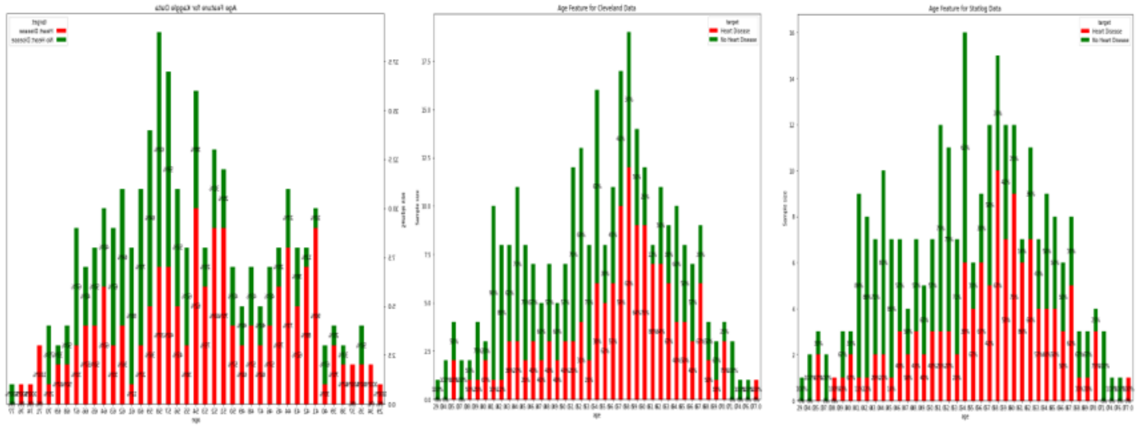


Figure 3.2: Age Feature for Kaggle, Cleveland and Statlog datasets (from Left to Right)

3.4.2 Exercise Induced Pain Factor

We note here that the exercise induced angina (exang) correlates with chest pain (cp) variable, although the later has 4 levels and was not induced. Because different individuals who experience the same pain level may likely be classified as different levels of pains since pain level is quite subjective. For this reason, we choose exang over cp to compare against the target variable.

The Kaggle data shows that 70% of individuals who show no sign of angina during exercise were classified as 'Heart disease' vs 23% of individuals who show some sign of angina. This would mean that, according to the Kaggle data, those who experience angina pain during exercise are less likely to be classified as 'heart disease' contrary to those who do. Naturally, the opposite is true as shown in the feature distribution plots (Figure 3.3) of the Cleveland data and the Statlog data.

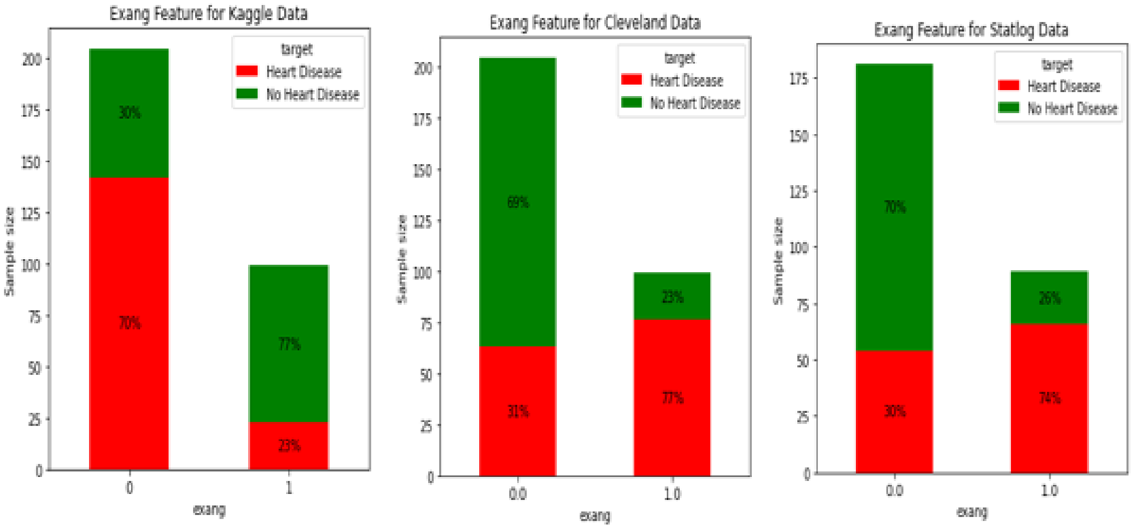


Figure 3.3: Exercise Induced Pain Feature Plot for three datasets

3.4.3 Resting Blood Pressure Factor

With Kaggle data, we find that individuals with lowest resting blood pressure are more likely to be classified as having heart disease than those who have higher resting blood pressure. See Figure 3.4. Once again, the opposite is true as shown in the feature distribution plots of the Cleveland and Statlog datasets.

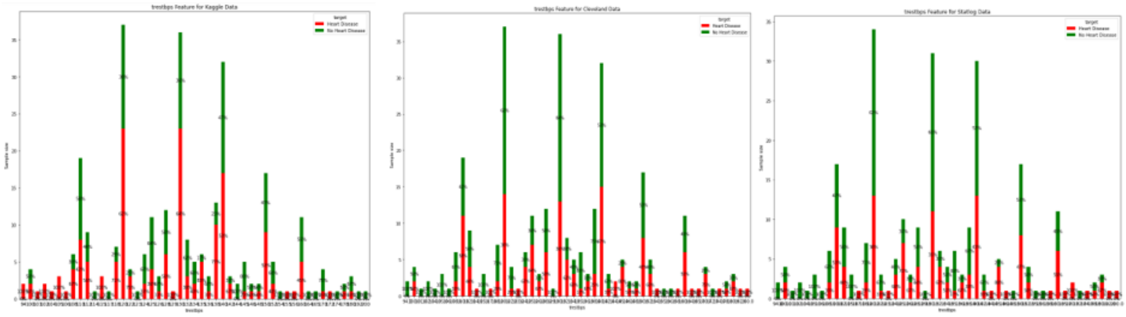


Figure 3.4: Exercise Induced Pain Feature Plot for three datasets

3.4.4 Conclusion

Given the pairwise graphical comparison of the records of the three features from both Cleveland and Kaggle data, we conclude that the target variable coding for the Kaggle data should be reversed; or else, it must be interpreted as '0' for 'Heart Disease' and '1' for 'No Heart Disease'. With this recoding or reinterpretation, the Kaggle data is essentially the Cleveland data (with 6 newly imputed values for the missing data). For the purpose of this research, we retain the original Cleveland data.

3.5 Features distributions of Statlog and Cleveland datasets

Here, we decide to compare the Statlog data to Cleveland data, as both come from the UCI Machine Learning archive.

3.5.1 Numerical features density plots

We observe similar density plots across all numerical features for both datasets as shown in their respective distribution plots. (See Figure 3.5 and Figure 3.6)

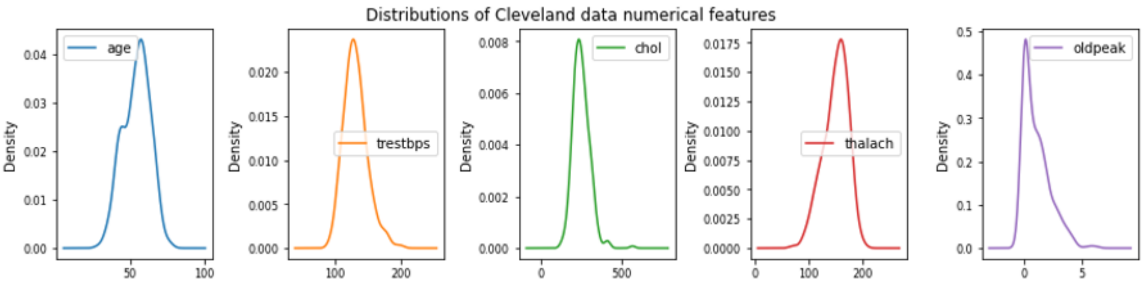


Figure 3.5: Distribution of Cleveland numerical features

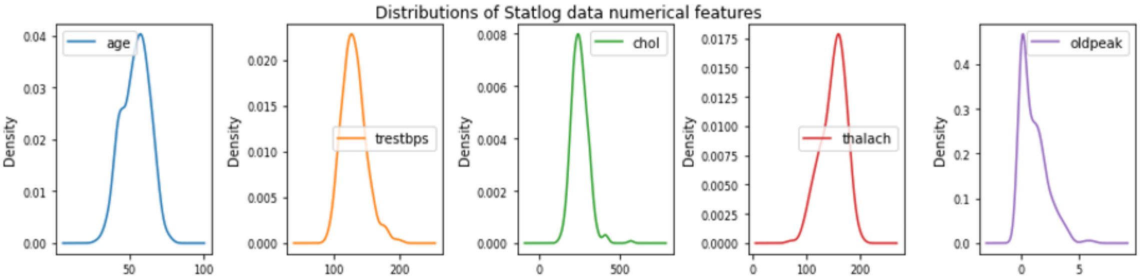
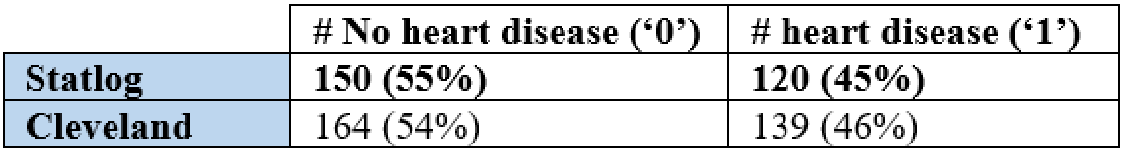


Figure 3.6: Distribution of Statlog numerical features

3.5.2 Categorical Features Distributions

3.5.3 Target

As we compare the number of records in each class of the target variable, we find that both data share approximately the same proportion of patients classified as 'Heart disease' (or 'No Heart disease') as shown in Table 3.7.



	# No heart disease ('0')	**# heart disease ('1')**
Statlog	**150 (55%)**	**120 (45%)**
Cleveland	164 (54%)	139 (46%)

Figure 3.7: Target variable distributions across both datasets

3.5.4 Gender

Both datasets (as shown in Figure 3.8) appear to have the same gender distributions: 68% (or 206) Males vs 32% (or 97) Females for the Cleveland data, while

Statlog data registers 67.8% (or 183) Males vs 32.2% (or 87) Females.

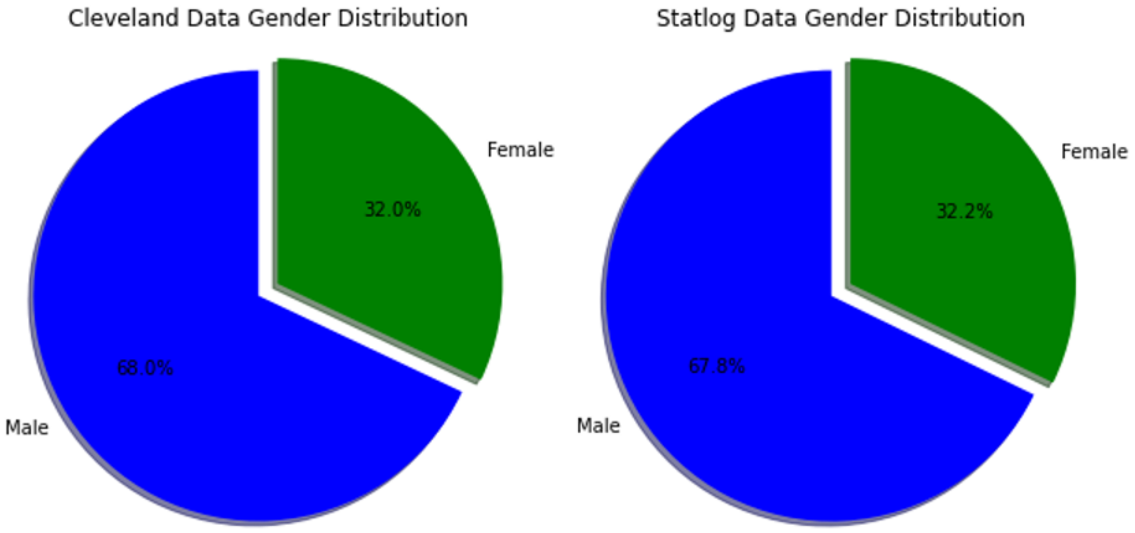


Figure 3.8: Cleveland and Statlog datasets Gender distribution

Further, when compared against the target variable, both datasets show very similar distributions: More percentage of women (between 74%-77%) are more likely to be classified as 'No Heart Disease' compared to men (45%).

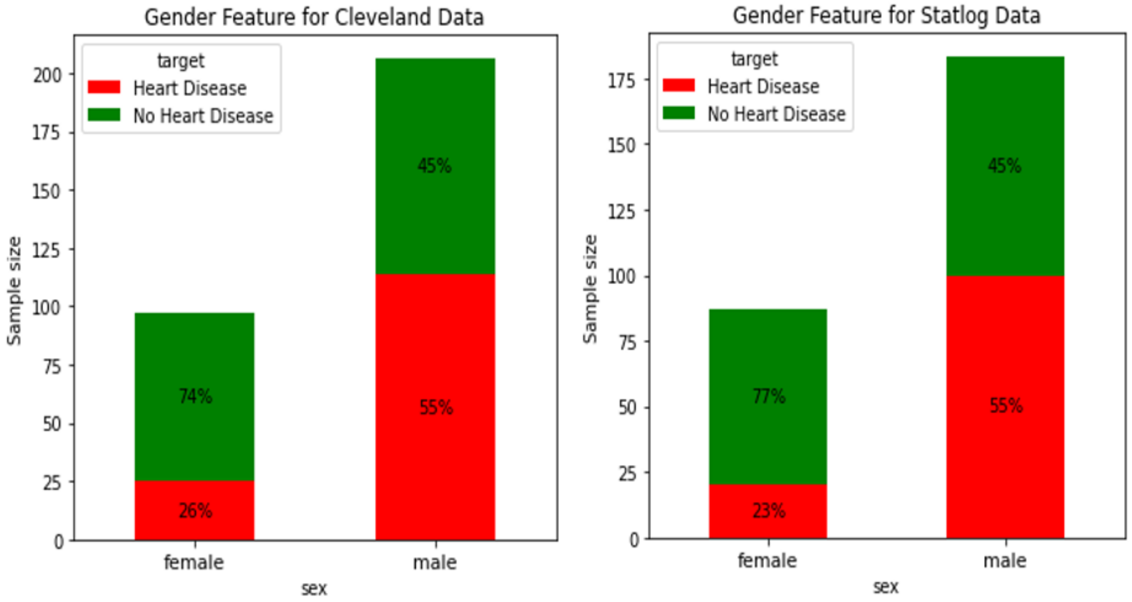


Figure 3.9: Gender Feature distribution against the Target

3.5.5 Chest Pain

According to the American Heart Association [`https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain/angina-pectoris-stable-angina`] chest pain (aka angina pectoris) is a common condition that affects several million people in the United States. Yet most people are not aware of its different symptoms. It is the discomfort that is noted when the heart does not get enough blood or oxygen. This is often due to a blockage or plaque buildup in the coronary arteries. In which case, a partially or completely blockage will prevent the heart from getting enough oxygen.

As we analyze both data (see plot) we find that, a the majority of individuals (between130-140 records) have expressed a high level (level '4') of chest pain. Of this group, a clear majority of them (between 71%-73%) are classified as having a 'Heart disease'.

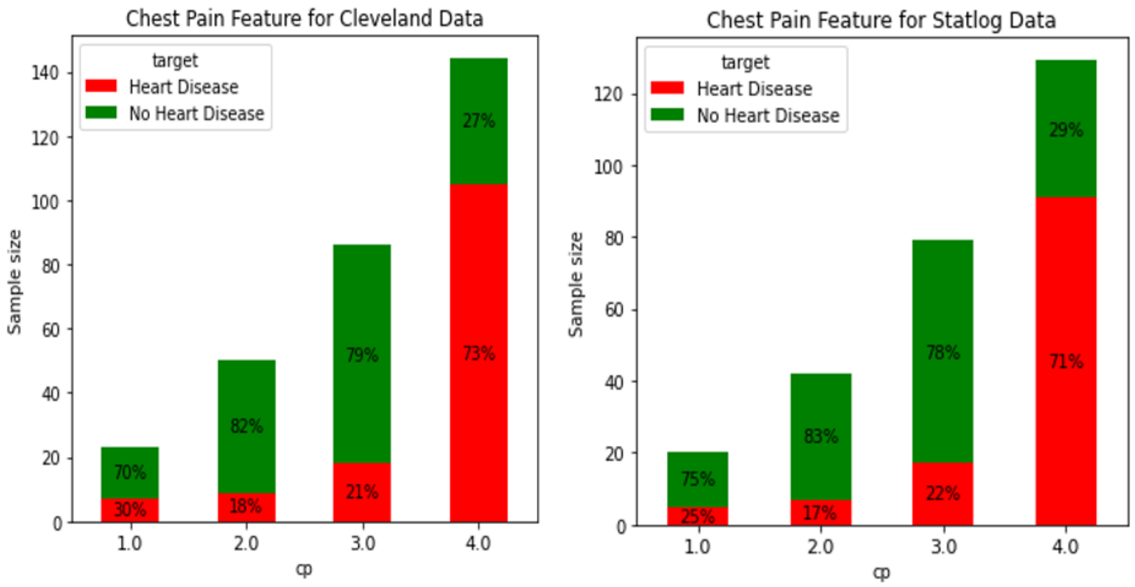


Figure 3.10: Chest Pain Feature distribution against the Target

3.5.6 Fasting Blood Sugar

For most people, 80 to 99 milligrams of sugar per deciliter before a meal and 80 to 140 mg/dl after a meal is normal. The American Diabetes Association [`https://www.diabetes.org/a1c/diagnosis`] recommends that most nonpregnant adults with diabetes should have 80 to 130 mg/dl before a meal and less than 180 mg/dl at 1 to 2 hours after beginning the meal.

From Figure 3.11, we observe that the majority of the patients do not show any sign of high blood sugar (or diabetes)—In fact, a slightly higher majority of individuals who have a high blood sugar (>120mg/L) are actually classified as "No heart disease' in the Statlog data. There is no difference in the percentage of classification for those do not have a high blood sugar. It appears that, overall, diabetes or the sugar level alone, is not a decisive factor in the classification of heart disease patients.

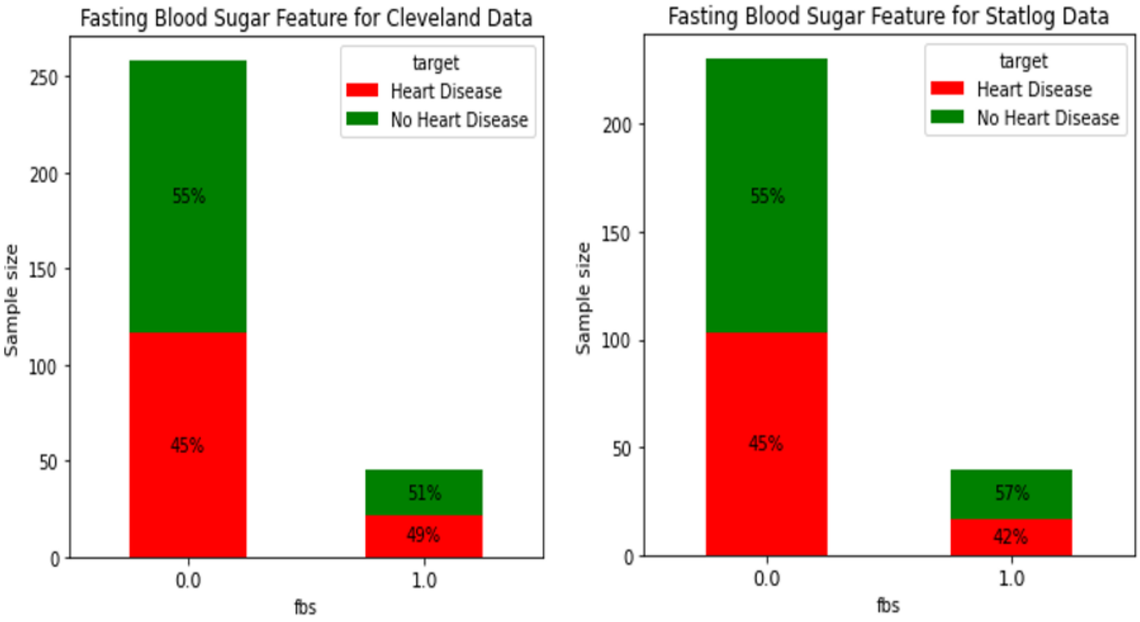


Figure 3.11: Fasting Blood Sugar Feature distribution against the Target

3.5.7 Resting Electrocardiogram Feature

An electrocardiogram (ECG or EKG) is a simple test that measures an individual heart's electrical activity. Typically, each heartbeat is triggered by an electrical signal that starts at the top of your heart and travels to the bottom and when a heart is showing signs of disease, it affects its electrical activity.

Both data show individuals who have normal EKG result are more likely to be classified as 'No Heart Disease' than those who show some abnormality. Further, the plots also show that, having an abnormal EKG result does not increase significantly the odds of being classified as 'Heart Disease', especially in the Statlog data. Still, the odds are slightly bit higher (10%) for the Cleveland data.

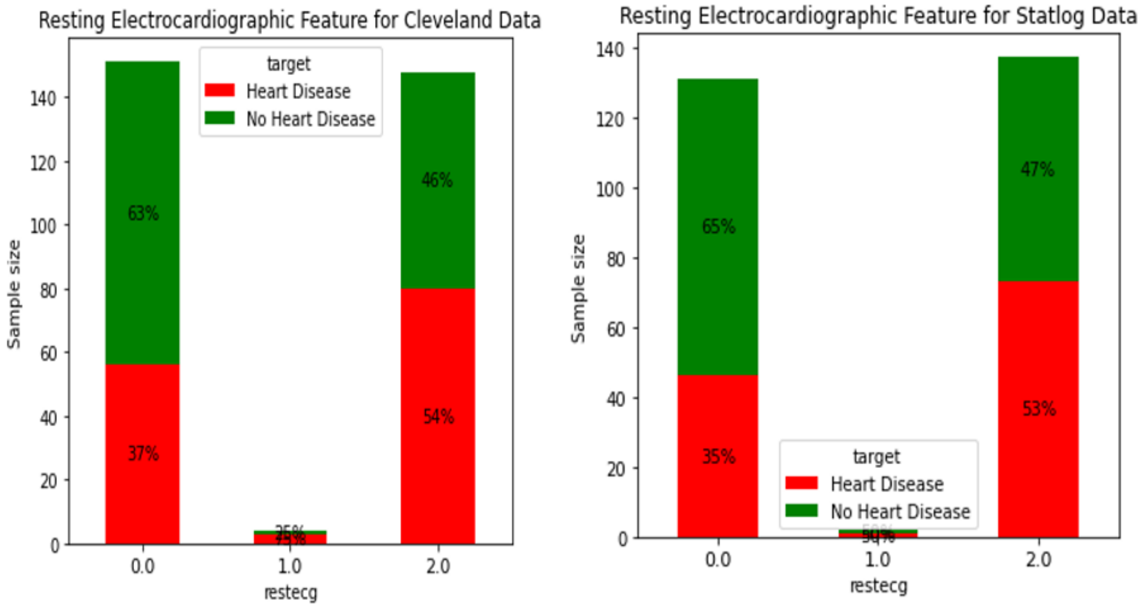


Figure 3.12: Resting ECG Feature distribution against the Target

3.5.8 Exercise Angina

Angina or chest pain or discomfort may occur during activities such as climbing stairs, or becoming upset or even going outside into the cold air according to the American Heart Association. In Section 1.1, we compared this feature's distribution for both datasets against the one found in Kaggle. See Figure 3.3. There, we found that, not only both datasets share the same distribution but about 70% of individuals who show no sign of exercise angina are classified as 'No Heart Disease'.

3.5.9 Slope

An ST segment or slope is the flat section of the EKG between the end of an S wave (the J point) and the beginning of the T wave. The ST Segment represents the interval between ventricular depolarization and repolarization.The most important cause of ST segment abnormality (elevation or depression) is myocardial ischaemia or infarction.

Once again, we see a similar trend between the classification of patients in both datasets, as shown in Figure 3.13. It is evident that, those who have ‘low’ slope are classified as ‘No Heart Disease’ at a rate of 3 to 1 compared to those who have a mid to high slope.

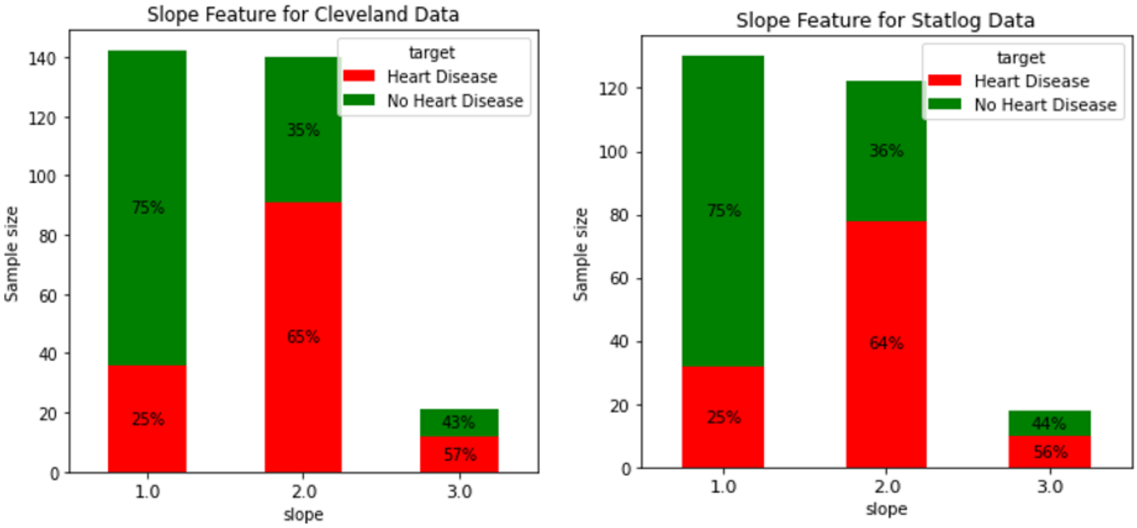


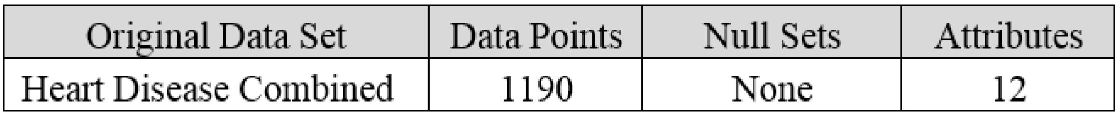
Figure 3.13: Slope Feature distribution against the Target

3.5.10 Conclusion

Given the pairwise graphical comparison of the records of all the (12) features from both Cleveland and Statlog data, we noted the unusually close similarities between these data. Upon further inspections (through merging and removal of duplicates), we found that the Statlog data is actually a subset of the Cleveland; in which case, only the Cleveland datasets (303=270+33) contains exactly 33 records that cannot be found in the Statlog data. We decide to drop the Statlog data and keep the Cleveland data for further analysis. Thus, any datasets that includes both of these data is riddled with duplicates and should **not** be recommended for analysis.

3.6 Potential Missing datasets

Recently [Nov 2020], Siddhartha has created and uploaded a dataset in IEEE data site aka dataport which can be found at `https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive`. For the purpose of this research, we refer to his data as ORIG data. This data is declared to be a combination of the following (5) datasets: Cleveland (303 obs.), Hungarian (294 obs.), Switzerland (123 obs.), Long Beach VA (200 obs.) and Statlog Heart Data (270 obs.); a total of 1190 observations and 12 attributes as shown in Table 2.2 and summarized by Table 3.1.



Original Data Set	Data Points	Null Sets	Attributes
Heart Disease Combined	1190	None	12

Table 3.1: ORIG dataset

The attributes are the same as those listed in Table 2.4 i.e., those in Table 2.3 minus two columns: number of major vessels (ca) and patients heart rate (thal). So, the ORIG data contains the previously mentioned (2) datasets, namely Cleveland and Statlog; this explains the unusually high level of duplicates ($\sim$ 40%). We proceed to identify these two datasets in the ORIG data. Once identified, we remove those records from the ORIG data and the remaining dataset is called the MISS (for “missing”) dataset. The MISS data has 613 observations, 12 features, with 0 duplicate.

Yet, we have no way of recovering the missing three datasets in the original UCI dataset repository. We have reached out to Siddhartha to find out about the original subsets of data but we have not received a response from him.

3.6.1 MISS dataset Features Exploration

We found that most the numerical features’ distributions appear normal except ‘chol’ and ‘oldpeak’. A closer look at the cholesterol variable shows that there are 172 (about 28% of the data) records with ‘zero’ values for the cholesterol in the data. Likewise, ‘oldpeak’ shows 267 (about 46% of the data) records with ‘zero’ values for the ST depression. Still, it is important to point out that the Cleveland data has no ‘zero’ cholesterol record but does have 99 (about 33% of the data) ‘zeros’ records for ‘oldpeak’ records.

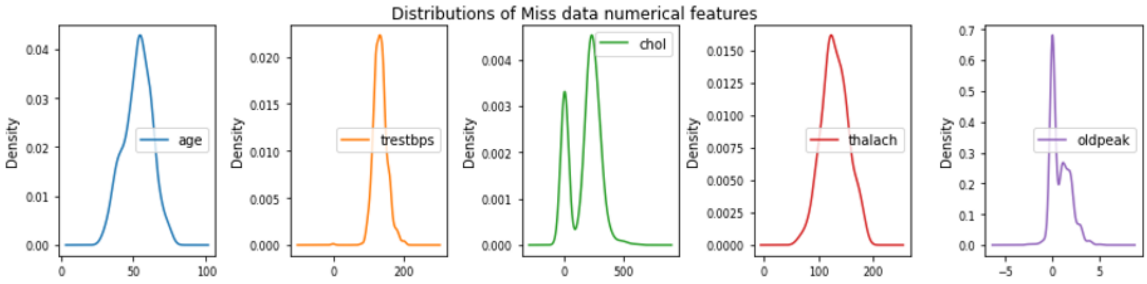


Figure 3.14: Distribution of MISS data numerical features

Moreover, we noticed that the data has an unusually high male proportion (85% male vs 15% female) and any analysis that includes ‘sex’ will potential be biased due to a lack of a reasonable female representation.

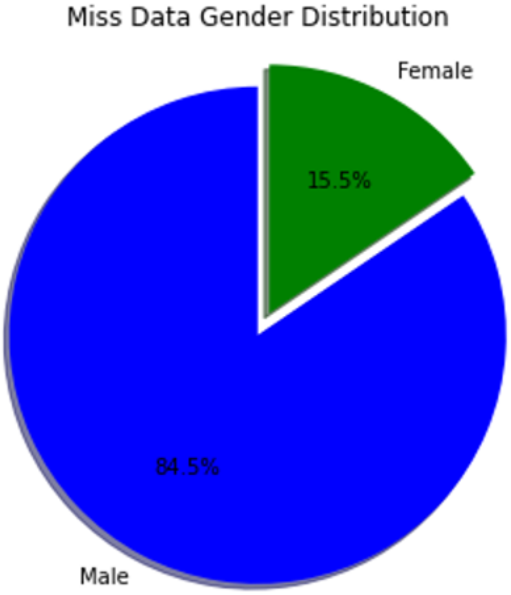


Figure 3.15: Distribution of Gender features

3.6.2 Conclusion

Given some of the earlier anomalies (highly skewed gender distribution, and unusual values for ’chol’) found in MISS data and our inabilities to authenticate the ORIG data, we decided to drop the MISS data for our predictive model. For the rest of this thesis, we rely only on the Cleveland data for our analysis.

Chapter 4 Predictive Models

Predictive analytics generally seek to extract information from the raw data in order to predict trends or indicate certain patterns of behavior. Here we rely on standard statistical data modeling such as logistic regression and a well-known machine learning technique called Random Forest. Fundamentally, we are trying to capture the relationships between Heart Disease and features such as Age, Sex, Cholesterol,etc...Some of the features are more important than others so we rely on Random Forest features' selections to select the best classifiers. We begin by introducing the reader to some common statistics, models, and technical terms.

4.1 Basic Statistics and Machine Learning

4.1.1 Level of significance

Also known as **alpha level**, this value is used as a probability cutoff for making decisions about the null hypothesis. Its value represents the probability we are willing to place on our test for making an incorrect decision in regards to rejecting the null hypothesis. In other words, it is the level of risk we are willing to take as we reject a possibly correct hypothesis. For example, a significance level of 0.05 indicates a 5% risk of concluding there is a statistically significant result or difference when there is none.

4.1.2 P-value and Confidence Interval

P-values are the probability of obtaining an effect or a relationship at least as extreme as the one in the sample data, as we assume the truth of the null hypothesis.

When a p-value is less than or equal to the significance level (typically 0.05), we reject the null hypothesis.

The range of values, for which the p-value exceeds a specified alpha level is called **confidence interval**. In other words, this interval gives a range of values within which lies a true (population) parameter. So, with an estimated parameter at $\alpha = 0.05$, a confidence interval indicates that, with repeated samplings (identical studies in all respects except for random error), we are "confident" that, in spite of margin-of-error (or deviations), 95% of the parameter estimates will lie within this interval. With the margin-of-error we can state that the interval includes the true population parameter.

4.1.3 Correlation

A simple correlation measures the relationship between two (ideally normally distributed) variables. For our thesis we used Pearson's r which measures a linear relationship (or association) between two continuous (numeric) variables without taking into account other variables. For each pair of variables (X_i, X_j) Pearson's correlation coefficient is computed using

$$r = \frac{\sum\limits_{i=1}^{n}(x - \overline{x_i})(y - \overline{y_i})}{\sqrt{\sum\limits_{i=1}^{n}(x - \overline{x_i})^2 \sum\limits_{i=1}^{n}(y - \overline{y_i})^2}}.$$

Its value range between -1 and 1 and $|r| \sim 1$ indicates a strong dependence or correlation and $|r| \sim 0$ indicates a strong independence between the variables.

The objective of any data analysis is to extract information (or accurate estimation) from the original (raw) data. Typically, we seek to determine whether or not there is statistical relationship between a response variable (Y) and explanatory variables (X_i). One way to answer this question is to use some regression analysis in order to *model* its relationship. By modeling we try to predict the outcome (Y) based on values of a set of predictor variables (X_i). There are several types of regression analysis and each type of the regression model depends on the type of the distribution of Y. They are often used to assess the impact of multiple variables (a.k.a. covariates and factors) in the same model. Here, we focus on two of these which we define next.

4.1.4 Linear regression

This is an extension of the simple correlation. In regression, one or more variables X_i (*predictors* or *factors* or *independent variables* or *inputs*) are used to predict an outcome Y_i (*response* or *target* or *criterion* or *dependent variable* or *output*). In practice, a linear regression model or equation returns estimates of the coefficients of a linear equation that involves one or more independent variables that best predict the values of an ouput or the dependent variable which must be quantitative continuous or scale. It is often written as

$$E(Y_i) = \beta_0 + \beta X_i \quad \text{or} \quad Y_i = \beta_0 + \beta X_i + \epsilon_i$$

for each i observation or data point with errors ϵ_i.

Regression coefficients or coefficient estimates β_i represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

The *p-value* for each term tests the null hypothesis that the coefficient is equal to zero (no effect). Thus, a low p-value (< 0.05) indicates that we can reject the null hypothesis, in which case the corresponding predictor is likely to be a meaningful addition (or is *statistically significant*) to your model. Likewise, a larger (insignificant) p-value suggests that changes in the predictor are not associated with (or do not help explain) changes in the response. Thus, for our analysis, we use the coefficient p-values to determine which variables are useful for our final model.

As it is true for any model, part of the process involves checking to make sure that the data we want to analyze can actually be done using the chosen model. For a linear model it is required that, for each value of the independent variable, the distribution of the dependent variable must be normal. Typically, we plot the errors (residuals) to see if they follow a normal distribution. A QQ- plot is an example of such a residual plot that can be used to reveal biased results more effectively than a simple computation. Further, the variance of the distribution of the dependent variable should be constant for all values of the independent variable. Finally, the relationship between the dependent variable and the independent variables should be linear, and all observations should be independent. In brief, the residuals of a good model should be normally and randomly distributed.

In the event the response variable takes a form where the residuals look completely different from a normal distribution, it is preferable to consider another class of models known as *generalized linear models (GLM)*; in which case the response variable Y_i follows an exponential family distribution. Logistic regression is an example of a GLM as we define it, next.

4.1.5 Binomial Logistic regression

Binomial Logistic regression which is simply called a *logistic regression* estimates the probability of an occurrence of an event Y_i based on a set of predictors X_i. The basic mathematical concept behind logistic regression is *logit* which is the natural logarithm (ln) of an odds; and odds are ratios of probability "success" p (for instance, an ambulance was needed) to probability "failure" $1 - p$ (when no ambulance was needed, for instance). Thus, given a response categorical variable Y and m predictors X_i, we have

$$logit(Y) = log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i \tag{4.1}$$

where β_0 is the Y intercept (i.e., mean of Y independent of X_i's) and β_i's are the *regression coefficients* (or *parameter estimates*) for each predictor X_i, for $i = 1, \ldots, m$.

We note that, by taking exponential (or anti-log) of both sides of equation 4.1, we derive the equation to predict the probability of the occurrence of an outcome of interest as follows:

$$\begin{aligned} p &= Probability\ (Y = outcome\ of\ interest \mid X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_m = x_m) \\ &= \frac{e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}}, \end{aligned}$$

where $e \sim 2.71828$ is the natural base.

Interpretation :

(i) The sign $(\pm)$ of a coefficient (or slope) β_j gives the direction of the relationship (negative or positive) between the predictor X_j's and the logit of Y.

(ii) The intercept or log average odd $\beta_0 = log(\frac{p}{1-p})$ is an estimate of the model (null model) if we consider no predictor; this is also known as *unconditional log odds* of the response. Thus, the **average odd** is e^{β_0} and the **average probability of success**, p is $\frac{e^{\beta_0}}{1+e^{\beta_0}}$.

(iii) The coefficient β_j, for some predictor X_j. Fixing the levels of the remaining predictors X_k, $k \neq j$, this value gives the log(odds) of the effect of X_j on Y (beyond the average) for each unit increase (in a scale variable) or in comparison to a fixed (base) level in X_j. Thus, for a predictor X_j, the **estimated odds** value is e^{β_j} and the **percentage change** in odds (per unit increase or relative to a base level) is

$$(e^{\beta_j} - 1) \times 100\%.$$

As related to inferential statistics, a *null hypothesis* would state that, for some $\beta_j = 0$, $j > 0$, i.e., there is no linear relationship between logit of Y and X_j, in the population. So, rejecting such a null hypothesis would imply that a linear relationship exists between logit of Y and X_j. As indicated earlier for linear regression, we will rely on the p-values and the alpha level of .05, to help make our decision on the significance of the coefficients.

4.1.6 Multinomial Logistic regression

Multinomial logistic regression (or *multinomial regression*) is used to predict a nominal dependent variable (with two or more factors or categories) given one or more independent variables. As such, it is an extension of binomial logistic regression to allow for a dependent variable with more than two categories.

4.1.7 R-squared

Also known as **coefficient of determination**. it is a statistical measure of how close the data are to the fitted regression line. In other words, it is the percentage of the response variable variation that is explained by a linear model in which case

$$R^2 = \frac{Explained\ variation}{Total\ variation} \times 100$$

0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data but there are risks of "overfitting" or bias, which makes the model less adaptable to a different data taken under a similar circumstance.

4.1.8 Pseudo R-squared

As opposed to an R-squared value that is obtain from evaluating a model built on a continuous response, such an indicator does not make sense for models built on an ordinal response where the variance is fixed instead. However, a similar metric (in scale) called a “Pseudo” R-squared is used for models such as logistic regressions. In which case, the higher the value the better model but they are only meaningful when comparing these values for distinct models. There are several such pseudo R-squared values but SPSS software returns the values for Nagelkerke, and Cox & Snell (Pseudo) R-squareds.

4.1.9 Confusion Matrix

In the area of *machine learning* when it comes to statistical classification we often rely on a confusion matrix (or *error matrix*) which gives the performance of a classifier or supervised learning algorithm; neural network, which we define later, is an example of a classifier. The **confusion table** or **confusion matrix** is a 2 matrix with the number of **true positives** (TP; hit) and **true negatives** (TN; correct rejection) on row 1 and the number of **false positives** (FP; false alarm or Type I error) and **false negatives** (FN; miss or Type II error) on row 2, respectively by columns. The performance of a classifier will be measured with the following statistics:

4.1.10 Accuracy

It is a parameter that is designed to determined whether or not a test accurately measures what it is supposed to measure. In which case, it is the ration of correctly classified patients or subjects in the entire record. Thus, **Accuracy** is given by

$$\frac{TP+TN}{TP+FP+TN+FN}.$$

4.1.11 Recall or Sensitivity

It is the measure of the proportion of actual positives (TP) that are correctly identified (e.g., the percentage of injured bikers who are correctly identified as being injured). Thus, **sensitivity** or true positive rate (TPR) is given by

$$TPR = \frac{TP}{TP+FN}.$$

4.1.12 Specificity

It is the measure of the proportion of actual negatives that are correctly identified (e.g., the percentage of patients with no heart disease that are correctly identified as 'No Heart Disease'). Thus, **specificity** or true negative rate (TNR) is given by

$$TNR = 1 - FPR = \frac{FN}{FP + TN}.$$

When one is measuring the proportion of actual positives that are correctly identified, it is called **Precision** . In which case, the formula becomes

$$\frac{TP}{TP + FP}.$$

4.1.13 Receiver Operating Characteristic (ROC)

This is a plot of the diagnostic ability of the classifier system as we vary its discrimination threshold (or cut-points). Thus, a curve is obtained as we plot the true positive rate (TPR) against the false positive rate (FPR) at various cut points. In general, the closer the curve is to the top left corner in the plane, the better the classification as shown in Figure 4.1.

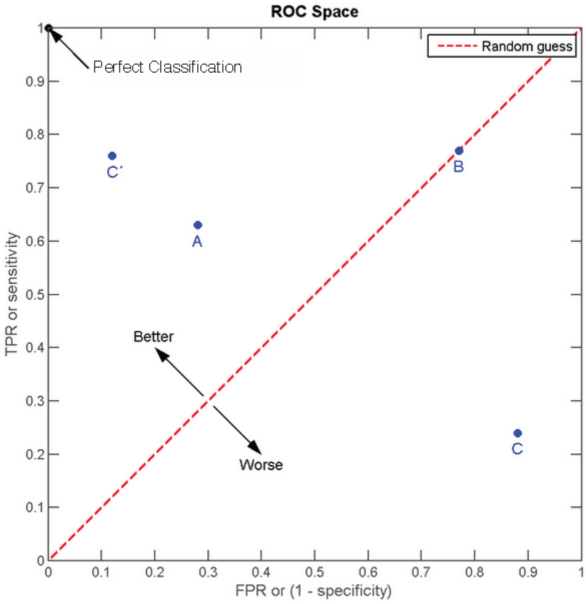


Figure 4.1: An ROC curve space

In order to check the performance of our classifier, we will rely on the **AUC (Area Under Curve)** of the ROC curve; this is a measure of discrimination or diagnostics. As such, a higher AUC, the better the model at distinguishing between say, injured bikers vs non-injured bikers, following an accident. Thus, an excellent classifier has AUC ~ 1 while a poor classifier has AUC ~ 0.

4.1.14 Akaike's Information Criterion

Akaike information criterion (AIC) (Akaike, 1974) is a a criteria that is used to estimate the likelihood of a model to predict/estimate future values based on a "large enough" (usually $n > 40$) sample size. A *relatively* good model is the one that with *minimum* AIC among all the other models. It is computed with the formula

$$AIC = -2log(L) + 2k,$$

where k is the number of model parameters (the number of variables in the model plus the intercept) and L is the value of the likelihood; the Log-likelihood is a measure of model fit–The higher the number, the better the fit. This is automatically generated from a statistically R output.

4.1.15 Neural Network

This is a sophisticated classifier that is applied to a data when the nature of the relationship between the predictors and the response is not clear; this relationship is learned through repetitive "training" methods. For example, *gradient methods* such as *gradient descent* (on a loss function) are used to train multilayer networks by updating weights to minimize loss.

4.1.16 Random Forest

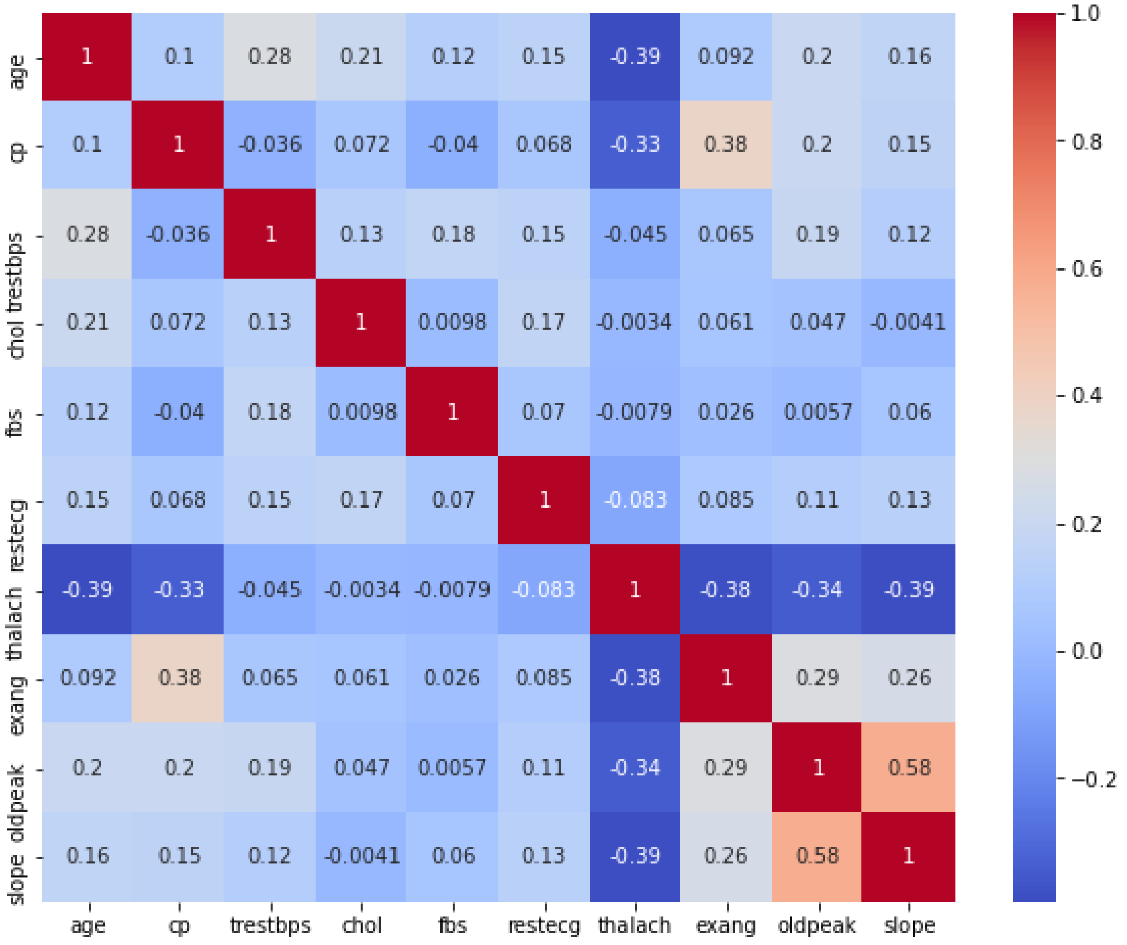
This is a collection of decision trees also called a forest that classifies for each tree a new object based on the variables. From there, each object classification receives a vote where selection is based on the highest votes received. In other words, this algorithm takes a subset of the data points along with a subset of variables and constructs a decision tree. The tool amalgamates the build of the decision trees in order to make a more accurate and stable prediction. From the input, the out of predictions performs the model. The model for this algorithm is

$$P(c|f) = \sum_{1}^{n} P_n(c|f).$$

4.2 Analysis Results

The regression analysis are done using R computing language. By default, the category that R chooses to be the *reference or baseline* is the first category listed **alphabetically** or numerically (if coded 0, 1, ...). For our regression output for instance, it is Female ('0') and 'Heart Disease' that are use used as baselines. The machine learning, Random Forest, is run in Python (Jupyter notebook).

A quick numerical variables' correlation check is first performed. By default, Python computes the Pearson's correlation coefficient. As shown in Figure 4.2, each row and column represents a continuous variable, and each value in this matrix is the correlation coefficient (Pearson's r value) between the variables represented by the corresponding row and column. We found that most attributes are not strongly correlated. Slope is mildly correlated with oldpeak with a correlation coefficient of $r = 0.58$.

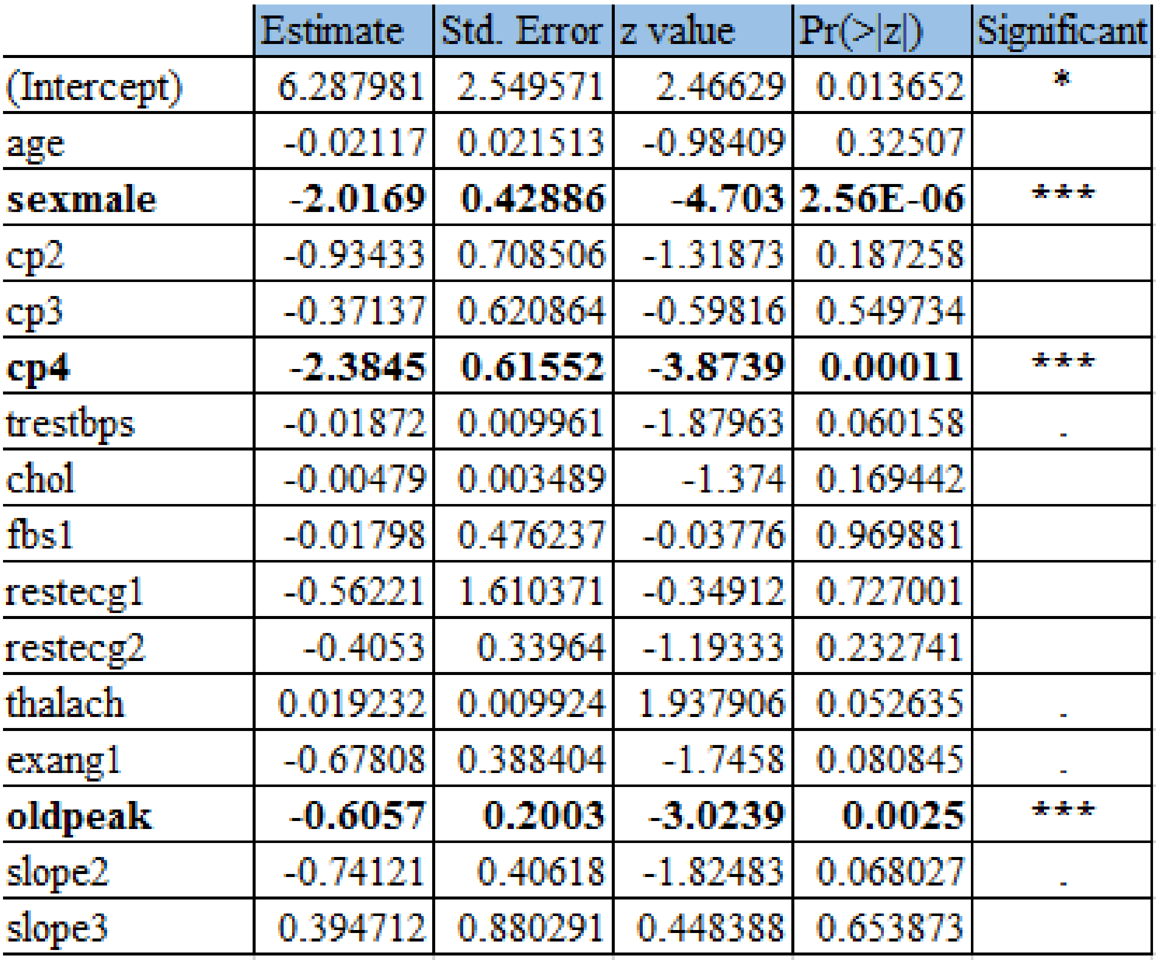


	age	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
age	1	0.1	0.28	0.21	0.12	0.15	-0.39	0.092	0.2	0.16
cp	0.1	1	-0.036	0.072	-0.04	0.068	-0.33	0.38	0.2	0.15
trestbps	0.28	-0.036	1	0.13	0.18	0.15	-0.045	0.065	0.19	0.12
chol	0.21	0.072	0.13	1	0.0098	0.17	-0.0034	0.061	0.047	-0.0041
fbs	0.12	-0.04	0.18	0.0098	1	0.07	-0.0079	0.026	0.0057	0.06
restecg	0.15	0.068	0.15	0.17	0.07	1	-0.083	0.085	0.11	0.13
thalach	-0.39	-0.33	-0.045	-0.0034	-0.0079	-0.083	1	-0.38	-0.34	-0.39
exang	0.092	0.38	0.065	0.061	0.026	0.085	-0.38	1	0.29	0.26
oldpeak	0.2	0.2	0.19	0.047	0.0057	0.11	-0.34	0.29	1	0.58
slope	0.16	0.15	0.12	-0.0041	0.06	0.13	-0.39	0.26	0.58	1

Figure 4.2: Correlations Heat Map of Continuous Variables in Cleveland

4.2.1 Regression Output of Full Model of Cleveland Data

After ensuring that each variable has the proper coding and structure, we run a logistic regression on all the (12) features against the target variable. The regression output is shown in Table 4.1. The predictors which are significant with a p-value $\leq .05$ are highlighted. The logistic regression coefficients give the change in the log odds of getting 'Heart Disease'.



	Estimate	Std. Error	z value	Pr(>|z|)	Significant
(Intercept)	6.287981	2.549571	2.46629	0.013652	*
age	-0.02117	0.021513	-0.98409	0.32507	
sexmale	**-2.0169**	**0.42886**	**-4.703**	**2.56E-06**	***
cp2	-0.93433	0.708506	-1.31873	0.187258	
cp3	-0.37137	0.620864	-0.59816	0.549734	
cp4	**-2.3845**	**0.61552**	**-3.8739**	**0.00011**	***
trestbps	-0.01872	0.009961	-1.87963	0.060158	.
chol	-0.00479	0.003489	-1.374	0.169442	
fbs1	-0.01798	0.476237	-0.03776	0.969881	
restecg1	-0.56221	1.610371	-0.34912	0.727001	
restecg2	-0.4053	0.33964	-1.19333	0.232741	
thalach	0.019232	0.009924	1.937906	0.052635	.
exang1	-0.67808	0.388404	-1.7458	0.080845	.
oldpeak	**-0.6057**	**0.2003**	**-3.0239**	**0.0025**	***
slope2	-0.74121	0.40618	-1.82483	0.068027	.
slope3	0.394712	0.880291	0.448388	0.653873	

Table 4.1: Cleveland Data Regression Output

There are three variables namely sex, cp (level 4), and oldpeak, that are significant at a 5% level. There are three other variables that very close to be statistically significant at a 5% level. They are trestbps, thalach and exang. The error estimates for each of these values is low and the AIC value of the model is 269 which is relatively low. The deviance residuals are close to 0 and roughly symmetrical.

Further, we have decided to test the classification/decision making of a machine, given a logistic model prediction on all the features. We train on 75% of the data and test on 25% of the data, to get the output in Figure 4.3.

4.2.2 Model Classification Testing

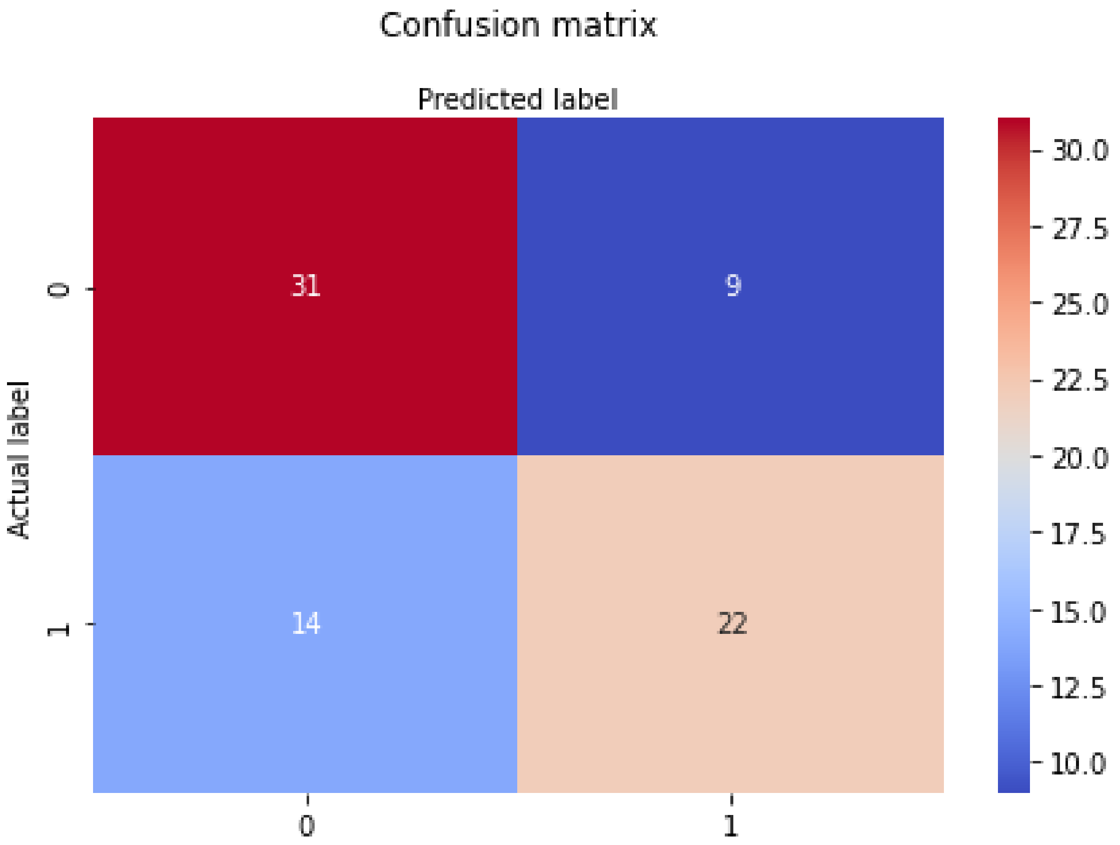


Figure 4.3: Confusion Matrix of Cleveland Data

For simplicity we summarize the output in the next table:

$n = 76$	Predicted: No Heart Disease	Predicted: Heart Disease	
Actual: No Heart Disease	$TN = 31$	$FP = 9$	40
Actual: Heart Disease	$FN = 14$	$TP = 22$	36
	45	31	

Table 4.2: Confusion Matrix of Cleveland Data Summarized

From the confusion matrix table (Figure 4.2), we obtained:

(a) Accuracy: This determines how often is the binary classifier correct–In which case we have $ACC = \frac{TP+TN}{Total} = \frac{31+22}{76} \sim 0.70$, i.e., 70% of the times. Consequently, its *misclassification rate* also known as *error rate* is 1-ACC which is about 30%.

(b) Precision: This helps determine how often is the classifier true prediction of a 'Heart Disease' out of its total 'Heart Disease' prediction. It is, $\frac{TP}{Total\ Predicted``\ Heart\ Disease\ "} = \frac{22}{31} \sim 0.71$ i.e., 71% of the times.

(c) Recall: This determines how often does the classifier predict a 'Heart Disease' when the individual actually has a 'Heart Disease' condition–It is in fact, $\frac{TP}{Total``\ Heart\ Disease\ "} = \frac{22}{36} \sim 0.60$

Given, the relatively low Recall, we use Random Forest algorithm on Cleveland data to help select the six most important features in order to help improve the classification rate.

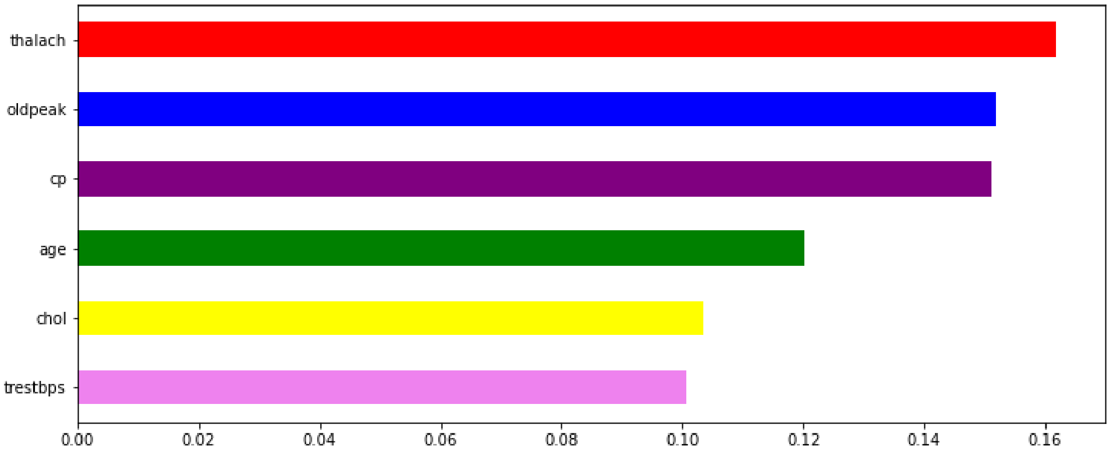
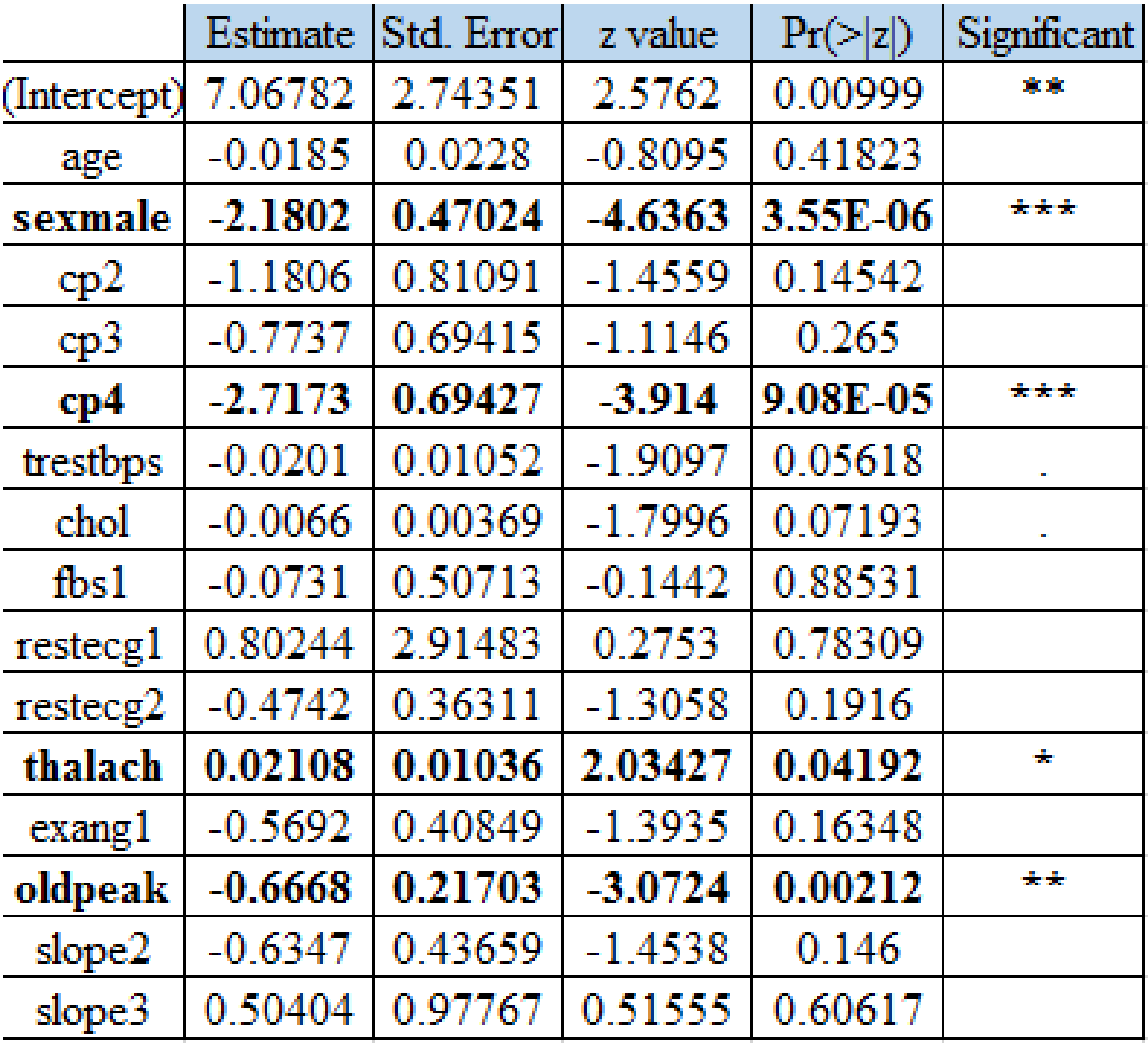


Figure 4.4: Random Forest Feature Importance on Cleveland

According to the output in Figure 4.4, 'thalach', 'oldpeak' and 'cp' are most important factors, and each is statistically significant in a larger model as built previously; we did not include 'thalach' in part because the risk is 5.2%. We also note that Age is a more important factor than 'sex' which is less important than 'restbps'.

4.2.3 Regression Output of Full Model of Statlog Data

After ensuring that each variable has the proper coding and structure, we run a logistic regression on all the (12) features against the target variable. The regression output is shown in Table 4.3.



	Estimate	Std. Error	z value	Pr(>|z|)	Significant
(Intercept)	7.06782	2.74351	2.5762	0.00999	**
age	-0.0185	0.0228	-0.8095	0.41823	
sexmale	**-2.1802**	**0.47024**	**-4.6363**	**3.55E-06**	***
cp2	-1.1806	0.81091	-1.4559	0.14542	
cp3	-0.7737	0.69415	-1.1146	0.265	
cp4	**-2.7173**	**0.69427**	**-3.914**	**9.08E-05**	***
trestbps	-0.0201	0.01052	-1.9097	0.05618	.
chol	-0.0066	0.00369	-1.7996	0.07193	.
fbs1	-0.0731	0.50713	-0.1442	0.88531	
restecg1	0.80244	2.91483	0.2753	0.78309	
restecg2	-0.4742	0.36311	-1.3058	0.1916	
thalach	**0.02108**	**0.01036**	**2.03427**	**0.04192**	*
exang1	-0.5692	0.40849	-1.3935	0.16348	
oldpeak	**-0.6668**	**0.21703**	**-3.0724**	**0.00212**	**
slope2	-0.6347	0.43659	-1.4538	0.146	
slope3	0.50404	0.97767	0.51555	0.60617	

Table 4.3: Statlog data Full Model Regression Output

The result shows that **sex, cp, chol, fbs, exang and oldpeak** are statistically significant–surprisingly, the log mean is not statistically significant, compared to the full model of the Cleveland data. Similarly, as we have done it for the Cleveland data, we run a Random forest on the data to select optimal features. The result is shown in Figure 4.5

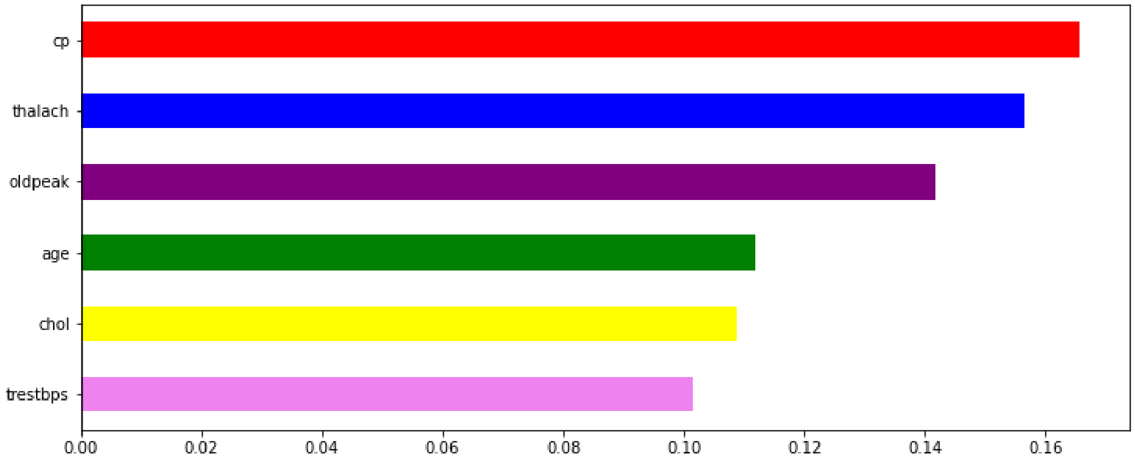
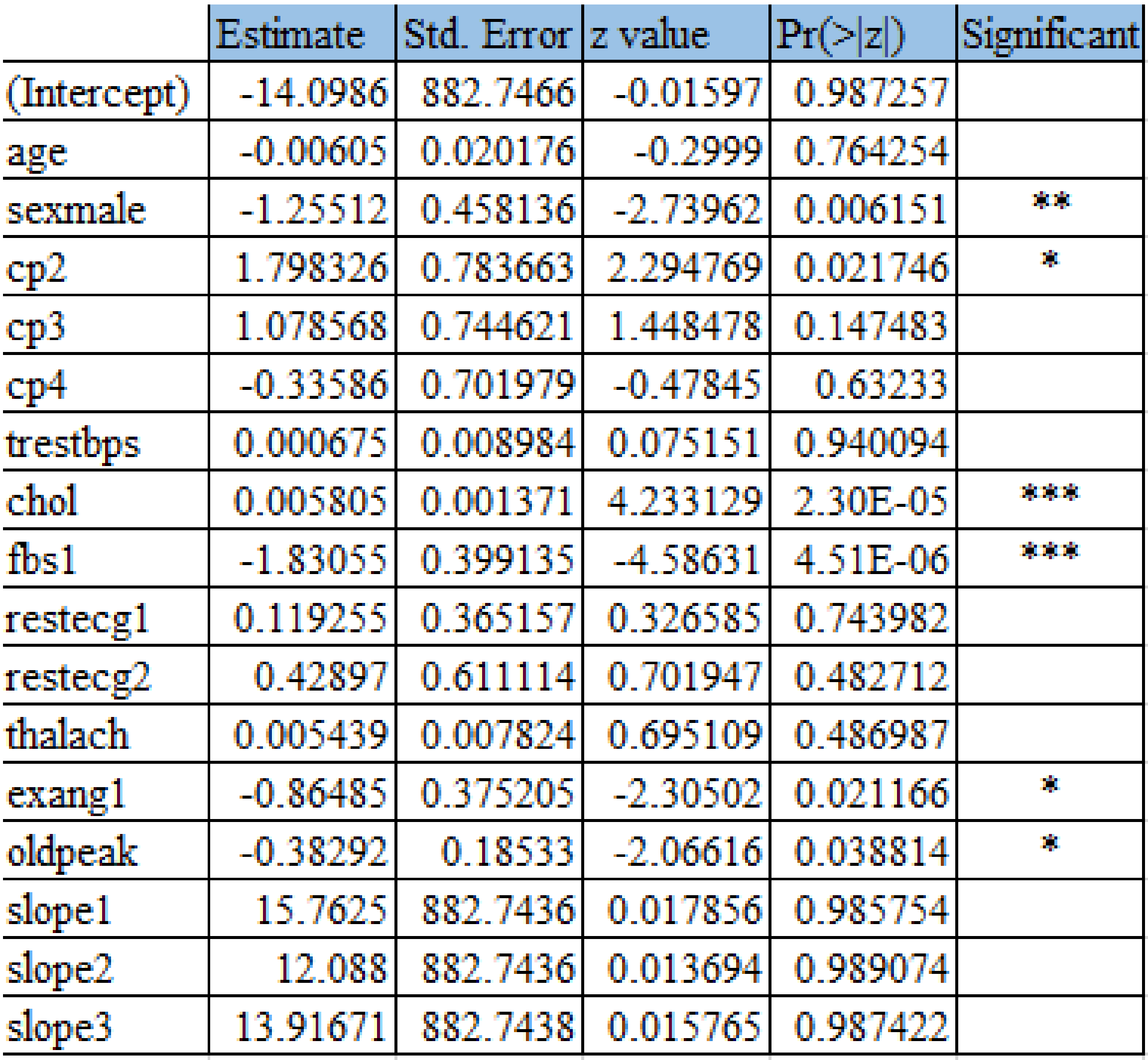


Figure 4.5: Random Forest Feature Importance on Statlog data

Here, we found that all the top 6 features found in the feature selection for the Cleveland data are appearing in the Statlog data, in the exact same order of importance. They are, in order of importance, cp, thalach, oldpeak, age, chol and trestbps. Finally, we consider the (yet to be authenticated), MISSING data.

4.2.4 Regression Output of Full Model of Missing Data

After ensuring that each variable has the proper coding and structure, we run a logistic regression on all the (12) features against the target variable. The regression output is shown in Table 4.4.



	Estimate	Std. Error	z value	Pr(>|z|)	Significant
(Intercept)	-14.0986	882.7466	-0.01597	0.987257	
age	-0.00605	0.020176	-0.2999	0.764254	
sexmale	-1.25512	0.458136	-2.73962	0.006151	**
cp2	1.798326	0.783663	2.294769	0.021746	*
cp3	1.078568	0.744621	1.448478	0.147483	
cp4	-0.33586	0.701979	-0.47845	0.63233	
trestbps	0.000675	0.008984	0.075151	0.940094	
chol	0.005805	0.001371	4.233129	2.30E-05	***
fbs1	-1.83055	0.399135	-4.58631	4.51E-06	***
restecg1	0.119255	0.365157	0.326585	0.743982	
restecg2	0.42897	0.611114	0.701947	0.482712	
thalach	0.005439	0.007824	0.695109	0.486987	
exang1	-0.86485	0.375205	-2.30502	0.021166	*
oldpeak	-0.38292	0.18533	-2.06616	0.038814	*
slope1	15.7625	882.7436	0.017856	0.985754	
slope2	12.088	882.7436	0.013694	0.989074	
slope3	13.91671	882.7438	0.015765	0.987422	

Table 4.4: Missing data Full Model Regression Output

The result shows that **sex, cp, chol, fbs, exang and oldpeak** are statistically significant–Here also, the log mean is not statistically significant, compared to the full model of the Cleveland data. Once again, we run a Random forest on the data to select optimal features. The result is shown in Figure 4.6

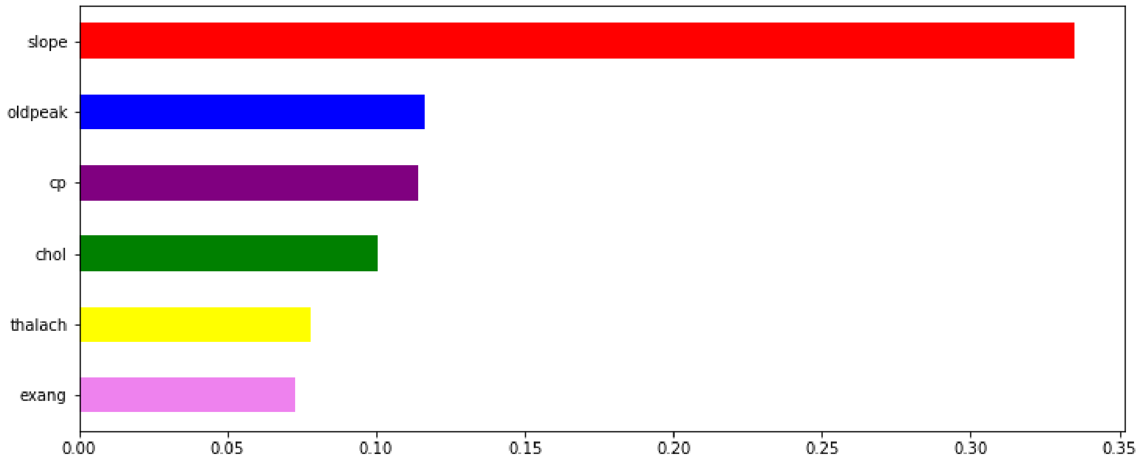
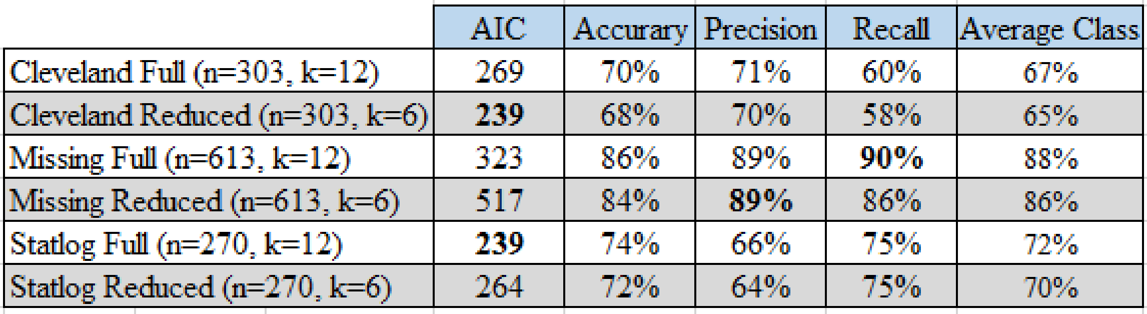


Figure 4.6: Random Forest Feature Importance on MISSING data

Here, we found that not all the top 6 features found in the feature selection for the Cleveland and Statlog datasets have appeared in the list; age is replaced by **slope** which becomes the most dominantly important feature. Also, **exang** replaces trestbps. The remaining important common features for all three datasets are: **oldpeak, cp, chol, thalach**, which are also found to be important in the previous two models. Of all four features, only cholesteral (chol) is not statistically significant at a 5% risk level for the Cleveland data.

4.2.5 Models Comparison and Conclusion

For each of the datasets, Cleveland, Statlog and Missing, we run a logistic regression that includes all features; these are considered *full* models and we record each AIC value. The top 6 features selected through Random forest for each dataset are used to build a second model; these are considered *reduced* models. Their AIC values are also recorded. Finally, for each model built (both full and reduced), we compute the Accuracy, Precision and Recall values as we have shown for the Cleveland (full) model. (see Table 4.3). The results of these models are summarized in Table 4.5

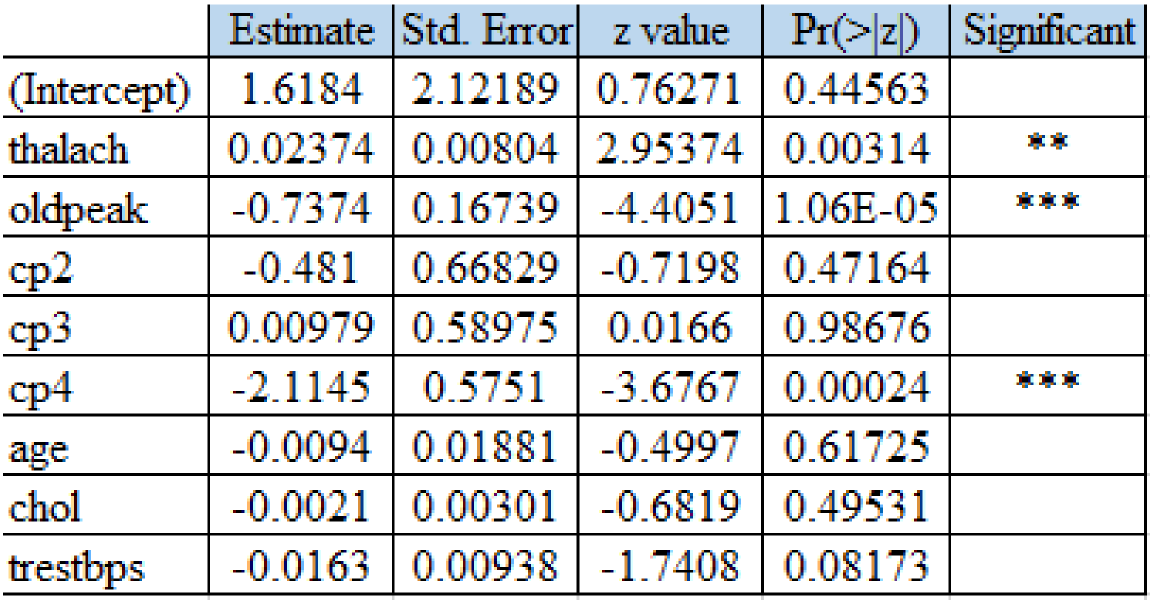


	AIC	Accurary	Precision	Recall	Average Class
Cleveland Full (n=303, k=12)	269	70%	71%	60%	67%
Cleveland Reduced (n=303, k=6)	**239**	68%	70%	58%	65%
Missing Full (n=613, k=12)	323	86%	89%	**90%**	88%
Missing Reduced (n=613, k=6)	**517**	84%	**89%**	86%	86%
Statlog Full (n=270, k=12)	**239**	74%	66%	75%	72%
Statlog Reduced (n=270, k=6)	264	72%	64%	75%	70%

Table 4.5: Classifications Criteria Summary of the Models

We found that, the best predictive models are the reduced Cleveland model and the full Statlog models. However, between these two models, the full model on Statlog has a slightly better accuracy and recall percentage compared to the Cleveland data and yet, the Cleveland data is more precised in classifying heart disease patients.

We think either model is better for statistical analysis. Moreover, it is also clear that, the bigger the dataset, the better for machine classifiers. So, we have decided the Cleveland data on the set of six attributes is the winner; although by a slightly minor margin!



	Estimate	Std. Error	z value	Pr(>|z|)	Significant
(Intercept)	1.6184	2.12189	0.76271	0.44563	
thalach	0.02374	0.00804	2.95374	0.00314	**
oldpeak	-0.7374	0.16739	-4.4051	1.06E-05	***
cp2	-0.481	0.66829	-0.7198	0.47164	
cp3	0.00979	0.58975	0.0166	0.98676	
cp4	-2.1145	0.5751	-3.6767	0.00024	***
age	-0.0094	0.01881	-0.4997	0.61725	
chol	-0.0021	0.00301	-0.6819	0.49531	
trestbps	-0.0163	0.00938	-1.7408	0.08173	

Table 4.6: Final (Cleveland Reduced) Model Regression Output

There are three important features that turn out to be significant: thalach, oldpeak, cp (level 4)–chol, age and trestbps are not statistically significant. From the regression table, we obtain the following:

Model Equation:

$Log(odds\ of\ Heart\ Disease) = 1.6 + 0.024 \cdot thalach - 0.74 \cdot oldpeak - 2.11 \cdot cp(level4)$

Interpretations:

- In average, the expected odds value (or average odds value) of getting 'Heart Disease' is quite high ($e^{1.6} \sim 5$ i.e., a 5 to 1 odds ratio). Thus, the average probability is $\frac{e^{1.6}}{1+e^{1.6}} \times 100\% \sim 49.5\%$ of getting 'Heart Disease', according to the Cleveland Data. In other words, one in two people will likely become 'heart patient'.

- A unit change in thalach value increases the odds of getting 'Heart Disease' by a factor of $e^{0.024} \sim 1.02$ above the average, holding the effect of the factors constant. In which case, the probability of getting 'Heart Disease' increases by about $(e^{0.024} - 1) \times 100\% \sim 2.4\%$ compared to the average.

- Also, when compared to patients with typical angina (cp-level 1), the average odds of getting heart disease for individuals who experience no pain (cp-level 4 or asymptomatic) decreases by a factor of $e^{-2.11} \sim 0.12$. In which case the probability of getting 'Heart Disease' while showing no chest pain (cp-level 4) decreases about $(1 - e^{-2.11}) \times 100\% \sim 88\%$ compared to those who experience typical angina. Other pain levels (2-3) are not statistically significant.

Chapter 5 Conclusion and Future Research

We investigate several heart disease datasets that are often used in both statistical analysis and machine learning. We found several deficiencies and duplicates which undoubtedly weaken any inference or predictive model built on these datasets. We solved several of these issues or at least proposed solutions. Further, we analyze the most likely valid datasets and determine using Random forest and logistic regression as the best dataset for machine learning and statistical analysis: the Cleveland data on a reduced (features) model. Finally, our analysis reveals three leading factors for heart disease: thalach (maximmum heart rate), oldpeak and cp (chest pain). Future work can combine Cleveland data and 'Missing' data to further enhance the classification rate of any classifier. Other classifiers such as gradient boosting, genetic algorithm, support-vector machines and even neural network could be used on the final combined data to further determine features' associations.

Bibliography

[1] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Retrieved from: `https://www.sciencedirect.com/science/article/abs/pii/S0169260713000801`

[2] American Heart Association. (2018). Retrieved from: `https://www.heart.org/en/impact-map`

[3] Center for Disease Control. (2020a). Interactive Atlas of Heart Disease and Stroke. Retrieved from: `https://nccd.cdc.gov/dhdspatlas/`

[4] Center for Disease Control. (2020b). Blood Pressure. Retrieved from: `https://www.cdc.gov/bloodpressure/facts.htm`

[5] El-Bialy, R.,Salamay M. A., Karam O. H., & Khalifa M. E.(2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. Retrieved from: `https:\www.sciencedirect.com/science/article/pii/S1877050915029622`

[6] Hartman, B.(2017). What Everybody Ought to Know About Cholesterol. Retrieved from: `http:\longevityfacts.com/what-everybody-ought-to-know-about-high-cholesterol-levels/`

[7] Heart Foundation (HF). (2021). Retrieved from: `https://www.heartfoundation.org.au//heart-age-calculator`

[8] Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R., & Aha, D. W. (1988). Heart Disease Data Set. Retrieved from:`https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

[9] King, H. T. (2019). What Are the Signs of High Blood Pressure?. Retrieved from:`https://www.keckmedicine.org/what-are-the-signs-of-high-blood-pressure/`

[10] King, R. D. (1992). Statlog Project Data Set. Retrieved from:`https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29`

[11] Mayo Clinic (MC). (2021). Heart disease. Retrieved from:`https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118`

[12] Medina, M.(2020). 4 Types of Heart Disease - and How to Help Prevent Them. Retrieved from:`https://www.keckmedicine.org/4-types-of-heart-disease-and-how-to-help-prevent-them/`

[13] National Heart, Lung, and Blood Institute. (2020). Coronary Heart Disease. Retrieved from:`https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease`

[14] Nelson, J.(2017). Thallium Stress Test. Retrieved from:`https://www.healthline.com/health/thallium-stress-test`

[15] Qin, Y., Yan, G., Qiao, Y., Ma, C.,Liu, J., & Tang, C. (2019). Relationship between Random Blood Glucose, Fasting Blood Glucose, and Gensini Score in Patients with Acute Myocardial Infarction. Retrieved from: `http://dx.doi.org.proxy.lib.odu.edu/10.1155/2019/9707513`

[16] Sagi, R. (2018). Heart Disease UCI. Retrieved from:`https://www.kaggle.com/ronitf/heart-disease-uci`

[17] Sekhri, N., Feder, G. S., Junghans, C., Hemingway, H., & Timmis, A. D. (2007). How effective are rapid access chest pain clinics? Prognosis of incident angina and non-cardiac chest pain in 8762 Consecutive Patients. Heart. London Vol. 93, Iss. 4, 458. Retrieved from: DOI:10.1136/hrt.2006.090894

[18] USAFacts. (2020). Retrieved from:`https://usafacts.org/articles/top-causes-death-united-states-heart-disease-cancer-and-covid-19/`

[19] WebMD. (2020). Risk Factors for Heart Disease. Retrieved from:`https://www.webmd.com/heart-disease/risk-factors-heart-disease`

[20] Whelton, P. K., Carey, R. M., Aronow, W. S., CaseyJr, D. E., Collins, K. J., Himmelfarb, C. D., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., SmithJr, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., WilliamsSr, K. A., Williamson, J. D., & WrightJr, J. T. (2017). ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. Hypertension. 2018. 71:e13–e115. Retrieved from:`https://www.ahajournals.org/doi/10.1161/HYP.0000000000000065`

[21] Ziad, O., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 1216-1219. doi:http://dx.doi.org.proxy.lib.odu.edu/10.1056/NEJMp1606181

[22] Ziasabounchi, N., & Askerzade, I. N. (2014). A Comparative Study of Heart Disease PredictionBased on Principal Component Analysis and Clustering Methods.

[23] Zriqat, I. A., Altamimi, A. M., & Azzeh M. (2017). A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods.

Appendix

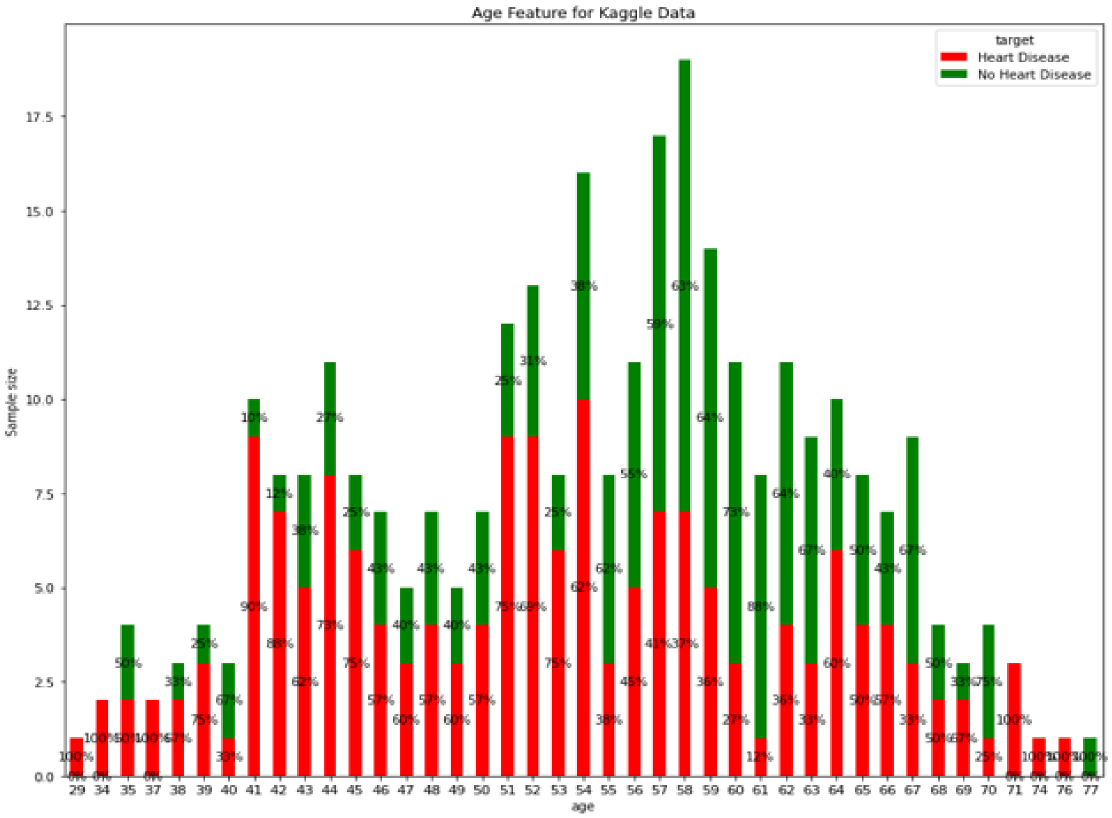


Figure 5.1: Age Feature for Kaggle data

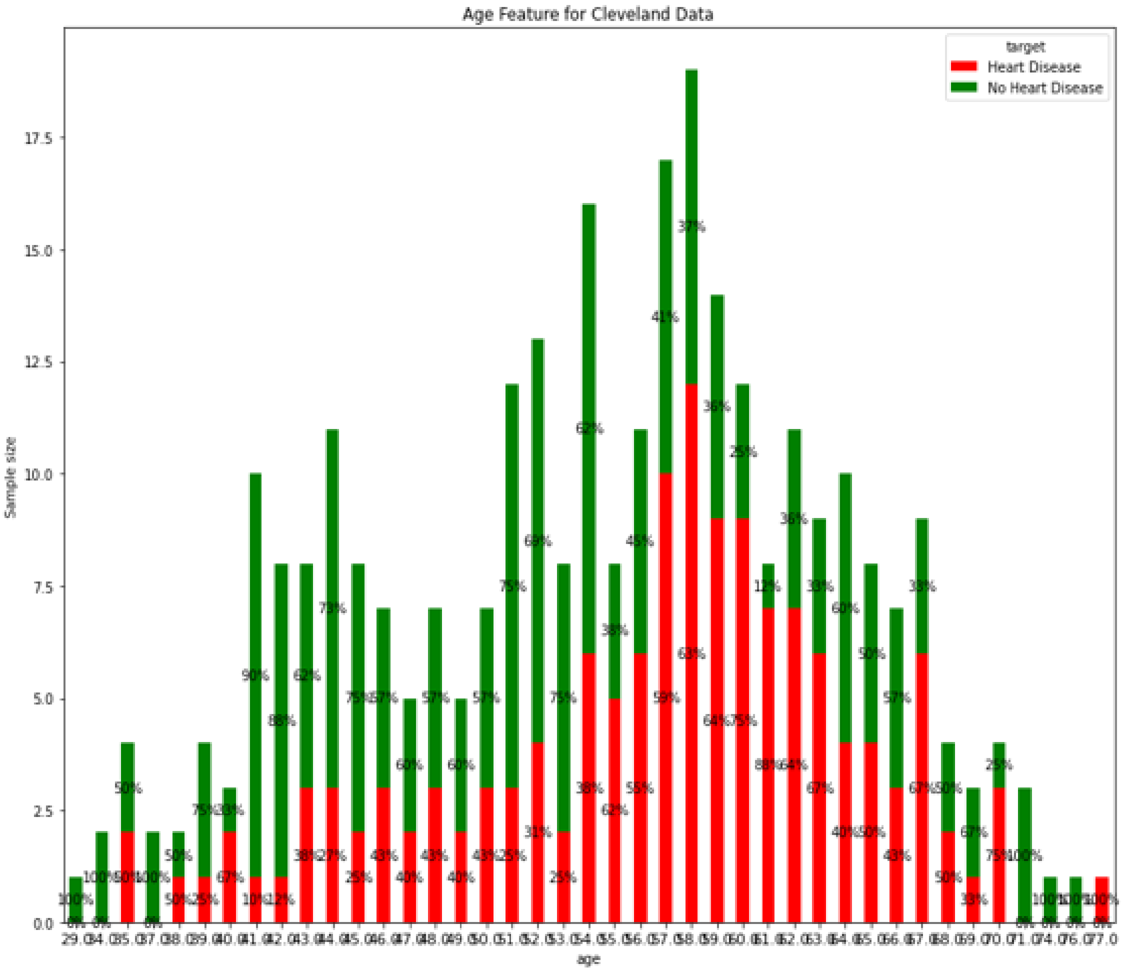


Figure 5.2: Age Feature for Cleveland data

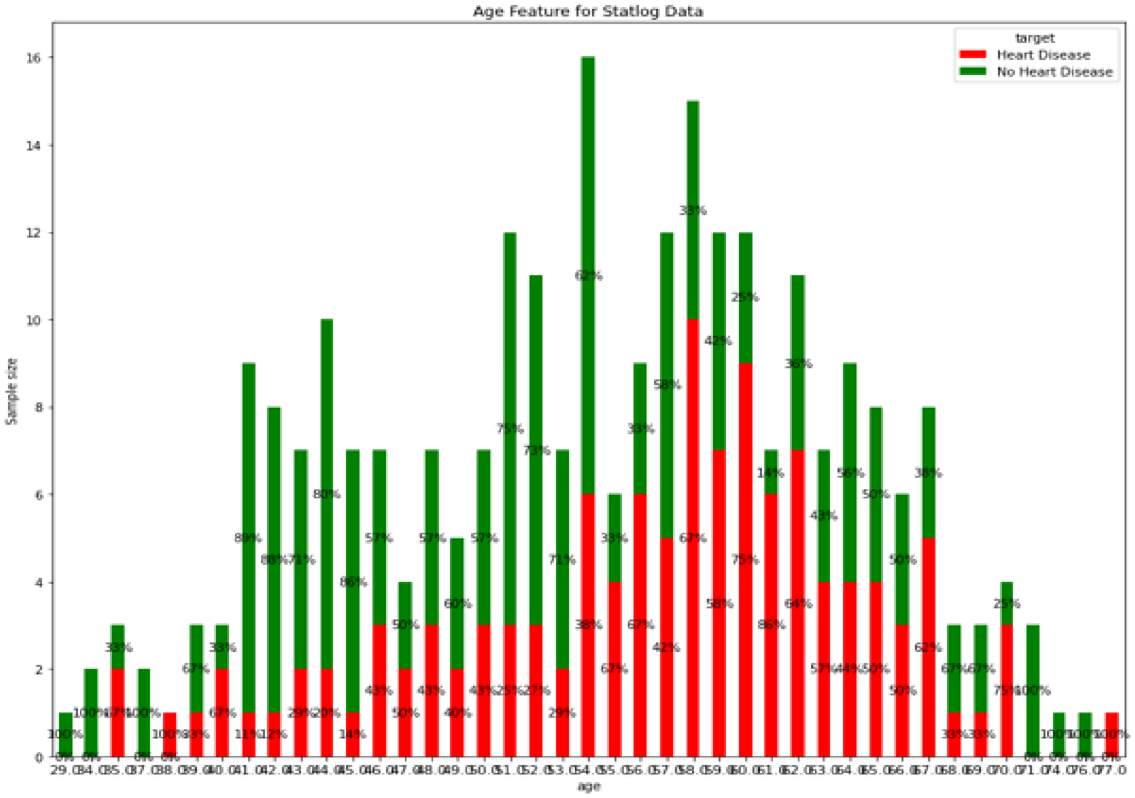


Figure 5.3: Age Feature for Statlog data