NILE UNIVERSITY

Predicting Cardiovascular Disease in Patients

A Math201s Project Report

By

Rawan Mohamed Elfaramawy 202001762

Mohamed Abdelmaged Essawey 202000440

Nadeen Mohamed Farid 202002561

Esraa Negm Sayed 202000799

Sama Ahmed Okasha 202000452

Nouran Hady Shaaban 202001903

Mariam Amr Barakat 202000210

Submitted in partial fulfillment of the requirements

for Math-201 Project

January 6, 2022

ABSTRACT

It is acknowledged that cardiovascular diseases are one of the eras epidemics, thus it is vital to facilitate the diagnosis process and increase its accuracy. In this paper, we explored the definition and general outline of cardiovascular diseases and then dove deep into analyzing and preprocessing our dataset using probability and statistical concepts to extract major insights that were represented in the form of graphs. Lastly, we modeled our dataset using three algorithms: logistic regression, KNN, and SVM. All three produced reasonably acceptable accuracies; however, the former stood out the most.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SECTION I. Introduction

According to WHO [1], cardiovascular disease accounted for 32% of all global deaths in 2019 which is about 17.9 million people. Fortunately, most cardiovascular diseases can be prevented by lifestyle modifications. Nevertheless, technological tools are required to predict people's probability of having cardiovascular diseases. To clarify, when doctors and patients have access to programs that forecast the likelihood of getting cardiovascular disease, not only will it guide doctors to establish a plan for the patient to lower his chances of CVD and other consequent diseases, but also it will aid the patient with visuals that convince him of his destiny if he didn't modify his lifestyle. Such tools may ameliorate patients' consciousness about the disease, which will propel them to take measurable actions.

Our vision for this paper is to create an artificial intelligence (AI) model, which predicts if someone has cardiovascular disease with reasonable accuracy. This is done by using machine learning techniques, specifically logistic regression, SVM and KNN. We will be able to improve the percentage of the model accuracy by training it using the available data posted online and by using probability and statistical concepts.

The paper embarks by having an overview of what is cardiovascular diseases, what causes them and previous research that attempts to utilize datasets to produce an effective model that can effectively predict if this patient has CVD or not. Then, in the paper's core, we commenced by analyzing and comprehending our data, then we preprocessed the data to prepare it for the modelling phase. In the EDA phase, we have collected various insights that helped us verify scientific facts. To add, we have explored modelling using logistic regression, SVM and KNN, and logistic regression turned out to have the highest accuracy rate when used on testing data points.

The main target of our paper is to make a defending mechanism from heart attacks, which can be utilized by doctors to see how likely a patient may have cardiovascular disease. This will be done by asking the patients questions like gender, age, and other heart related measurements.

SECTION II. Background and Literature Review

A. What is cardiovascular disease

According to WHO [1], cardiovascular diseases are a collection of disorders that are related to both the heart and the blood vessels, and it includes coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. To elucidate their danger, all are associated with damage in arteries, which indirectly affects the brain, heart, kidneys, and eyes [1].

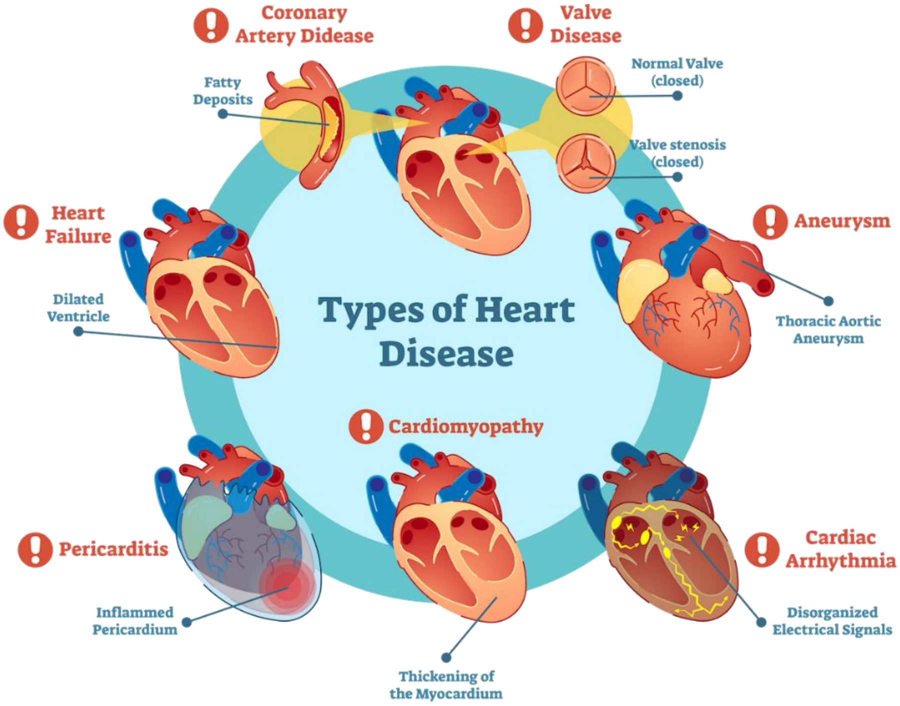
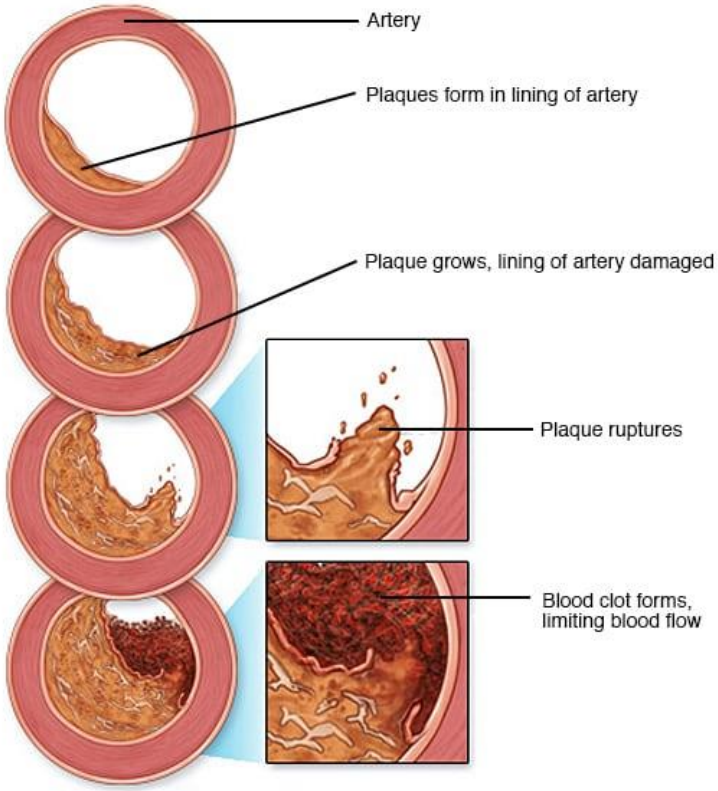


Fig. 1. Types of heart diseases. Adapted from [2]

Figure 1 above explains the myriad types of heart diseases as it is illustrated the problems are viewed in a combination of thickening of walls of the heart, problems in electric synchronicity, valves do not close properly, or most importantly fat deposits inside the coronary artery. All latterly described, are exceedingly harmful and need an early diagnosis to prevent sudden death and other complications.

B. Causes of cardiovascular diseases

At the beginning of the 20th century, the killer, cardiovascular disease, has been widespread especially in the United States. The increasing trend propelled scientists to investigate what are the causes for this sudden epidemic, Kolata and Marx [3] concluded that the risks factors that cause the disease are high blood pressure, high concentration of serum cholesterol, and cigarette smoking. Furthermore, another researcher by Lundell [4] claimed that eating foods that contain high fat and sugar contents cause heart diseases.


Artery
Plaques form in lining of artery
Plaque grows, lining of artery damaged
Plaque ruptures
Blood clot forms,
limiting blood flow

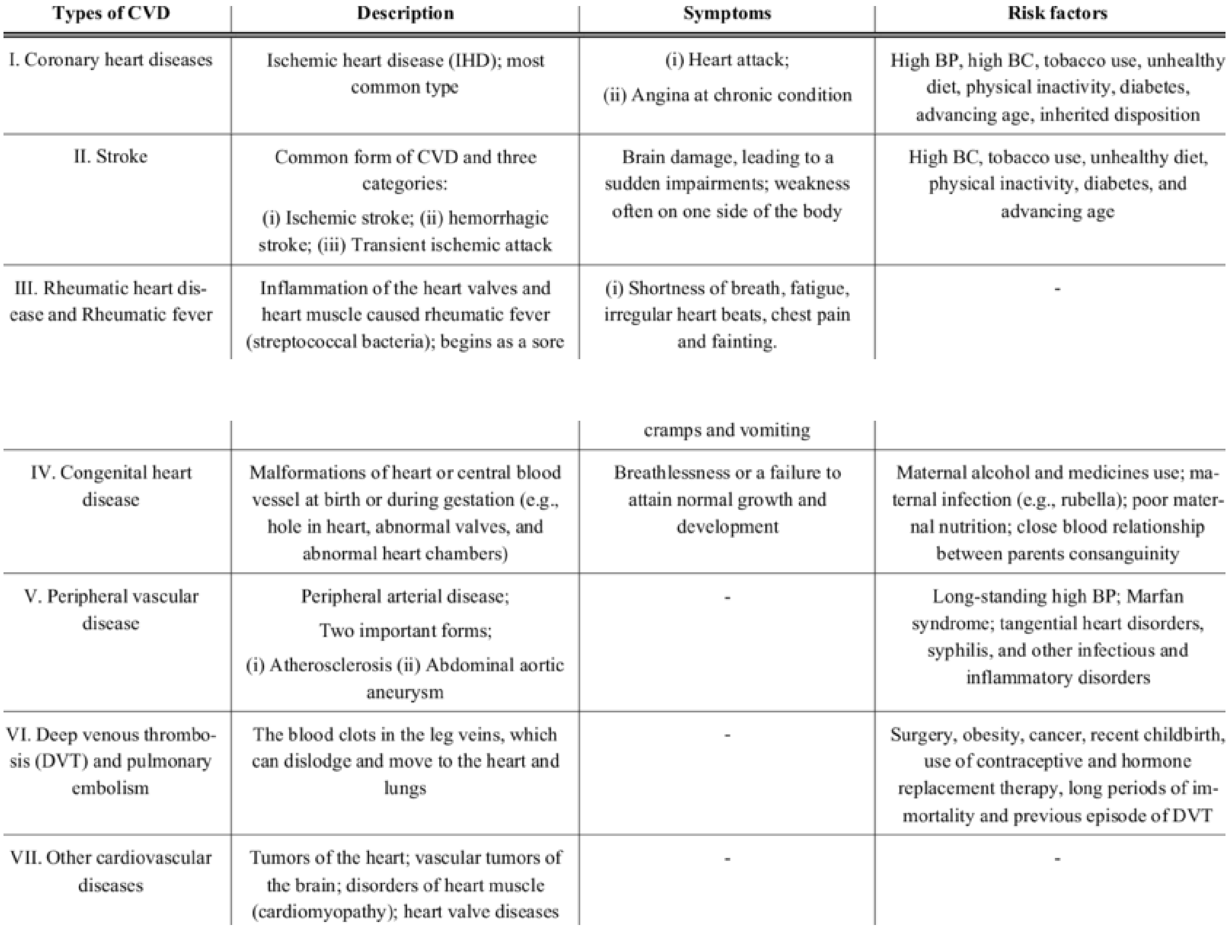
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Fig. 2. The development of atherosclerosis. Adapted from [5]

Figure 2 illustrates how the above causes originate heart diseases. Mainly, the poor lifestyle of patients leads to the constant accumulation of fatty plaques in the arteries, alongside other factors like high blood pressure, this cause the lining of the arteries to get damaged. Consequently, the immune system will respond and start forming blood clots; those blood clots decrease the blood flow and may later lead to heart attacks [6].

C. Factors that predict cardiovascular disease

There are specific symptoms and features in a patient that indicate the presence of heart disease. This includes chest discomfort, shortness of breath, numbness, having diabetes, having hypertension, and having high cholesterol levels [6]. It is crucial to emphasize that, having diabetes vigorously raises the risk of heart disease even when glucose levels are under control [7]. To further explain the relation between heart diseases and diabetes, diabetes results in weakening the blood vessels, this will escalate the rate of damage in vessels and lead to an increase in the clots [6]. Moreover, hypertension indicates the development of heart diseases because the increase in blood pressure also leads to an increase in blood clots. Lastly, high cholesterol levels are a feature in people with heart diseases since the increase in lipids in arteries will narrow the vessels, leading to an increase in blood pressure, again increasing blood clots [6]. Figure 3 below goes into greater depth with the symptoms of each type of CVD. Also, it links it with the risk factors for further clarification.



Types of CVD	Description	Symptoms	Risk factors
I. Coronary heart diseases	Ischemic heart disease (IHD); most common type	(i) Heart attack; (ii) Angina at chronic condition	High BP, high BC, tobacco use, unhealthy diet, physical inactivity, diabetes, advancing age, inherited disposition
II. Stroke	Common form of CVD and three categories: (i) Ischemic stroke; (ii) hemorrhagic stroke; (iii) Transient ischemic attack	Brain damage, leading to a sudden impairments; weakness often on one side of the body	High BC, tobacco use, unhealthy diet, physical inactivity, diabetes, and advancing age
III. Rheumatic heart disease and Rheumatic fever	Inflammation of the heart valves and heart muscle caused rheumatic fever (streptococcal bacteria); begins as a sore	(i) Shortness of breath, fatigue, irregular heart beats, chest pain and fainting. cramps and vomiting	-
IV. Congenital heart disease	Malformations of heart or central blood vessel at birth or during gestation (e.g., hole in heart, abnormal valves, and abnormal heart chambers)	Breathlessness or a failure to attain normal growth and development	Maternal alcohol and medicines use; maternal infection (e.g., rubella); poor maternal nutrition; close blood relationship between parents consanguinity

V. Peripheral vascular disease	Peripheral arterial disease;
	Two important forms;
	(i) Atherosclerosis (ii) Abdominal aortic aneurysm
		-	Long-standing high BP; Marfan syndrome; tangential heart disorders, syphilis, and other infectious and inflammatory disorders

VI. Deep venous thrombosis (DVT) and pulmonary embolism	The blood clots in the leg veins, which can dislodge and move to the heart and lungs	-	Surgery, obesity, cancer, recent childbirth, use of contraceptive and hormone replacement therapy, long periods of immortality and previous episode of DVT
VII. Other cardiovascular diseases	Tumors of the heart; vascular tumors of the brain; disorders of heart muscle (cardiomyopathy); heart valve diseases	-	-

Fig. 3. Cardiovascular diseases symptoms and risk factors. Adapted from [8]

D. Ways to prevent heart disease

For a holistic view, it was necessary to mention how patients can avoid and lower their chances of getting cardiovascular diseases, as this can prevent about 75% of the deaths [1]. All the provided suggestions are lifestyle modifications and not medications.

It is highly suggested to have daily exercise for 75 to 150 minutes since the increase in muscle strength will indirectly strengthen the heart muscle and allow the heart to pump blood at higher efficiency [9]. Additionally, diet is another prominent issue; it is recommended to cut sugar and saturated fats and replace them with vegetables, fruits, and whole grains [9]. This will help in decreasing low-density lipoprotein cholesterol and maintaining a BMI of 20 to 25 [9]. Lastly, the elimination of smoking is a must as smoking increases the risk of CVD by 30%, which indicates it is the riskiest factor to consider [9].

E. Previous research

For background information, two previous papers will be discussed briefly with the technologies they used. Both papers have greatly inspired us to apprehend the variety of probability approaches used to find the likelihood that the patient has cardiovascular disease.

In the first paper [10], researchers were able to analyze data and predict heart disease using different machine learning algorithms. First, they started by preprocessing the data and removing all the outliers to improve the accuracy of the future model. Then they used feature selection algorithms to select vital features as LASSO, and mRMR, this process was vital to simplify the model and have a shorter training time. Moreover, they used a cross-va method called Absolute Shrinkage and Selection Operator, which means that the team reserved a part of the training data to test the model's effectiveness. Finally, many machine learning classifiers algorithms were used and compared. To start, logistic regression with classification had an accuracy of 84%, SVM had an accuracy of 86%, and the worst model was ANN with an accuracy 73%. In conclusion, the best algorithms were logistic regression and SVM [10].

In the second research paper [11], utilized backpropagations of Multi-Layer-Perceptron (MLP) and Artificial Neural Network (ANN) to predict CVD. Firstly, to find tangible patterns, they used CVD well-known Cleveland dataset; afterwards, the data pre-processing was carried out on the data of the 297 patients, to remove outliners and clean the data. Subsequently, through continuous trial and error they produced a prediction model for heart disease with Hybrid Random Forest with Linear Model (HRFLM), which had a noticeable accuracy of 88.7% [11].

SECTION III. Methodology

A. Methodological Approach

As explained before, our main objective in this paper is to develop a model that predicts whether the patient has cardiovascular disease or not. To develop an accurate model, it was crucial to first research for a reliable and unbiased dataset, that links patient features with whether they have heart diseases or not; therefore, we chose the famous Cleveland dataset [12].

Our team aimed to move in a systematic approach; firstly, we commenced by understanding our dataset, its features, and whether each column was discrete or continuous. This helped us later to use the correct probability and statistical concepts. Subsequently, we indulged in Exploratory data analysis (EDV), this greatly assisted us to comprehend the data at a deeper level. To clarify, in this stage, we pre-processed the data, by investigating the missing values, outliers, and wrong values. After cleaning up the dataset, we proceeded by data visualization, which included the usage of Probability Mass Function (PMF) for discrete features, Probability Density Function (PDF) for the continuous features, and correlation Matrix for understanding the connections between each column and the rest. This stage truly gave us various insights about the correlations between the features which were eye-opening. Lastly, we modeled the data using Machine Learning algorithms like logistic regression. This was to create a classification model that can predict whether the patient has heart disease or not.

We followed that approach described above for multiple reasons. Firstly, it was vital to clean up or dataset to wholeheartedly ensure that all the data presented is accurate. Then, for visualizations, we utilized probability functions to help us understand the density and distribution of the data, so that in the modeling phase we are in a better position to judge and accumulate insights. Lastly, we have decided that logistic regression is the best option for our problem, as it will help to classify our data into two categories by holistically finding the relationships between all parameters. All those efforts were to ensure high validity and reliability.

B. Data Collection

To gain better quantitative insight into heart disease prediction in patients, we have used the famous Cleveland Clinic heart disease dataset [12]. It included a total of 303 participants and was accompanied by 13 features and one target attribute for each patient. We have selected this data among other datasets for a couple of reasons. Firstly, the dataset was not extensively huge which allowed us as students to process the information with ease. Moreover, this dataset had minimal defects; in other words, there were no missing data. Lastly, similar research papers also utilized this dataset; thus, it was a superb opportunity to compare our results and accuracies with their research papers.

To go deeper into or dataset, the features are categorized into objective, subjective and examination features. This is thoroughly explained in table I.

Table I Types of features

Objective	It is information on facts such as gender.
Examination	It is the results and the findings of medical examination.
subjective	It is the data and the information provided by the patients.

Moreover, table II describes all the features and what they scientifically mean.

Table II The definition of our dataset features

Gender	It’s either male or female.
CP (Chest Pain)	Angina is chest pain caused by fat depositing in the arterial walls of the heart, which leads to less oxygen-rich blood from reaching the heart. Consequently, leading to chest pain and discomfort, which may be expressed as pressure, squeezing, burning, or fullness. The types of chest pain are typical angina, atypical angina, non-anginal pain and asymptomatic [13].
Typical angina	It’s a reduction in blood flow to the heart and is linked to chest discomfort.
Atypical angina	It is a chest pain not related to heart.
Non-anginal	It’s typically esophageal spasms (non-heart related).
Asymptomatic	It is a chest pain not showing signs of disease.
Threstbps	It’s the person's blood pressure at rest with no exercise and normal pressure.
Col (Cholesterol)	This measures the amount of cholesterol in your blood vessels. These deposits eventually build up, making it difficult for enough blood to pass through your arteries. The reading of 240 mg/dL or higher is considered high [14].
FBS (Fasting Blood Sugar)	It is a test that measure and determine blood sugar level after a patient has fasted for at least eight hours. If fasting blood sugar > 120 mg/dl that’s means that the person has diabetes [15].
Restecg	It’s electrocardiographic findings at rest.
Thalach	It is the person's highest heart rate achieved.
Exang	Its angina caused by exercise.
Oldpeak	It’s a ST depression caused by exercise in relation to the resting state [16].
slope	It is the slope of the ST portion of the peak activity. The types of the slope are upsloping, flatsloping and downsloping. Upsloping is when exercise causes a better heart rate (uncommon). Flatsloping is a minor modification (A normal, healthy heart). Downsloping is indications of an unhealthy heart [16].
CA	The doctor observes the blood flowing through the colored vessel; the more blood motion, the better heart we will get (no clots) [16].

Thal (Thalassemia)	Thalassemia is a blood disorder, in which patients have inherently less hemoglobin. The Types of thalassemia are normal, fixed defect, and reversable defect. Fixed defect is when there is insufficient blood flow in some region of the heart, and reversable defect is when blood movement is not as expected at exercise [16].
Target	Patient either has CVD or not

C. Exploratory data analysis

For clarification, this stage was responsible to analyze and investigate our dataset. It helped us accumulate the main characteristics and trends in our data. Furthermore, this stage has two sub-stages preprocessing the data and visualizing it. In the end, this stage helped us to discover hidden patterns, spot outliners, and anomalous points, and make more accurate assumptions [17]. We used Jupyter Notebook to write our python code to utilize its powerful organization and visualization features, the code can be accessed from the appendix.

In the data preprocessing section, we will check if the 303-entity-data form is genuinely stable. No forbidden value was found in all the data frame and no missing data also known as non-null; however, some statistical measures should be calculated on the quantitative data as mean, minimum, maximum and interquartile range, for further understanding of the data [18]. Final output is indicated in the appendix displayed that minimum and maximum values were sensible for all features.

Table III Collection of standard equations to understand the dataset as a whole

Equation	Variables
$Mean = \frac{\sum x}{n}$ (1)	n is the total number of datapoints, and x is the individual value of a datapoint that will be summed with all the datapoints.
$Median = \frac{n+1}{2}$ (2)	n is the total number of datapoints. This formula results in the position of our median.
$IQR = Q_3 - Q_1$ (3)	Q_1 is the 25th percentile and Q_3 is the 75th percentile. The interquartile range describes the middle 50% of values [18]
$Minimum = Q_1 - 1.5 * IQR$ (4)	Q_1 is the 25th percentile and IQR is the interquartile range.
$Maximum = Q_3 - 1.5 * IQR$ (5)	Q_3 is the 75th percentile.
$Standard\ deviation = \sqrt{\frac{\sum(x-\mu)^2}{n}}$ (6)	n is the total number of datapoints and μ is the mean. The std is used to understand how far the datapoints are from the mean.

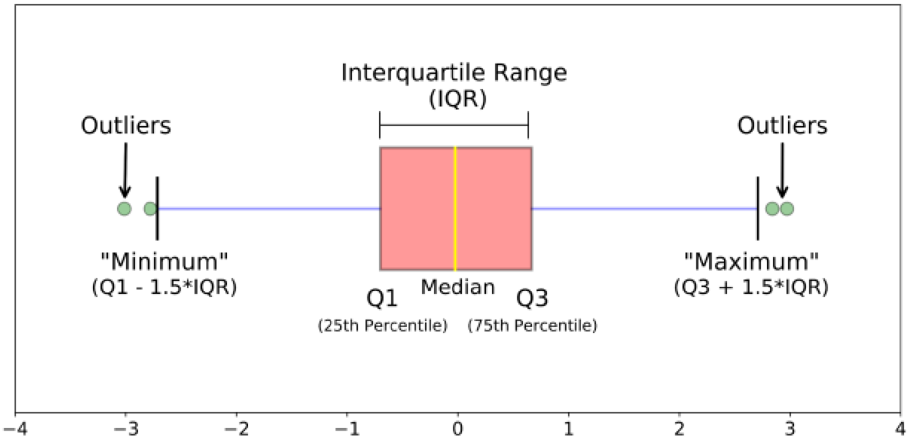


Fig. 4. Explains IQR and outliers

Figure 4 acts as an aid to further understand what is meant by the interquartile range. This was used in the boxplot calculations to help us remove outliers. If a datapoint is lower than the calculated minimum provided in equation 4 or is higher than the calculated maximum in equation 5, then this datapoint is consider an outlier and will be discarded, as it acts as an anomalous point.

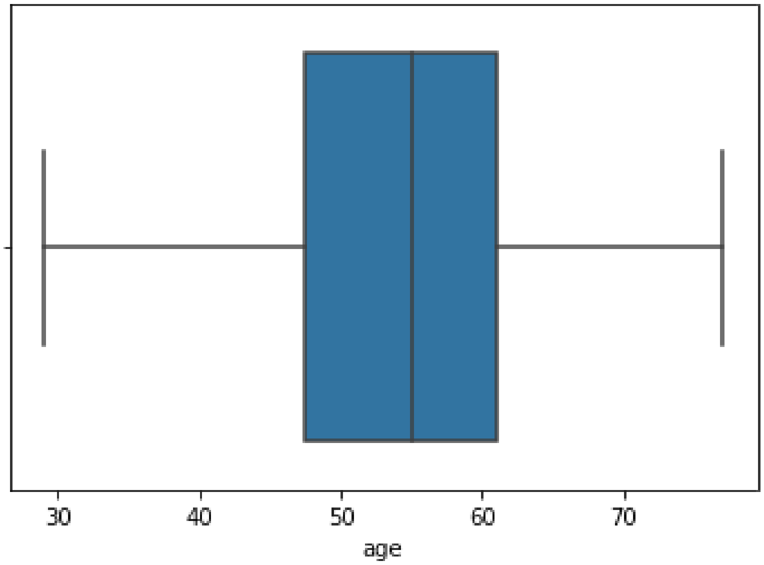


Fig. 5. Age boxplot with no outliers.

Starting with age figure 5, we found that its balance point was 54.3, with a minimum value equal to 29 and a maximum of 77 years. The standard deviation (std) was approximately 9, where most of the values range between 47.5 to 61 years, with 41 unique values. These numbers show that the age’s data is stable. Moving on to the gender, we found that the males are at 70%, while females are at a percentage of 30%.

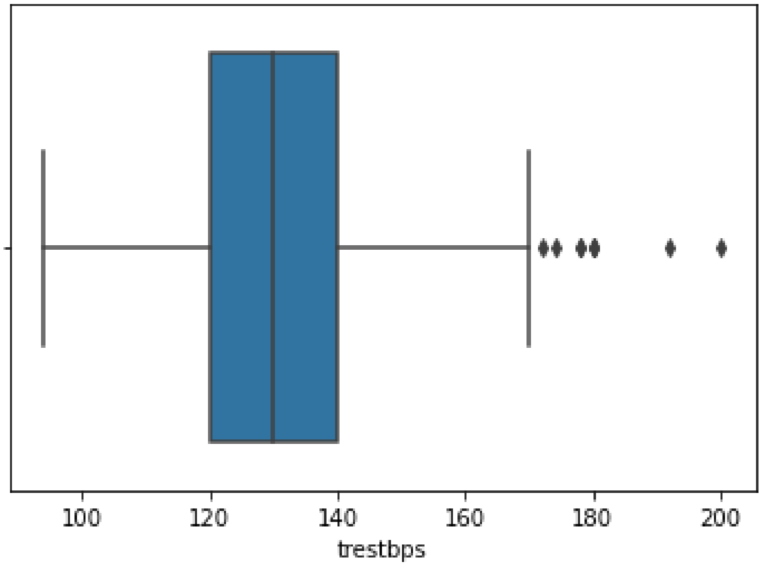


Fig. 6. Resting blood pressure boxplot with outliers.

Secondly in figure 6, the Resting blood pressure, we found that its mean is approximately 131, with a minimum value equal to 49 and a maximum value equal to 200, the standard deviation was equivalent to 17.5, where most of the value’s range between sharp 120 and 140, making the IQR = 20.0, with 94 unique values. While evaluating the data, nine entities were greater than the upper limit and were eliminated. After that, the resting blood pressure measurements are reliable.

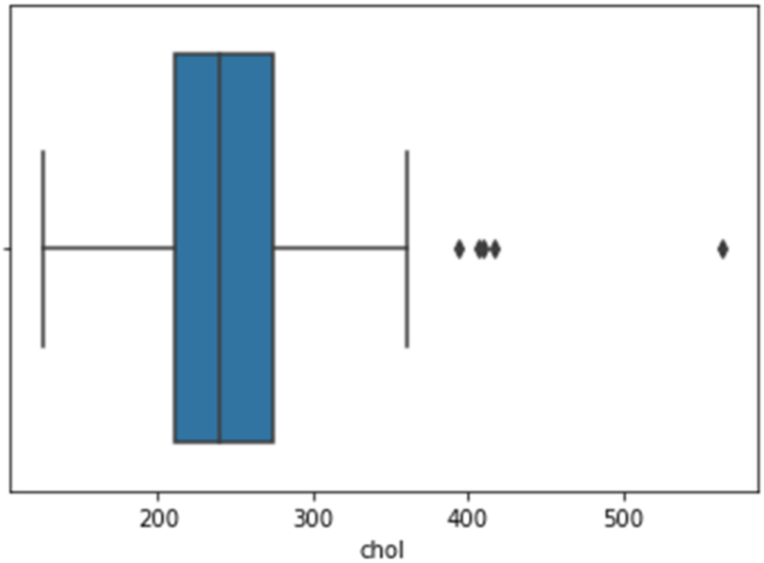


Fig. 7. Cholesterol boxplot with outliers.

Then in figure 7, the serum cholesterol, we found its average is approximately equal to 246.2 mg/dl, with a minimum value equal to 126 mg/dl and the maximum equal to 564 mg/dl, the standard deviation was equivalent to 51.8 mg/dl, where most of the value's range between 211 and 273.75, making the IQR = 62.75, with 152 unique values. While evaluating the data, five entities were greater than the upper limit and were eliminated. After that, the cholesterols' measurements are reliable.

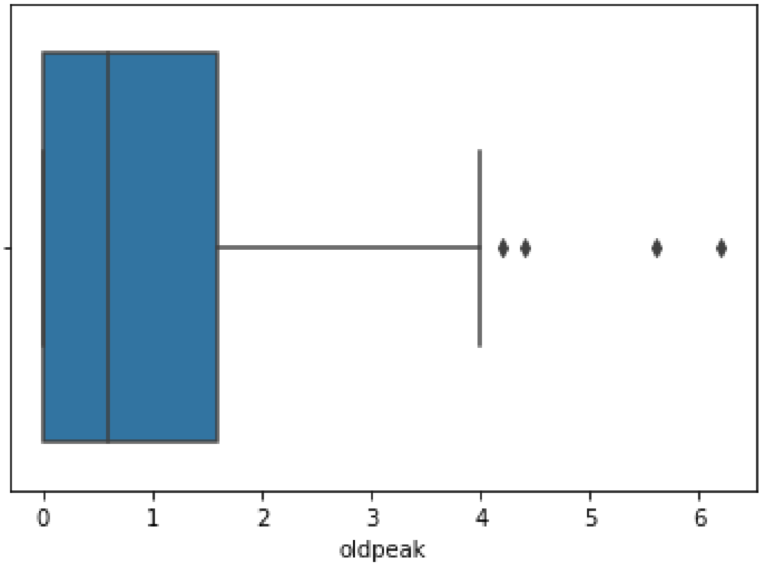


Fig.8. Old peak boxplot with outliers

After that, in figure 8, the ST depression induced by exercise relative to rest, we found its mean is equal to 1, with a minimum value equal to 0 and the maximum equal to 6.2, the standard deviation was equivalent to 1.1, where most of the value's range between 0 and 1.6, making the IQR = 1.6, with 40 unique values. While evaluating the data, four entities were greater than the upper limit and were eliminated. After that, the heart rates' measurements are reliable.

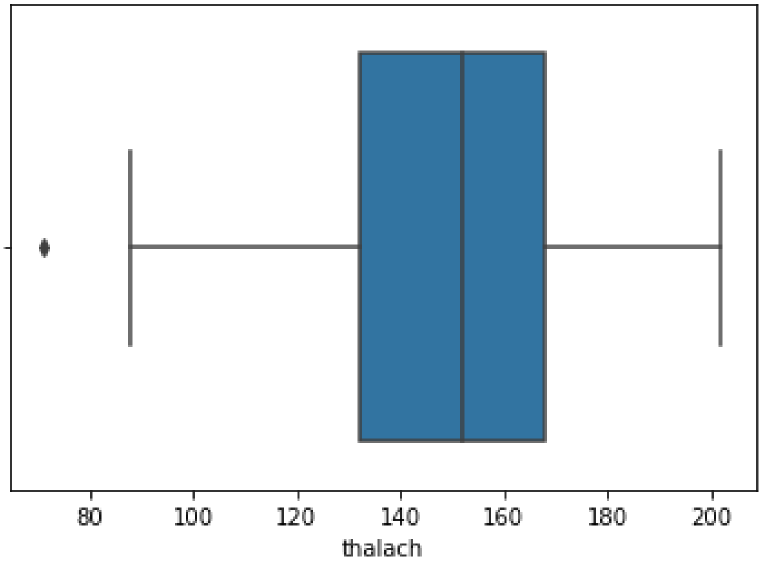


Fig. 9. Thelach boxplot with an outlier

Finally in figure 9, the maximum heart rate, we found mean is equal to 149.6, with a minimum value equal to 71 and the maximum value equal to 202, the standard deviation was equivalent to 22.9, where most of the value's range between 132 and 168, making the IQR = 36, with 91 unique values. While evaluating the data, one entity was greater than the upper limit and was eliminated. After that, the heart rates' measurements are reliable.

.

D.Data Visualization

We began by categorizing our features into discrete and continuous, so that we can use the correct probability approach. Table IV has categorized all the features accordingly.

Table IV Categorizing features

Discrete Features	Continuous Features
Sex	Age
Chest pain (cp)	Resting blood pressure (trestbps)
Fasting blood sugar (fbs)	Serum cholesterol (chol)
Target	ST depression induced by exercise relative to rest (oldpeak)
Resting electrocardiographic (restecg)	Maximum heart rate achieved (thalach)
Slope of the peak exercise ST segment (slope)	
Number of major vessels (ca)	
Types of thalassemia (thal)	
Exercise induced angina (exang)	

1) *Discrete data distribution using Probability Mass Function:*

Probability Mass Function (PMF) is used when we have discrete random variables that have a specific discrete range. It is used to showcase the probability distribution of each discrete random variable.

$$R_x = \{x_1, x_2, x_3, \ldots\}$$

$$P_X(x_k) = P(X = x_k), for\ k = 1,2,3, \ldots, \qquad (7)$$

$$0 \leq P_X(x) \leq 1\ for\ all\ x$$

$$\sum_{x \in Rx} P_X(x) = 1$$

Above is the formula describing the Probability Mass function (7), in which Rx is the random discrete variables that are applicable for each feature. Then for P(X= x), when at the discrete random variable X, the probability will be x. It is worth mentioning that to ensure our PMF is correct, each x should be between 0 and 1 inclusively. Moreover, the summation of all the x should be 1. Below we have calculated the PMF of some of the discrete features. This was done by, using code to count the total occurrence of each discrete random variable, and dividing it by the total number of participants [18].

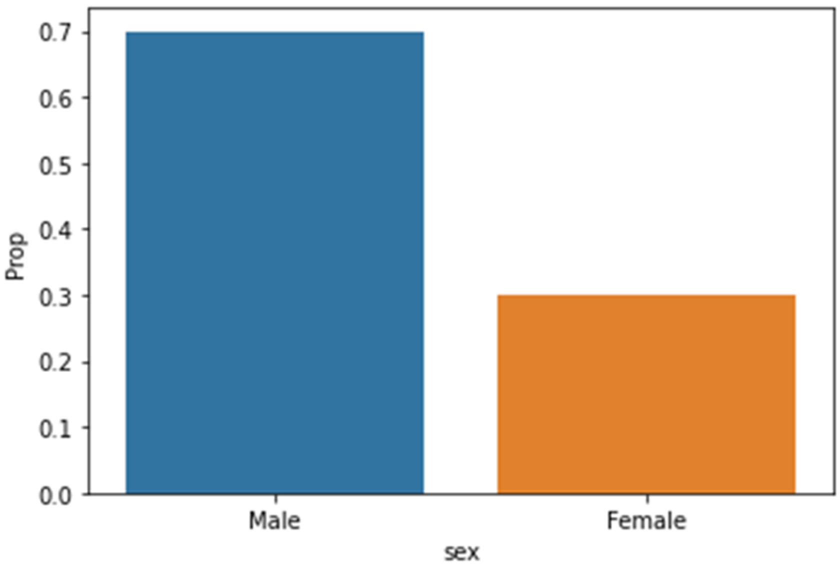


Fig. 10. Sex PMF

Figure 10 describes the PMF of the discrete variables, male and female. And it is obvious that 70% of the participants where males and the rest were females. This is considered a well-balanced probability that will not make our analysis biased, as it shows both genders.

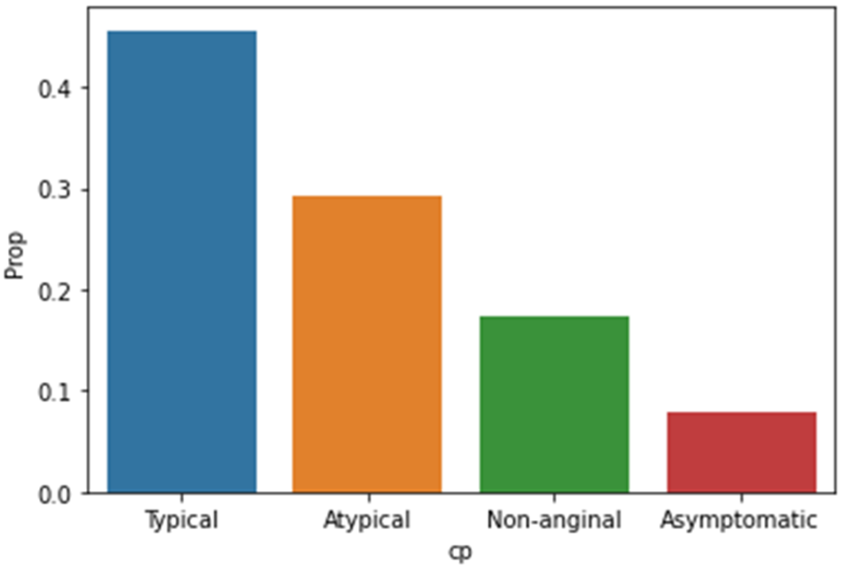


Fig. 11. Chest pain PMF

Figure 11 illustrates the PMF of the different types of chest pain, 45.6% of participants have typical chest pain, 29.3% have atypical angima, 17.3% have non-anginal pain, and the rest are asymptomatic. The data also showcase the variety of types of chest pain in patients, with the Asymptomatic chest pain as the least spread, followed by Non-anginal and Atypical chest pain.

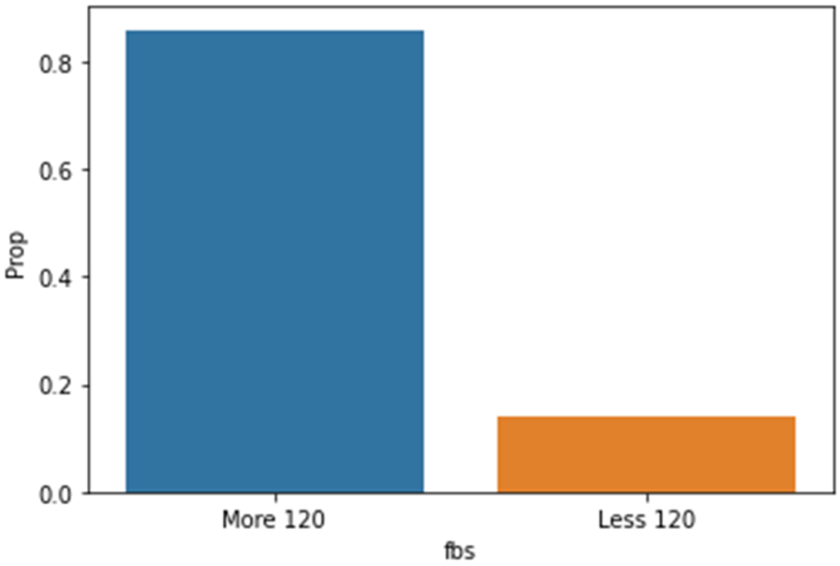


Fig. 12. Fasting blood sugar PMF

Figure 12 provides insight of the high density of patients with fasting blood sugar of more that 120, meaning that about 85.9% of our participants may be prediabetic or diabetic. This is important to consider when finding correlations between diabetes and heart disease.

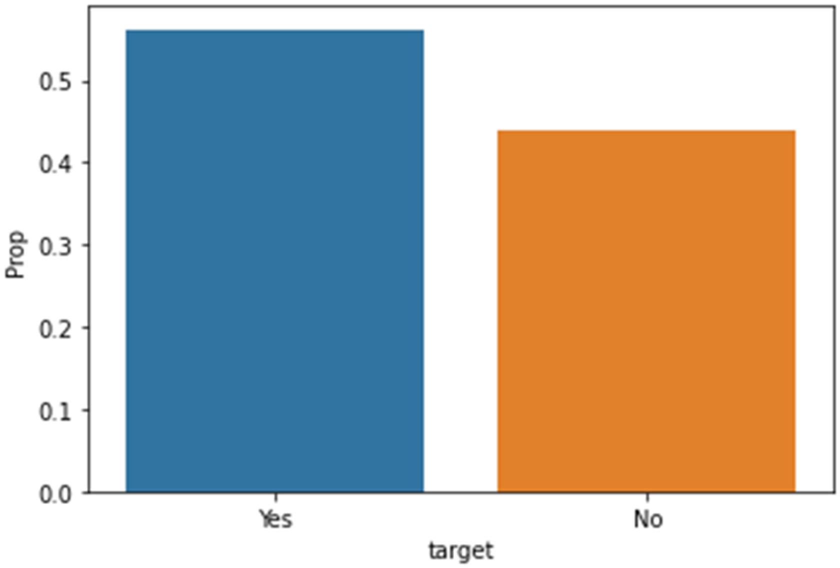


Fig. 13. Target PMF

Figure 13 illustrates the probability distribution of how many participants had heart disease, and how many did not. To our relief, 56.2% of participants had heart disease, and 43.8% did not have heart diseases. In other words, the sample is balance with both diseased and healthy people, which will allow us to further use correlation techniques with higher validity.

1) *Continuous data distribution using Probability Density Function:*

A continuous variable takes on an uncountable number of values, in other words, there are infinite possibilities for our continuous variable that can apply. We will explore the densities of our continuous features such as age, trestbps, chol , oldpeak. However, we will first explore how probability is calculated using PDF so that we can then graph it to represent the densities.

$$P\{a < X < b\} = \int_a^b f(x)dx \ = F(b) - F(a) \quad (8)$$

Above is the formula describing how probability of a range is obtained using the PDF. The X represents our continuous random variable, where a and b are the range of interest where we want to obtain the probability; also, f(x) represents our pdf. The probability can be obtained either by integrating the pdf of the random variable X or by using CDF. To elucidate, F(b) is used to obtain the accumulation of all x values equal to or less than b, then we subtract it with F(a) [18]. To represent this using a histogram, we chose our bins to be equal to 20, so that patterns appear more clearly.

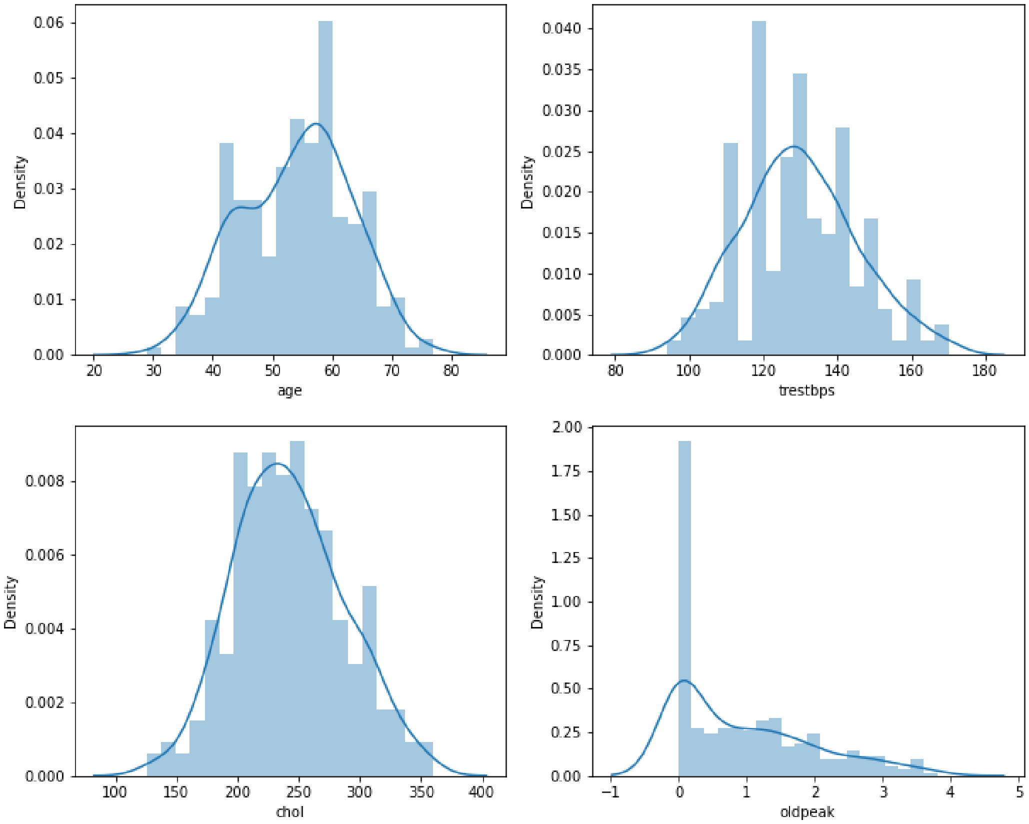


Fig. 14. All continuous features' density

In figure 14, those histograms represent the densities of the random variables age, trestbps, chol and oldpeak respectively. We can see that age, trestbps and chol have their distribution concentrated in the median. However, in the oldpeak histogram is right skewed.

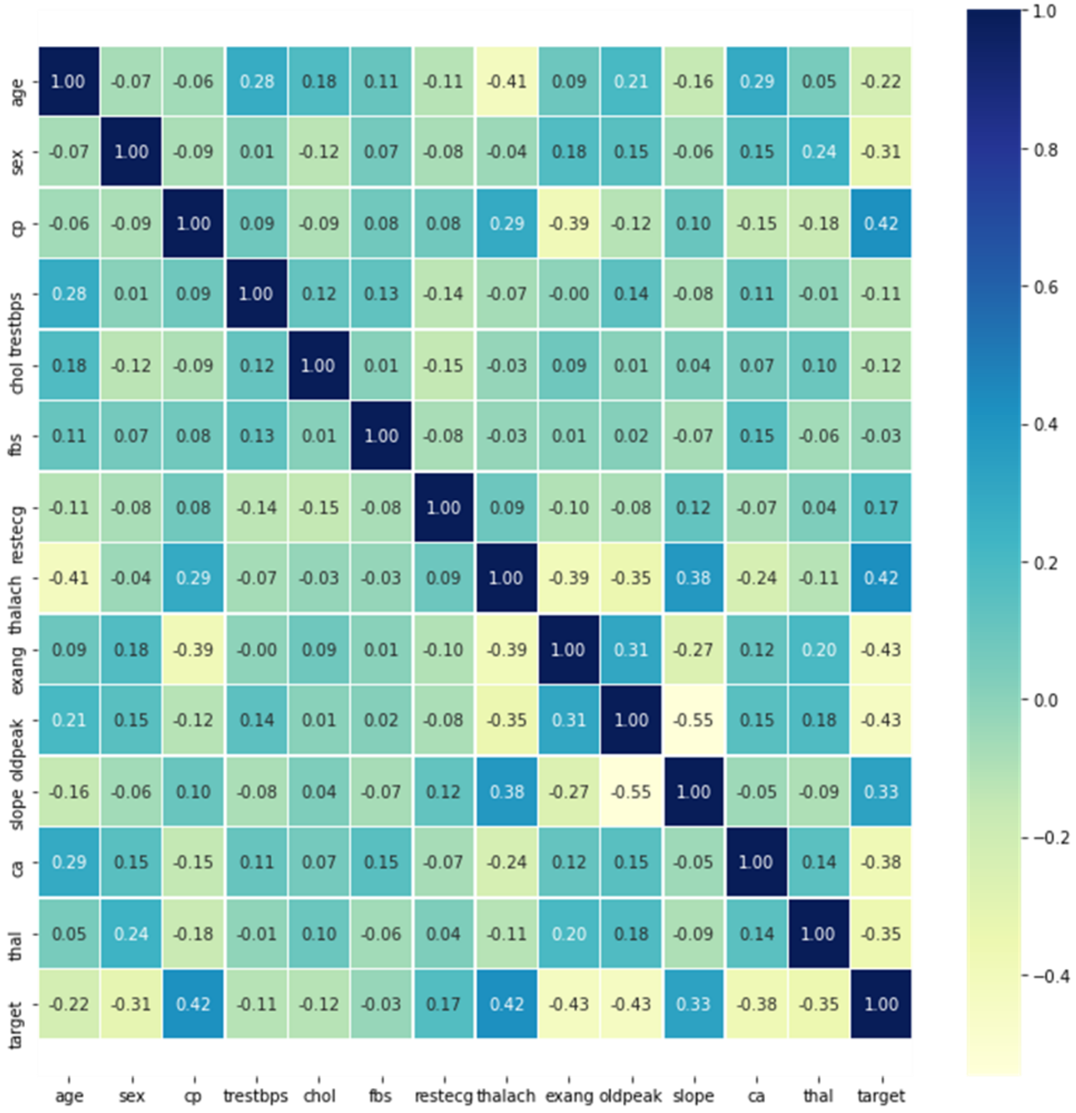
2) *Correlation Matrix:*

This is a statistical concept, that helps us visualize the correlation between all seemingly independent features. To elucidate, it describes the correlation among the variables in a symmetrically square matrix. It is worth noting that the diagonals are always equal to 1.0. Moreover, the correlation in this matrix ranges from -1 to 1. In which negative numbers indicate a negative correlation, and positive numbers indicate positive correlation. Also, as the absolute number increase the higher the correlation [19].

To create such matrix, using code we calculated the correlation between each feature and the rest features, using the correlation coefficient formula (8) shown below. Then this data is organized into a matrix to ease the comprehension of all correlations simultaneously.

$$Correlation\ Coefficent\ Formula = \frac{\sum[(X - X_m)(Y - Y_m)]}{\sqrt{[\sum(X - X_m)^2 \sum(Y - Y_m)^2]}} \qquad (8)$$

This formula first determines the covariance of the variables; in other words, measures the directional relationship between both variables. Then it is divided by the product of both variables' standard deviations. The X and Y are the data points in column x and y respectively, and the X_m and Y_m are the mean of the data sets [20].



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.07	-0.06	0.28	0.18	0.11	-0.11	-0.41	0.09	0.21	-0.16	0.29	0.05	-0.22
sex	-0.07	1.00	-0.09	0.01	-0.12	0.07	-0.08	-0.04	0.18	0.15	-0.06	0.15	0.24	-0.31
cp	-0.06	-0.09	1.00	0.09	-0.09	0.08	0.08	0.29	-0.39	-0.12	0.10	-0.15	-0.18	0.42
trestbps	0.28	0.01	0.09	1.00	0.12	0.13	-0.14	-0.07	-0.00	0.14	-0.08	0.11	-0.01	-0.11
chol	0.18	-0.12	-0.09	0.12	1.00	0.01	-0.15	-0.03	0.09	0.01	0.04	0.07	0.10	-0.12
fbs	0.11	0.07	0.08	0.13	0.01	1.00	-0.08	-0.03	0.01	0.02	-0.07	0.15	-0.06	-0.03
restecg	-0.11	-0.08	0.08	-0.14	-0.15	-0.08	1.00	0.09	-0.10	-0.08	0.12	-0.07	0.04	0.17
thalach	-0.41	-0.04	0.29	-0.07	-0.03	-0.03	0.09	1.00	-0.39	-0.35	0.38	-0.24	-0.11	0.42
exang	0.09	0.18	-0.39	-0.00	0.09	0.01	-0.10	-0.39	1.00	0.31	-0.27	0.12	0.20	-0.43
oldpeak	0.21	0.15	-0.12	0.14	0.01	0.02	-0.08	-0.35	0.31	1.00	-0.55	0.15	0.18	-0.43
slope	-0.16	-0.06	0.10	-0.08	0.04	-0.07	0.12	0.38	-0.27	-0.55	1.00	-0.05	-0.09	0.33
ca	0.29	0.15	-0.15	0.11	0.07	0.15	-0.07	-0.24	0.12	0.15	-0.05	1.00	0.14	-0.38
thal	0.05	0.24	-0.18	-0.01	0.10	-0.06	0.04	-0.11	0.20	0.18	-0.09	0.14	1.00	-0.35
target	-0.22	-0.31	0.42	-0.11	-0.12	-0.03	0.17	0.42	-0.43	-0.43	0.33	-0.38	-0.35	1.00

Fig. 15. Correlation Matrix

Figure 15 beautifully illustrates the outcome of our correlation matrix, as expected the diagonal has 1.0 correlation. What is most important is the obvious positive correlation between chest pain, thelach and slope with the target. Which gives us a clue that those maybe the most important factors that predicts the availability of heart disease.

3) <u>*Comparing target with features:*</u>

In this section we aim to compare specific features with the target to be able to apprehend graphically what the correlation matrix was trying to convey.

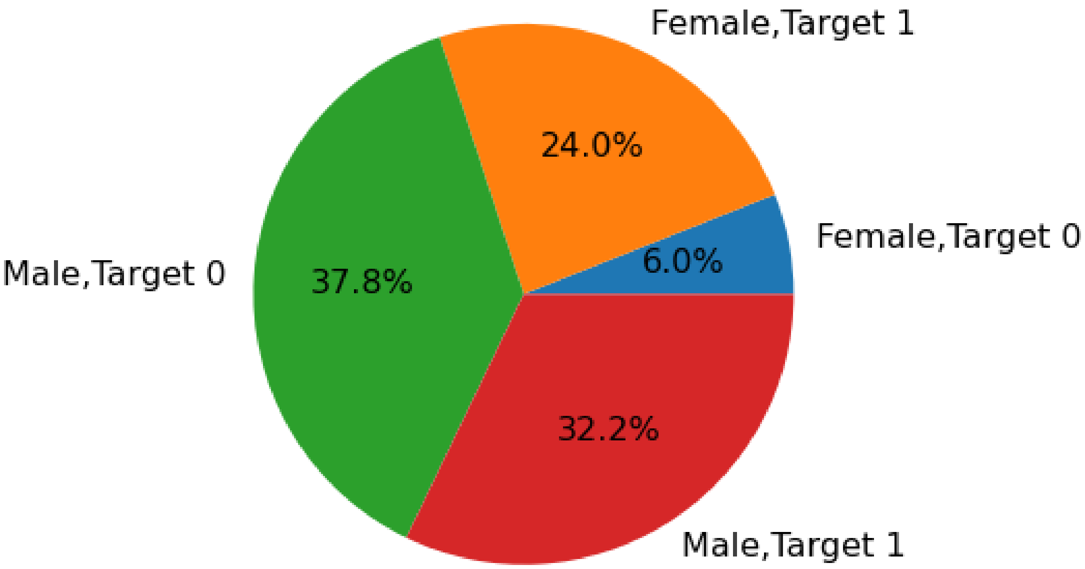


Fig. 16. Gender VS Target

Figure 16 illustrates the comparison of gender and target. It shows that females having heart disease are 24%; this concludes that the ratio of males having heart disease is 30.7%, a little bit higher than females. From afar, we can conclude that males have a higher chance of having heart diseases; however, we must count the fact that most of our participants were males, so this may have given us an inaccurate conclusion [21].

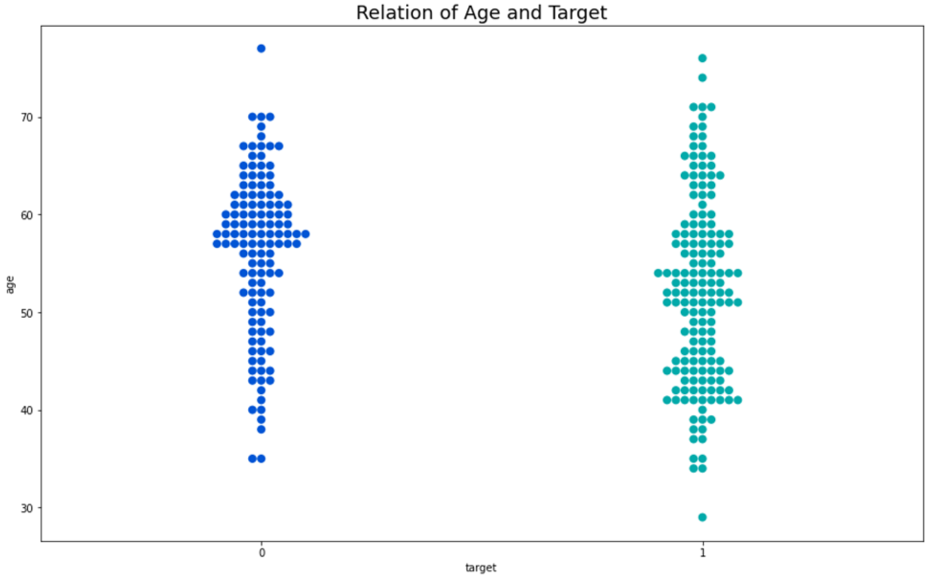


Fig. 17. Age swarm plot

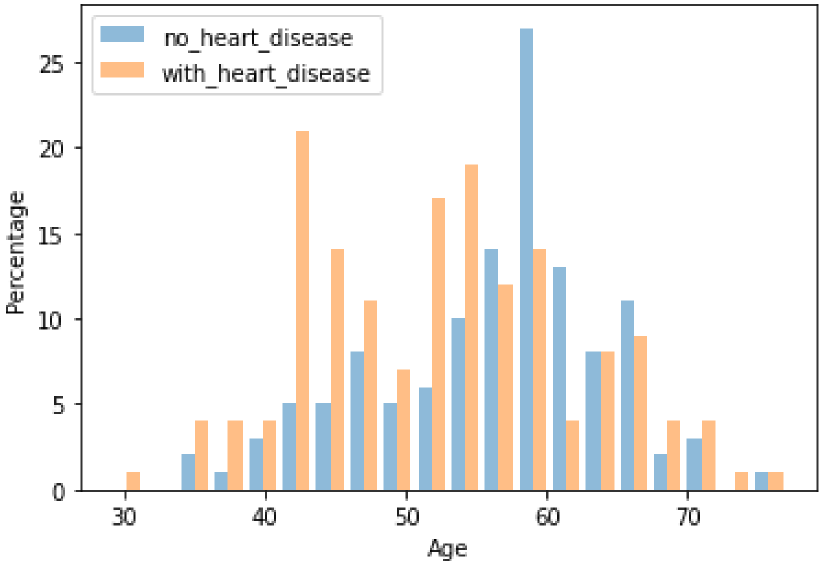


Fig. 18. Age VS Target

Figures 17 and 18 show the link between the age of the patient and having heart disease. The ratio gets higher over the age of forty to sixty. Therefore, people who are above forty are at a higher risk of getting heart diseases. However, when above 60 people who do not have heart disease are increasing.

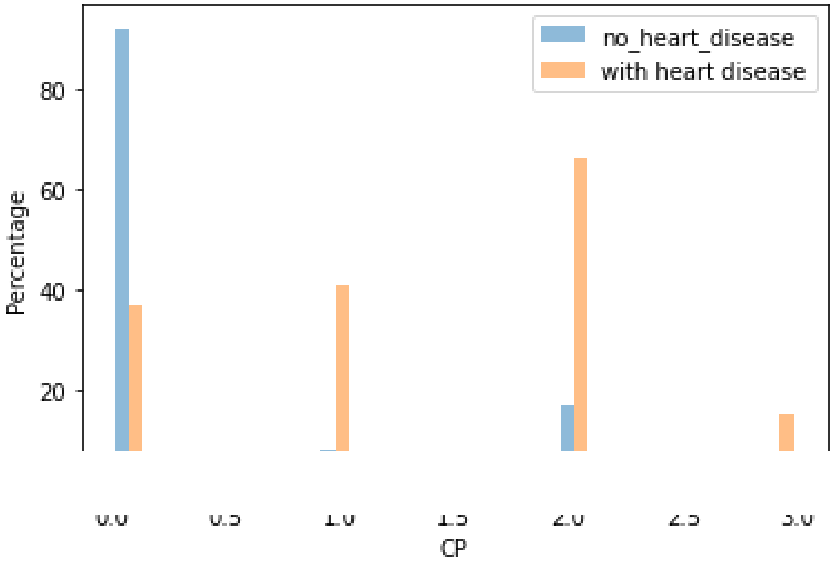


Fig. 19. CP VS Target

Figure 19 reveals the link between the chest pain and the degree of having heart disease. The results showed that people with cp equals to 1, 2, 3 are more likely to have heart disease than people with cp equals to 0. Moreover, people with cp 2 are much more likely to have heart diseases compare to the other types.

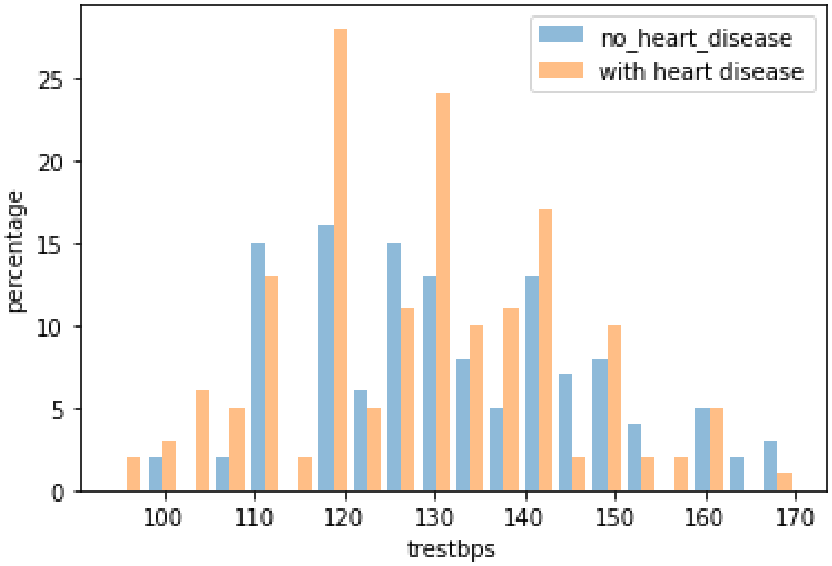


Fig. 20. Trestbps VS Target

Figure 20 clearly shows that having high blood pressure (Trestbps) is a great indicator for heart disease. Furthermore, whether the patients have heart disease or not, over 50% of patients had high blood pressure, this assuming that the ideal blood pressure is at most 120 mmHg.

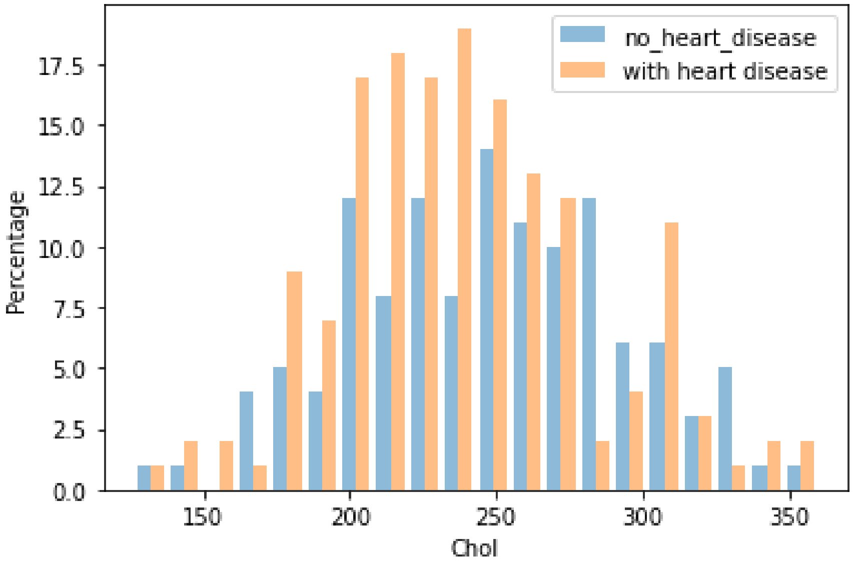


Fig. 21. Chol VS Target

Figure 21 displays the comparison between cholesterol level and the amount of people having CVDs and not having CVDs. The diagram conveys that most people having heart disease have a cholesterol of 200 to 275, which is above normal.

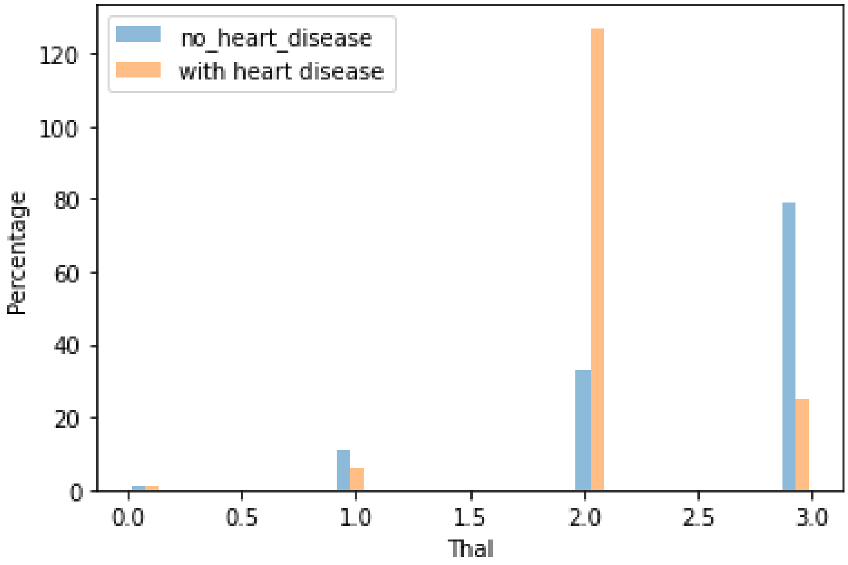


Fig. 22. Thal vs Target

Figure 22 demonstrates the relationship between the thal value and the number of people suffering from heart disease. People with a thal value of 2 are more prone to have heart disease than those with other degrees.

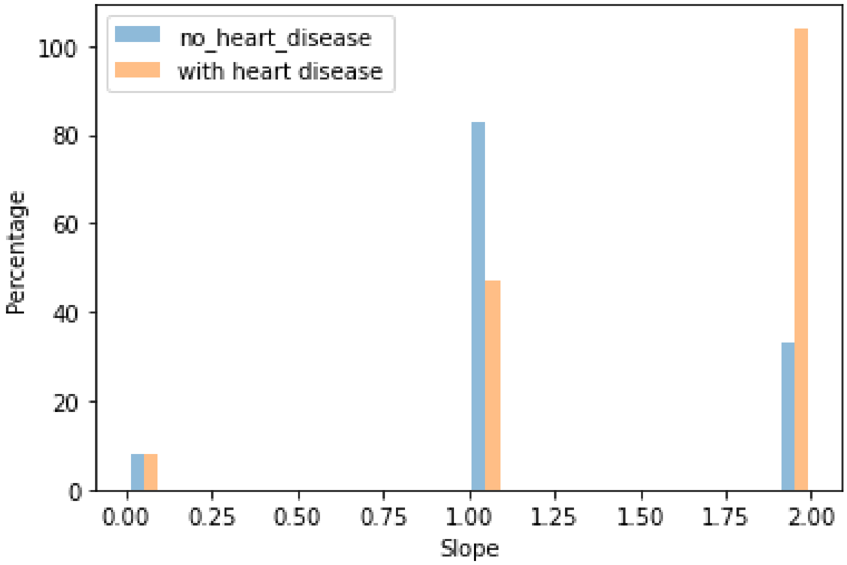


Fig. 23. Slope vs Target

Figure 23 indicates the relationship between the slope of a person, degree of heart pumping heath, and the presence of CVDs. The graph illustrates that people with a slope of 2 are more likely to have heart diseases than those with a slope of 0 or 1.

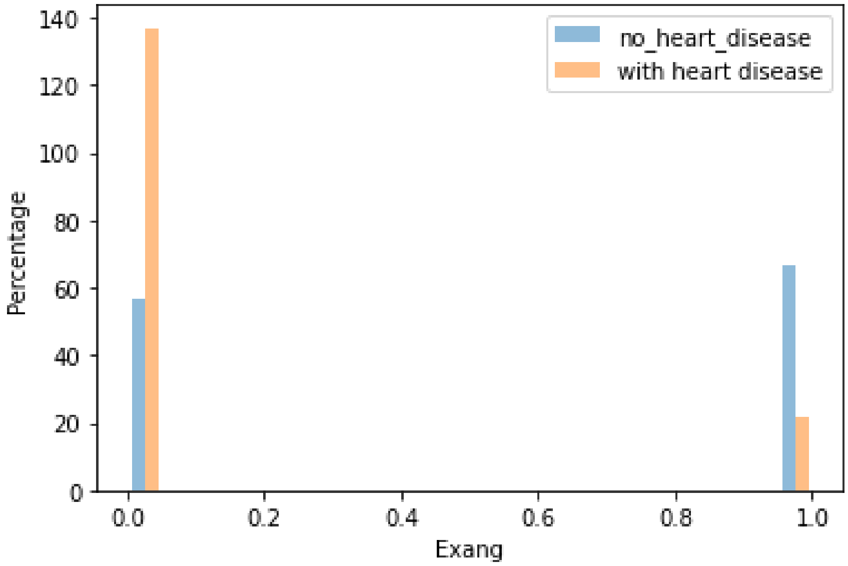


Fig. 24. Exang VS Target

Figure 24 clearly demonstrates the negative correlation of exang with the target. To clarify, you are more likely to have CVDs if your exang is 0; however, you are more likely to not develop CVDs if your exang is 1.

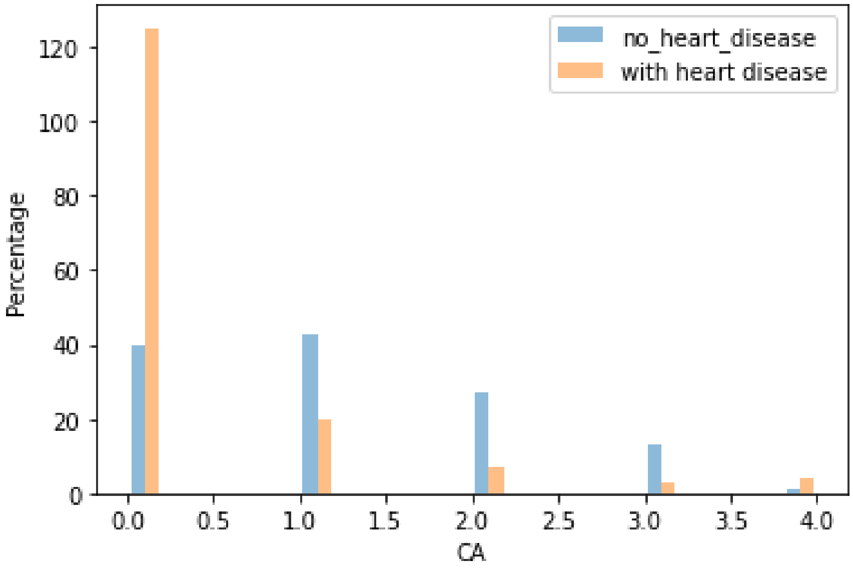


Fig. 25. CA VS Target

Figure 25 demonstrates that the lower the CA the more likely the patient will have CVDs as at CA equal to zero, the number of CVD patients is much higher than proceeding range.

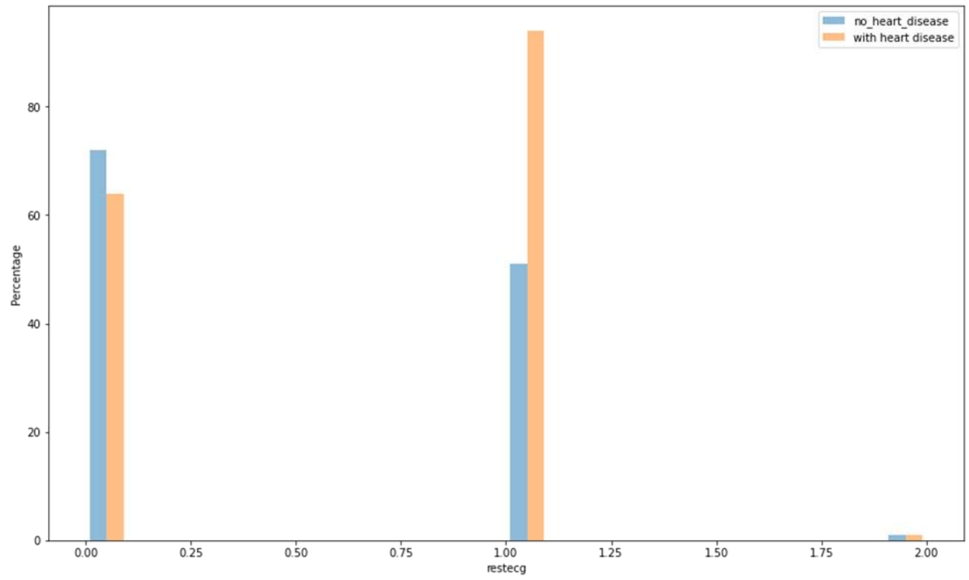


Fig. 26. Restecg VS Target

Figure 26 demonstrates that most people with restecg equal to 1, have a higher chance of getting heart disease. This graph conveys that restecg has a significant effect on heart disease prediction.

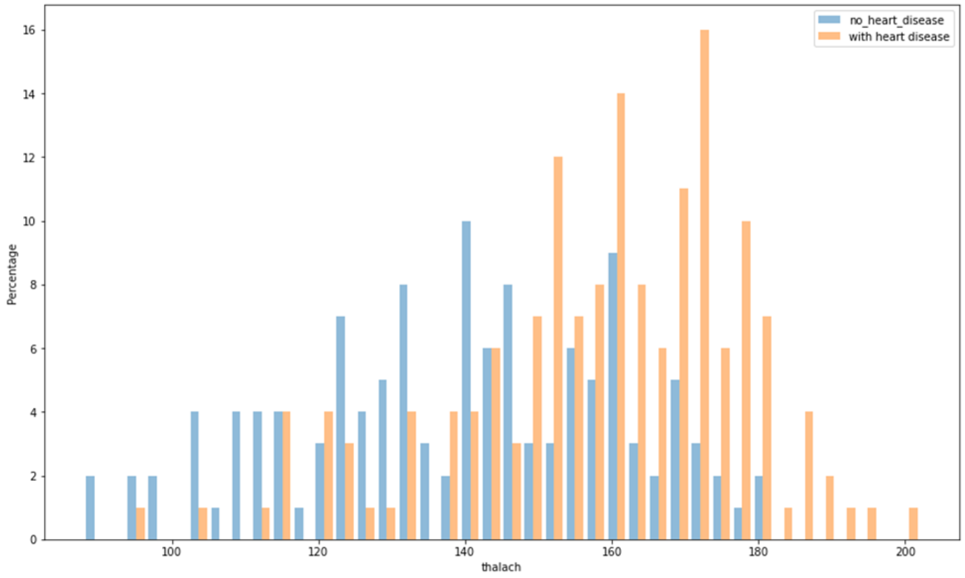


Fig. 27. Thalach VS Target

Figure 27 illustrates that patient with a heart rate greater than 150 are at a greater risk of CVD, this also conveys that thalach has a significant effect on heart disease prediction

.

E. Data Modeling

After our extensive data preprocessing phase, it is time to start modeling our data using multiple supervised learning algorithms until we find the one with the highest training and testing accuracy. To start, we first need to understand the type of output we expect to be able to choose the correct algorithms. To clarify, we expect our output to be whether the patient has heart disease or not, this is considered as a binary classification problem.

Before the modeling phase, we used feature selection to be able to make our model work on all features according to how much it affects the target. To elucidate, if a feature is highly correlated to the target than the rest, it will have a higher effect on the model. This will help our models to produce more sensible results as it will ignore features that are not related to the target.

In our paper, we have chosen logistic regression, Support Vector Machine and K-Nearest Neighbor. All those algorithms are utilized when our problem is of a binary classification nature, in other words, one class is considered normal, and the other is abnormal.

1) *Binomial Logistic Regression*

Our first choice was to implement binomial logistic regression, it is a type of classification model in which it tries to map the dataset to a suitable sigmoid function. Since the sigmoid function has its y-axis spanning from 0 to 1, this verifies that our graph will output a probability, that can then be changed to binary solutions. Logistic regression is considered a very fast algorithm thanks to its effective cost function that halts in the right time and avoids overfitting. It is worth adding that this type of logistic regression is called binomial as there are only two possible outcomes, either we have a patient with CVD or a patient without it [22].

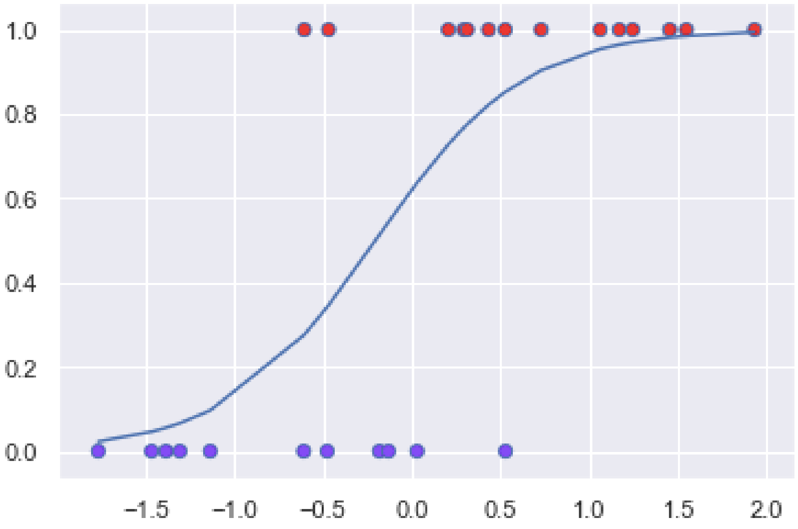


Fig. 28. Logistic regression graph. Adapted from [23].

Figure 28 demonstrates the notion of mapping the datapoints to a sigmoid function, until the lower purple values correspond to the lower left side of the sigmoid with a value of 0, and the red values correspond to the right upper part of the graph with a value of 1. This is done in an iterative process using a concept called “Maximum likelihood” using cost function [22].

$$Sigmoid\ Function\ g(z) = \frac{1}{1+e^{-z}} \tag{9}$$

Equation 9 is the sigmoid function mentioned above, the z is considered a value that we want to map to be between 0 and 1. This equation is crucial as it transforms our datapoints to probabilities and is used as the base of our logistic regression formula.

Understanding cost function is fundamental to comprehend how the model knows it found the best possible solution. The cost function is a mathematical equation that expresses the error in a machine learning model; this happens by continuously comparing the distance between the predicted values produced by the learning model and with the original data. This means that a high-cost function indicates that our model is inaccurate, and a low-cost function indicates that our model is more accurate.

$$Hypothesis\ \ h_{\theta}(x) = \frac{1}{1+e^{(-\theta^{T}x)}} \tag{10}$$

Equation 10 gives the mathematical equation to calculate the prediction of each step, which means the output is between 0 and 1 inclusively. The θ represents the parameters at play, and x is our input variable. The output is then compared with the actual output using the cost function, to minimize the cost function, we will use gradient descent algorithm.

To further understand the cost function, look at figure 29 below, if we had our output Y as 1 and our prediction is also 1 then our cost function is zero, in other words, we will not need to move anymore by gradient descent. However, if our output is 1 and the prediction is 0, then our cost function will be very large as our model needs to move much further from where it is to reach the minimum cost function.

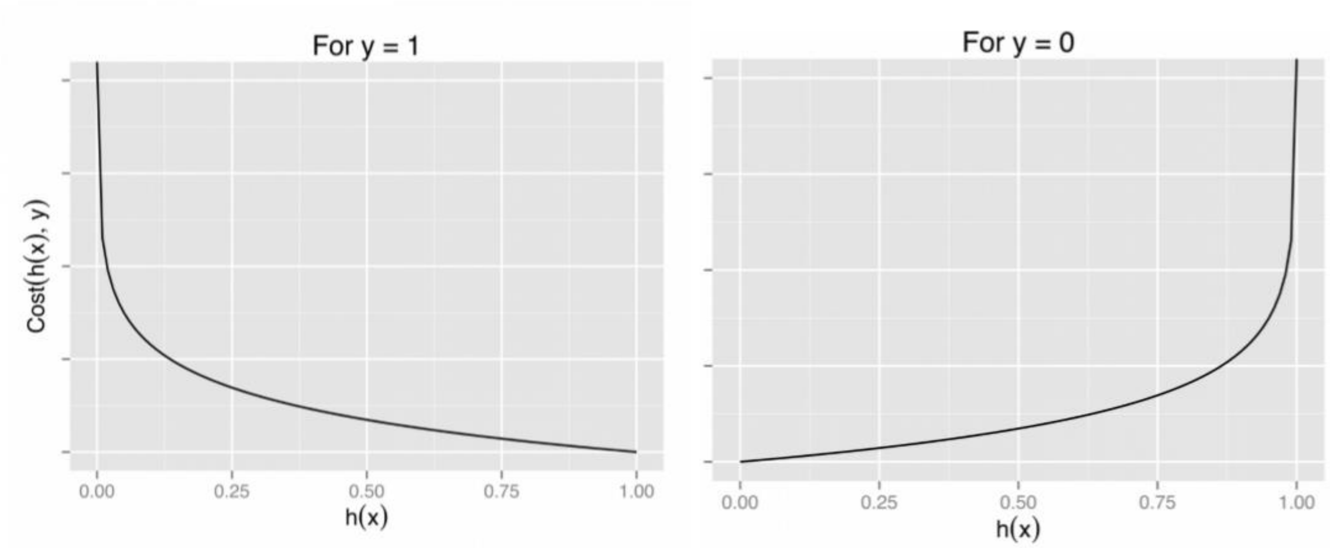


Fig. 29. Graphing hypothesis and cost function according to the output Y. Adapted from [24].

$$P = \frac{1}{1 + e^{-(a+bX)}} \qquad (11)$$

Equation 11 demonstrates the equation of logistic regression. The P is the probability of getting a 1, and a and b are parameters for our model.

2) SVM

A support vector machine (SVM) is a supervised linear machine learning model that solves two-group classification problems. They have two major advantages: speed and performance. This makes the algorithm well-suited to classify whether a patient has heart disease or not. This algorithm works by taking data points and using a formula to output the decision boundary. This is repeated until we maximize the correctly located datapoints [25].

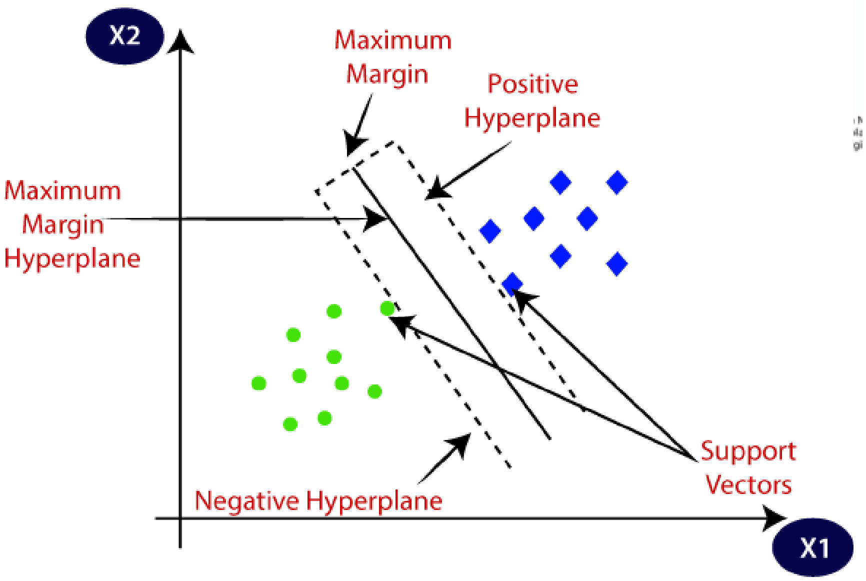


Fig. 30. SVM explanation. Adapted from [26].

Figure 30 hints how the SVM is implemented, the support vectors are the values nearest to the decision boundary. The decision boundary can also be called hyperplane, this is considered as the best boundary for our model.

3) KNN

KNN (K-Nearest Neighbor) is a non-parametric algorithm, which suggests it doesn't make any assumption on underlying data. It assumes the similarity between the new data and available cases and categorizes them accordingly. Then it stores all the available data and classifies a replacement datum supported the similarity. this suggests when new data appears we will derive to the solution by counting the different Ks nearest to the other data and make a decision based on the nearest mass of similar datapoints [27].

At the training phase KNN stores the dataset and when it gets new data, it classifies that data into a category that's almost just like the new data. First, it selects the amount K of the neighbors. Second, it calculates the Euclidean distance of K number of neighbors. Then it takes the K nearest neighbors as per the calculated Euclidean distance. Among these k neighbors, count the amount of the info points in each category. Finally, it assigns the new data points to the category of maximum neighbors [27]. Figure 31 explains this graphically.

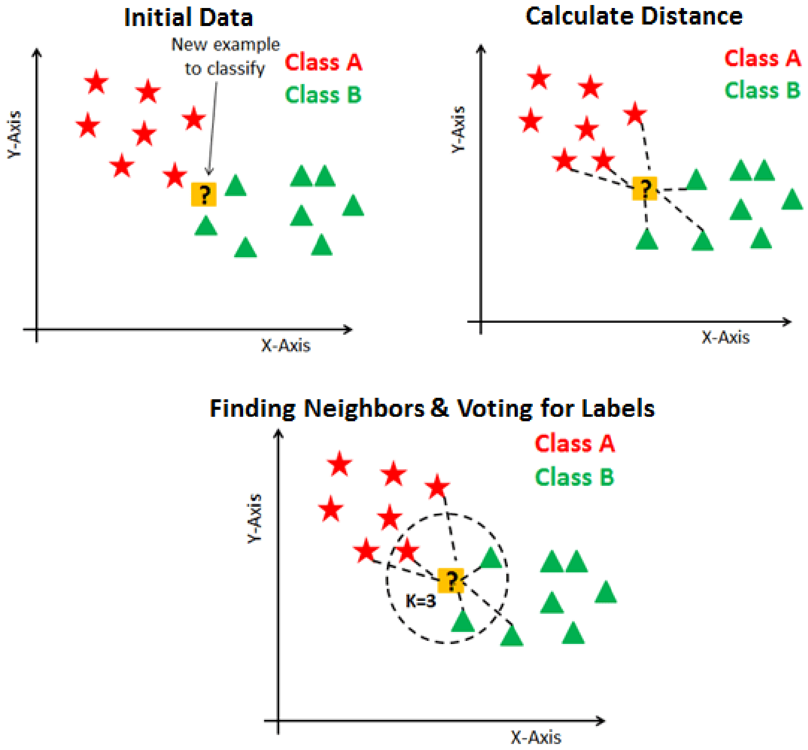


Fig. 31. KNN demonstration. Adapted from [28].

The algorithm works by finding the space between the mathematical values of those points. It computes the space between each datum and therefore the test data then finds the probability of the points being almost like the test data. Classification is predicated on which points share the very best probabilities.

the space function is often Euclidean distance. According to the Euclidean distance formula, the distance between two points in the plane with coordinates (x, y) and (a, b) is given by equation 12 [29].

$$\text{Dist }((x, y), (a, b)) = \sqrt{(x-a)^2 + (y-b)^2} \qquad (12)$$

SECTION IV. Results

Our results can be categorized into finding what features effect heart disease more, and the accuracies of the chosen classification models. After pre-processing and visualizing the data, the following five features turned out to be the most effective on our target and are mainly used in training our model, which are CP, Thal, CA, Exang and oldpeak as shown in figure 28 below. Those results were then crucial to integrate to our learning models, in other word, the model when learning must consider those five features the most, as in practice they are more correlated to the target than the rest of the features.

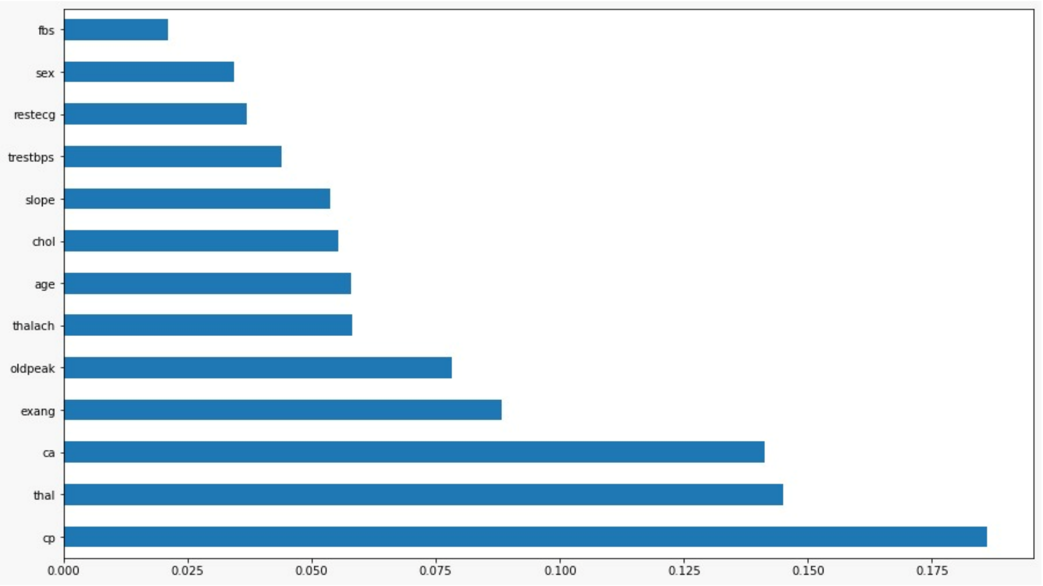


Fig. 32. Feature selection

We have worked using K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and finally, the logistic regression. They all have produced very reasonable accuracies. The least accurate algorithm was the KNN of a prediction percentage of 75.44%, and the SVM gave an accuracy of 80.28%. The highest algorithm in our case was logistic regression with an accuracy equal to 85% this is shown graphically below in figure 33.

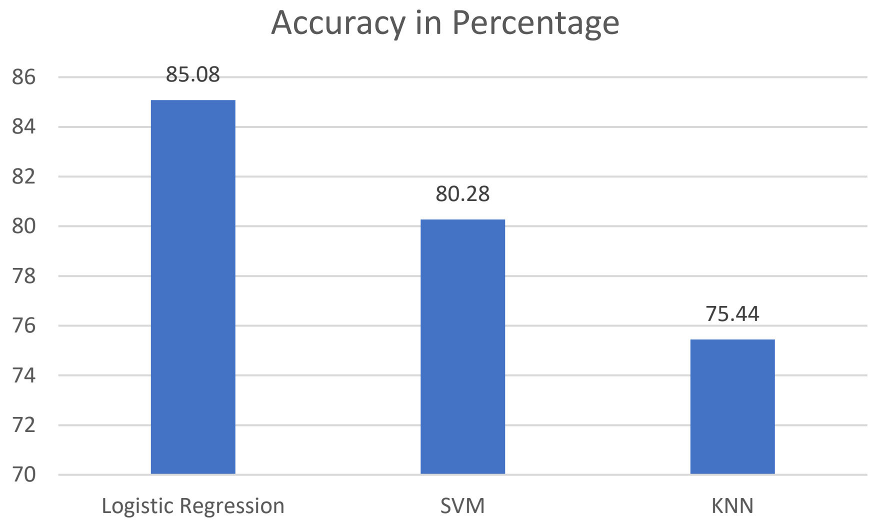


Fig. 33. Accuracy comparison

SECTION V. Conclusion

This paper hopes to have enhanced the readers’ understanding of the parameters that influence heart disease. It is worth noting that after extensive data analysis, we can conclude that the chest pain of the type atypical angina ,resting blood pressure, thal and ca are the most correlated to cardiovascular disease. To add, unlike popular beliefs, age did not directly correlate to having heart disease. This implies that, it is our responsibility to take care of our body and stop blaming age and external factors on our fate, and embark a journey of wellness that includes eating healthy and exercising.

In our paper, we suggested three classification models SVM, KNN, logistic regression. We found that machine learning algorithms are highly sophisticated mathematically, however they all followed a simple algorithm, try, compare, and improve. To clarify, those algorithms used a method to try different formulas, then it compares the values predicted with the actual output. Finally, it improves the equation by moving towards the solution.

For our results, the least accurate algorithm was the KNN with an accuracy of 75%, and then the SVM gave an accuracy of around 80%. Finally, the best algorithm -in our case- was logistic regression with an accuracy equal to 85%..

We are looking forward to integrating heart-disease-trained machine learning and deep learning models with multimedia to be utilized by patients and physicians. Also, we aim to collect more data to be able to improve the accuracy even more.

Resources

[1] "Cardiovascular diseases (CVDs)," *Who.int*, 01-Jun-2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).[Accessed: 29-Dec-2021].

[2] "University diagnostic medical imaging," University Diagnostic Medical Imaging, 14-Feb-2020. [Online]. Available: https://www.udmi.net/cardiovascular-disease-risk/.[Accessed: 29-Dec-2021].

[3] J. L. Marx and G. B. Kolata, Eds., Epidemiology of Heart Disease: Searches for Causes, vol. 197, no. 4264. American Association for the Advancement of Science, 1976. [Online]. Available: https://www.jstor.org/stable/1742465

[4] D. Lundell, "World renown heart surgeon speaks out on what really causes heart disease," Grc.com, 2012. [Online]. Available: https://www.grc.com/health/pdf/Dr._Dwight_Lundell.pdf.[Accessed: 29-Dec-2021].

[5] "Arteriosclerosis," Mayo Clinic, 16-Mar-2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/arteriosclerosis-atherosclerosis/symptoms-causes/syc-20350569.[Accessed: 29-Dec-2021].

[6] "Heart disease," Mayo Clinic, 09-Feb-2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118. [Accessed: 29-Dec-2021].

[7] R. Hajar, "Risk factors for coronary artery disease: Historical perspectives," Risk factors for coronary artery disease: Historical perspectives, vol. 18, no. 3, pp. 109–114, Nov. 2017. [Online]. Available: https://www.heartviews.org/article.asp?issn=1995-705X;year=2017;volume=18;issue=3;spage=109;epage=114;aulast=Hajar

[8] M. K. Nayak, "Types of cardiovascular diseases (CVDs), symptoms and risk factors," Researchgate.net, Aug-2015. [Online]. Available: https://www.researchgate.net/figure/Types-of-cardiovascular-diseases-CVDs-symptoms-and-risk-factors_tbl1_281082129. [Accessed: 29-Dec-2021].

[9] K. Montazeri, C. Unitt, J. M. Foody, J. R. Harris, A. H. Partridge, and J. Moslehi, "ABCDE steps to prevent heart disease in breast cancer survivors," Circulation, vol. 130, no. 18, pp. e157-9, 2014. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.114.008820

[10] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” Mob. Inf. Syst., vol. 2018, pp. 1–21, 2018. [Online]. Available: https://www.hindawi.com/journals/misy/2018/3860146/

[11] L. Chandrika and K. Madhavi, “A hybrid framework for heart disease prediction using machine learning Algorithms,” E3S Web Conf., vol. 309, p. 01043, 2021. [Online]. Available: https://www.e3s-conferences.org

[12] Ronit, “Heart Disease UCI.”, Kaggle. [Online].Available: https://www.kaggle.com/ronitf/heart-disease-uci [Accessed: 29-Dec-2021].

[13] “Angina,” nhs.uk, 22-Apr-2021. [Online]. Available: https://www.nhs.uk/conditions/angina/. [Accessed: 29-Dec-2021].

[14] J. Fletcher, “Cholesterol levels by age: Health ranges, what is high, and tips,” Medicalnewstoday.com, 23-Dec-2021. [Online]. Available: https://www.medicalnewstoday.com/articles/315900. [Accessed: 29-Dec-2021].

[15] CDC, “Diabetes tests,” Centers for Disease Control and Prevention, 11-Aug-2021. [Online]. Available: https://www.cdc.gov/diabetes/basics/getting-tested.html. [Accessed: 29-Dec-2021].

[16] “Cardiovascular Glossary A-Z (All),” Texas Heart Institute, 13-Feb-2018. [Online]. Available: https://www.texasheart.org/heart-health/heart-information-center/topics/a-z/. [Accessed: 29-Dec-2021].

[17] IBM Cloud Education, “What is Exploratory Data Analysis?,” Ibm.com. [Online]. Available: https://www.ibm.com/cloud/learn/exploratory-data-analysis. [Accessed: 29-Dec-2021].

[18] “Statistics and probability,” *Khan Academy*. [Online]. Available: https://www.khanacademy.org/math/statistics-probability. [Accessed: 30-Dec-2021].

[19] Data Science Team, “What is a correlation matrix?,” *DATA SCIENCE,* 28-Dec-2019. [Online]. Available: https://datascience.eu/mathematics-statistics/what-is-a-correlation-matrix/. [Accessed: 30-Dec-2021].

[20] “Correlation coefficient: Simple definition, formula, easy steps,” Statistics How To, 24-May-2021. [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/. [Accessed: 30-Dec-2021].

[21] “The heart attack gender gap,” Harvard Health, 12-Apr-2016. [Online]. Available: https://www.health.harvard.edu/heart-health/the-heart-attack-gender-gap. [Accessed: 05-Jan-2022].

[22] “Logistic regression: Loss and regularization,” *Google Developers*. [Online]. Available: https://developers.google.com/machine-learning/crash-course/logistic-regression/model-training. [Accessed: 06-Jan-2022].

[23] C. Maklin, “Logistic Regression in Python,” Towards Data Science, 03-Aug-2019. [Online]. Available: https://towardsdatascience.com/logistic-regression-python-7c451928efee. [Accessed: 06-Jan-2022].

[24] A. Pant, “Introduction to Logistic Regression,” *Towards Data Science*, 22-Jan-2019. [Online]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148. [Accessed: 06-Jan-2022].

[25] “Support vector machines (SVM) algorithm explained,” MonkeyLearn Blog, 22-Jun-2017. [Online]. Available: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/. [Accessed: 06-Jan-2022].

[26] “Support Vector Machine (SVM) Algorithm - javatpoint,” www.javatpoint.com. [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. [Accessed: 06-Jan-2022].

[27] “K-Nearest Neighbor(KNN) Algorithm for Machine Learning,” www.javatpoint.com. [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning. [Accessed: 06-Jan-2022].

[29] “KNN Classification,” *Datacamp.com*. [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. [Accessed: 06-Jan-2022].

[29] G. Blokdyk, *IBM docs: Complete self-assessment guide*. North Charleston, SC: Createspace Independent Publishing Platform, 2018.

APPENDIX

OUR TEAM’S CODE

The following hyperlinks includes our code, hover over to access the PDF version