Statistical Inference Final Project

Prediction of Heart Disease

Ali Khoramfar
Department of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
Email: khoramfar@ut.ac.ir

Abstract— **In the pursuit of advancing cardiovascular health outcomes, this project endeavors to bridge the gap between statistical inference and predictive analytics in heart disease diagnosis. Drawing upon the methodologies outlined in a seminal article on heart disease prediction using machine learning and deep learning, we aim to augment the scientific discourse with a rigorous statistical analysis. Through the application of visualization techniques, parametric inference, hypothesis testing, and regression analysis on the dataset utilized in the study, this project seeks to not only validate the original findings but also to uncover new insights that could refine predictive models. Our scientific inquiry is rooted in a critical examination of the data and methods, with an emphasis on enhancing the accuracy and efficacy of heart disease prediction. This endeavor not only highlights the importance of interdisciplinary approaches in medical research but also underscores the potential of statistical techniques in unlocking deeper understandings of complex health data.**

Keywords: statistical inference, heart disease prediction, machine learning.

I. INTRODUCTION

The intersection of statistical inference and healthcare has ushered in a transformative era in disease prediction and management, particularly in the realm of heart disease—the leading cause of death globally. This project sets out to explore the intricate dynamics between statistical methodologies and the prediction of heart disease, leveraging the dataset and findings from a seminal article that employs machine learning and deep learning approaches. The introduction of these computational techniques has significantly elevated the predictive accuracy, offering profound implications for early detection and preventive healthcare strategies. Our aim is to extend the existing research framework by applying a series of statistical inference techniques, including visualization, hypothesis testing, and regression analysis, to not only corroborate the article's conclusions but also to potentially unearth new insights that could further refine predictive models. This initiative embodies the synthesis of traditional statistical approaches with cutting-edge computational methodologies, symbolizing a step forward in the quest to mitigate the impact of heart disease through enhanced predictive analytics.

The article [1] presents a comprehensive analysis of heart disease prediction through a combination of machine learning and deep learning techniques. It emphasizes the significance of accurate prediction models in healthcare, detailing the use of various algorithms to analyze heart disease data. The study showcases the potential of integrating statistical and computational methods to enhance diagnostic processes, aiming to improve early detection and treatment strategies. This work stands as a testament to the power of artificial intelligence in revolutionizing the approach towards managing and understanding heart diseases, marking a significant contribution to medical informatics.

The article [1] delves into the alarming global prevalence of heart disease, attributing 17.9 million annual deaths to cardiovascular conditions, exacerbated by unhealthy lifestyles [2]. It discusses the American Heart Association's identification of symptoms often mistaken for aging, highlighting the diagnostic challenges [3]. Emphasizing the pivotal role of machine learning and artificial intelligence in healthcare, the article reviews various studies utilizing machine learning models for heart disease prediction, showcasing significant advancements in diagnostic accuracy. It addresses the challenges of high data dimensionality [4], advocating for feature engineering and selection to improve model performance [5-7]. Highlighting neural networks and dimensionality reduction techniques like PCA for enhanced predictive accuracy, the article underscores the importance of computational methods in early detection and management of heart disease, reflecting on the transformative impact of AI and machine learning in medical diagnostics.

II. METHODOLOGY

A. *Description of the Dataset.*

The researchers used the "Public Health Dataset" created in 1988, which combines data from four sources (Cleveland, Hungary, Switzerland, and Long Beach V). It has 76 different pieces of information about each patient, but most studies only use 14 specific ones. One important piece of information is the "target" field, which shows whether the patient has heart disease (0) or not (1). This section will now explain each of the 14 features used in the research and what they represent.

age: Age of the individual

sex: Sex of the individual (1 = male; 0 = female)

cp: Chest pain type (Value 0: asymptomatic; 1: atypical angina; 2: non-anginal pain; 3: typical angina)

trestbps: Resting blood pressure (in mm Hg on admission to the hospital)

chol: Serum cholesterol in mg/dl

fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

restecg: Resting electrocardiographic results (Values 0, 1, 2)

thalach: Maximum heart rate achieved

exang: Exercise-induced angina (1 = yes; 0 = no)

oldpeak: ST depression induced by exercise relative to rest

slope: The slope of the peak exercise ST segment (Values 0, 1, 2)

ca: Number of major vessels (0-3) colored by fluoroscopy

thal: Thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

target: Heart disease (1 = disease; 0 = no disease)

Categorical Variables: sex, cp, fbs, restecg, exang, slope, ca, thal, target

Numerical Variables: age, trestbps, chol, thalach, oldpeak

B. Preprocessing of the Dataset

The raw dataset had no missing values but suffered from outliers and an imbalanced class distribution. Applying data directly to models yielded poor results.

The authors meticulously preprocessed the dataset to address inherent issues and enhance the performance of their heart disease prediction model. While the dataset lacked missing values, two primary challenges were encountered:

Outliers and Imbalanced Distribution: The presence of outliers and a significant class imbalance (54.46% with heart disease) could mislead the model and hinder accurate predictions.

Skewed Distributions: Further analysis revealed non-uniform distributions in specific features ("thal" and "fasting blood sugar"), indicating potential overfitting or underfitting risks.

To address these concerns, they implemented a multi-pronged approach. Preprocessing steps included:

1) Outlier Detection and Removal:

The authors employed the Isolation Forest algorithm to effectively identify and eliminate outliers. Additionally, normalizing the dataset further mitigated overfitting concerns.

2) Distribution Balancing:

To counteract the class imbalance, they employed appropriate balancing techniques to ensure the model treats both classes fairly.

3) Skewness Correction:

Based on the distribution plots, the authors identified and tackled features exhibiting skewness ("thal" and "fasting blood sugar") using suitable transformation techniques.

4) Duplicate Value Detection and Removal:

They meticulously screened for duplicate values, which could inflate the model's confidence by causing test data to overlap with training data. These duplicates were carefully removed.

After applying the preprocessing steps we could execute, here's a summary of the dataset's transformation:

Original Dataset Size: 1025 entries

After Outlier Removal: Reduced to 976 entries

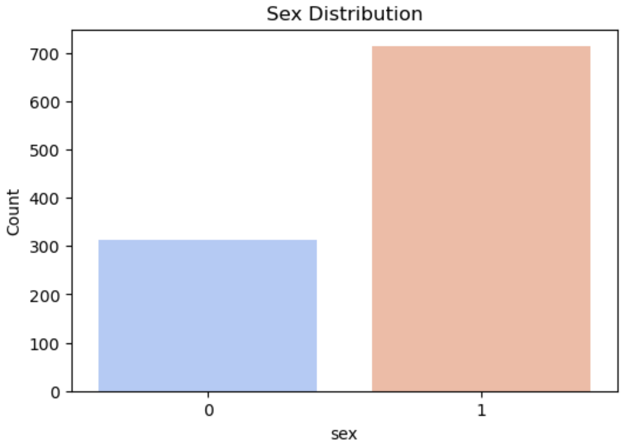
After Duplicate Removal: 288 unique entries

This significant reduction, especially after duplicate removal, suggests that there were many duplicate entries in the dataset.

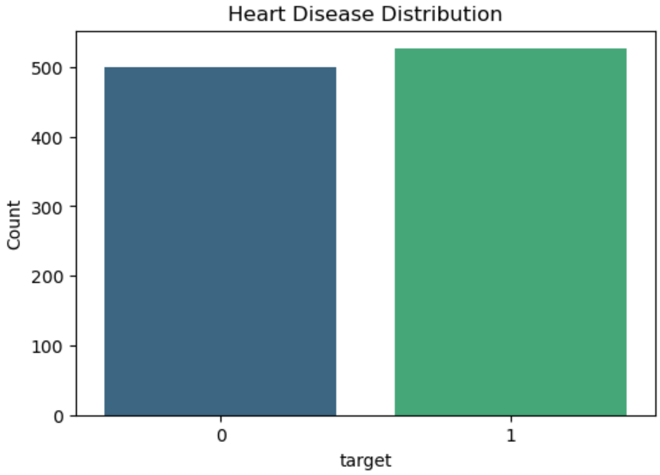
III. VISUALIZATION AND SUMMARIZATION

A. Visualization

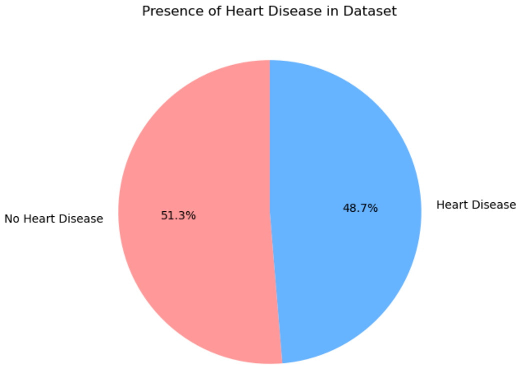
The dataset consists of several variables associated with heart disease, including both categorical and numerical types. We will visualize these variables individually using appropriate plots for each type:



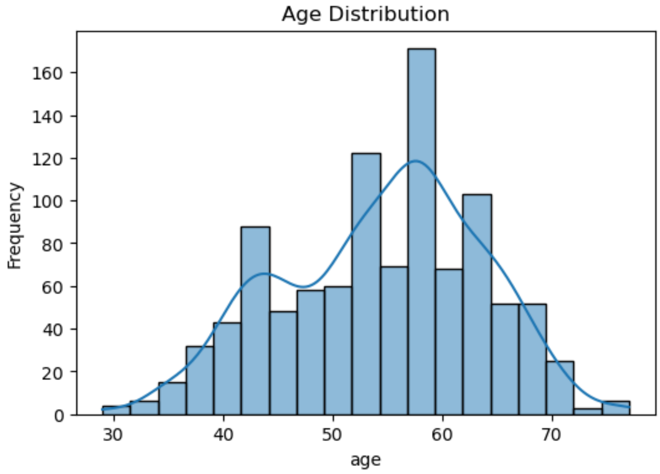
The bar chart above shows the distribution of the sex variable in the dataset, with 1 representing males and 0 representing females. There's a higher number of males compared to females. Heart disease happens more in males than females,which can be read further from Harvard Health Publishing.[8]



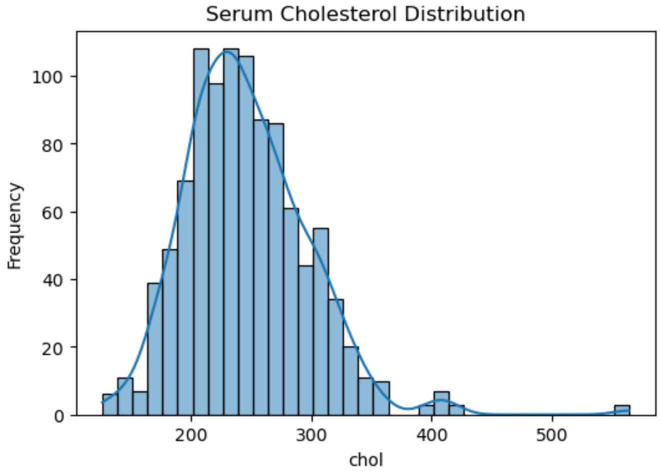
The bar chart for the target variable illustrates the distribution of heart disease within the dataset, showing a relatively balanced presence of individuals with and without heart disease.



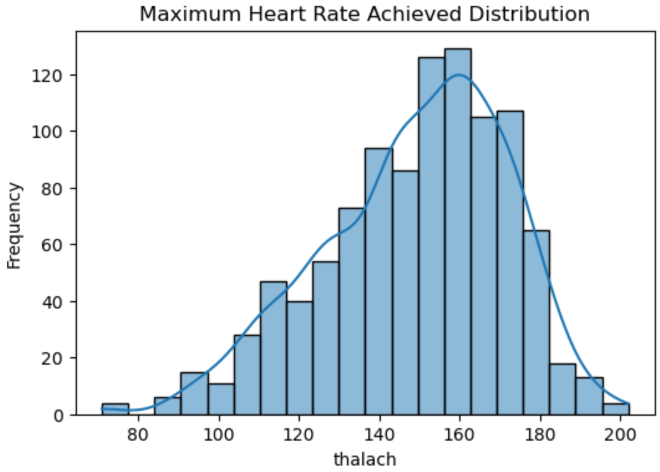
In this case, we can use the pie chart for a better review.



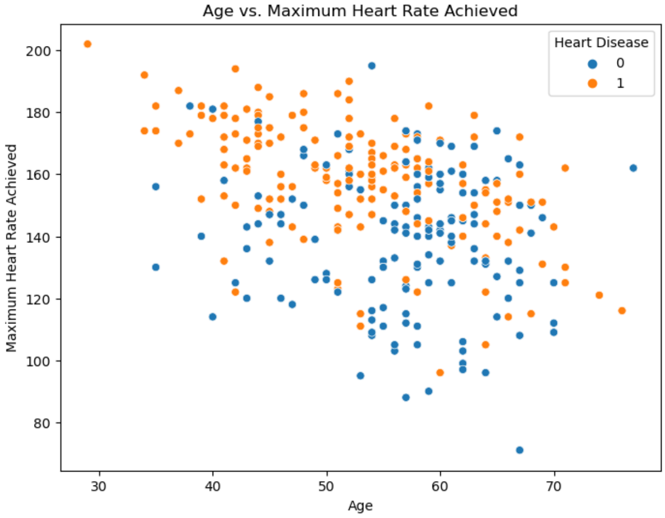
The histogram for age shows the distribution of ages within the dataset, indicating a fairly normal distribution with a slight skew towards older ages. The kernel density estimate (KDE) overlay provides a smooth representation of the distribution.



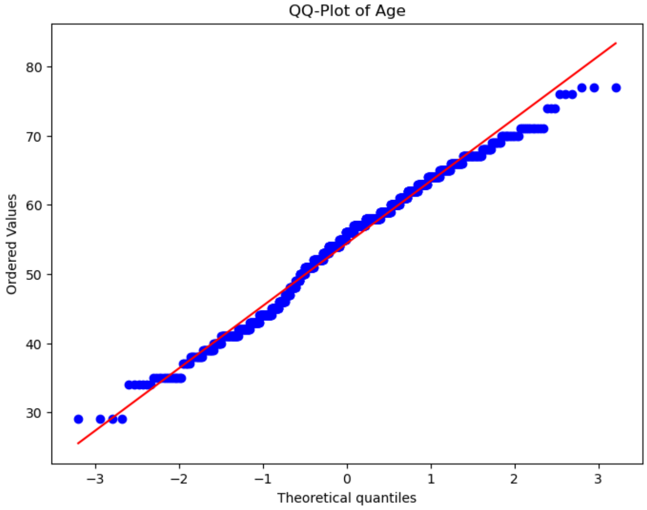
The histogram for chol (serum cholesterol) depicts its distribution within the dataset, displaying a normal distribution with a right skew. This skewness indicates that some individuals have significantly higher cholesterol levels.



The histogram for thalach (maximum heart rate achieved) reveals a distribution that is slightly left-skewed, with most individuals achieving a high maximum heart rate.

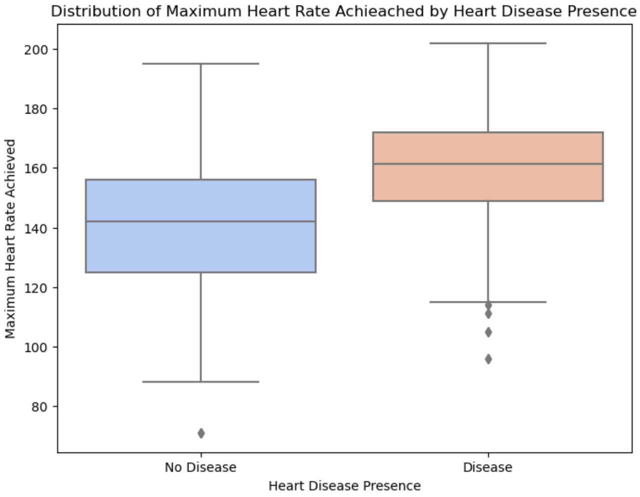


The scatter plot of age vs maximum heart rate achieved with heart disease reveals a potential relationship between these variables and the presence of heart disease. As age increases, the maximum heart rate achieved tends to decrease, and there's a



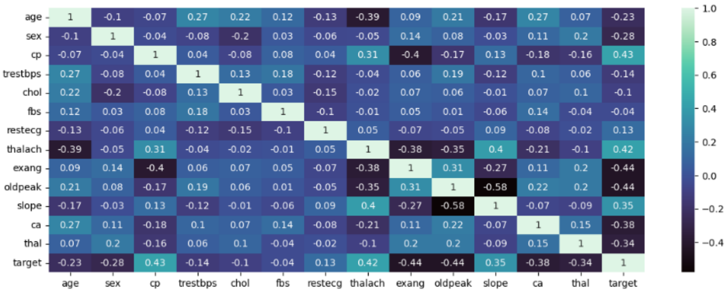
visible distribution difference between individuals with and without heart disease.

The QQ-plot of age is used to check the normality of the age distribution against a theoretical normal distribution. The points follow the line closely in the middle range but deviate at the ends, suggesting that while age is approximately normally distributed, it has some deviations from normality, especially in the tails.



This boxplot that visualizes the distribution of maximum heart rates between individuals with and without heart disease, aiding in the exploratory data analysis phase. This visualization would help identify if there are significant differences in the maximum heart rate achieved between individuals with and without heart disease, providing insights into the potential role of thalach in predicting heart disease.

another result of this part is a diagram from the heatmap created by plotting the correlation matrix.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.1	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
sex	-0.1	1	-0.04	-0.08	-0.2	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.2	-0.28
cp	-0.07	-0.04	1	0.04	-0.08	0.08	0.04	0.31	-0.4	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.1	0.06	-0.14
chol	0.22	-0.2	-0.08	0.13	1	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.1	-0.1
fbs	0.12	0.03	0.08	0.18	0.03	1	-0.1	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.1	1	0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05	1	-0.38	-0.35	0.4	-0.21	-0.1	0.42
exang	0.09	0.14	-0.4	0.06	0.07	0.05	-0.07	-0.38	1	0.31	-0.27	0.11	0.2	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1	-0.58	0.22	0.2	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.4	-0.27	-0.58	1	-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.1	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1	0.15	-0.38
thal	0.07	0.2	-0.16	0.06	0.1	-0.04	-0.02	-0.1	0.2	0.2	-0.09	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.1	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1

The heatmap displays correlation coefficients between pairs of variables, ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other tends to increase as well. A value close to -1 indicates a strong negative correlation, meaning that as one variable increases, the other tends to decrease. A value around 0 suggests no linear correlation between the variables.

B. *Identify dependent and independent factors*

1) *Age vs. Maximum Heart Rate:*

The scatter plot showed a noticeable trend where the maximum heart rate achieved decreases as age increases. This suggests a potential dependency between age and maximum heart rate, which could be an important factor in predicting heart disease. Additionally, the presence of heart disease seems to influence this relationship, indicating that age and maximum heart rate achieved could be dependent factors in the context of heart disease prediction.

2) *Normality of Age:*

The QQ-plot for age indicated that while age is approximately normally distributed, there are deviations, especially at the tails. This suggests that age, as a factor, has its peculiarities in distribution that could affect its role in predictive models.

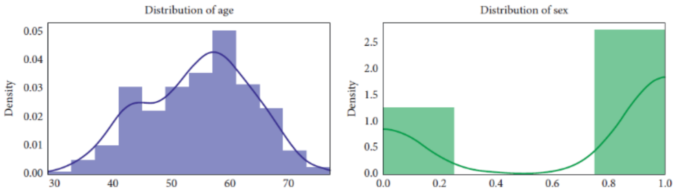
3) *Variable Relationships and Heart Disease:*

The scatter plot color-coded by heart disease presence highlighted that both age and maximum heart rate achieved are not only dependent on each other but also potentially influential in the context of heart disease. This interdependence and their relationship with the target variable suggest they are critical factors to consider in models predicting heart disease.

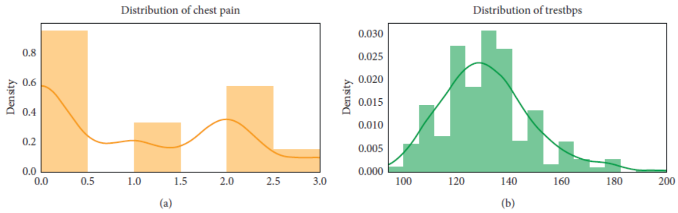
C. *Visualizations provided in the publication*

The publication [1] includes detailed visualizations to illustrate the distribution of heart disease within the dataset and the impact of various features on heart disease prediction. Specifically, it discusses class distributions, the importance of features related to heart disease, and the normalization of data to address overfitting. The visualizations provided, such as bar charts and density plots, help in understanding the dataset's characteristics and the preprocessing steps undertaken to improve model performance. Through these graphical representations, the authors highlight the skewness in certain features and the balanced approach taken to enhance predictive accuracy. Some of them are as follows:

Distribution of age and sex.



Distribution of chest pain and trestbps



The publication highlights visualizations that categorize features based on their importance for predicting heart disease. Features deemed important for heart disease prediction are emphasized through their impact on model accuracy and decision-making processes in machine learning algorithms. Conversely, features not important for heart disease are identified as having minimal to no impact on the models' predictive capabilities. These distinctions help in refining the models by focusing on significant predictors and excluding less relevant information, ultimately aiming to enhance the accuracy and efficiency of heart disease prediction.

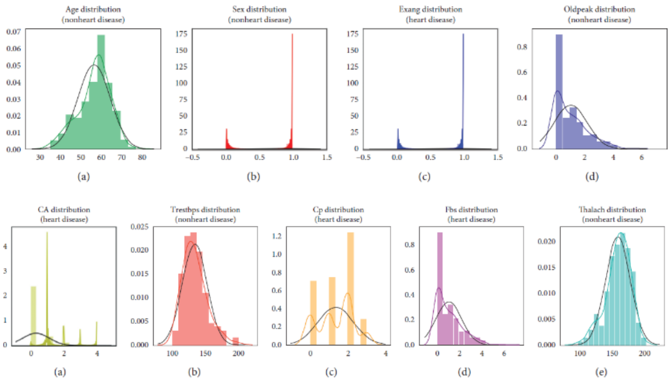


FIGURE 5: Features not important for heart disease.

D. Kolmogorov Smirnov Test

The KS test is a non-parametric test that compares a sample with a reference probability distribution or compares two samples. In this context, we're interested in testing if the age distribution in the dataset follows a normal distribution.

The test compares the empirical distribution function of the age variable with the cumulative distribution function (CDF) of the normal distribution.

The normal distribution parameters (mean and standard deviation) are derived from the age data itself.

The **KS Statistic** of **0.0796** indicates the maximum distance between the empirical distribution function of the age data and the cumulative distribution function (CDF) of a normal distribution with the same mean and standard deviation as the age data. This value by itself tells us there is some deviation, but it's the p-value that helps us determine the significance of this deviation.

The **P-value** is significantly low$(4.22 * 10^{-6})$ far below the common significance level of 0.05 used to assess statistical significance. The very low p-value suggests that we can reject the null hypothesis that the age distribution in the dataset follows a normal distribution. This means there is statistically significant evidence that the age data deviates from normality.

IV. PARAMETRIC INFERENCE AND ESTIMATION

A. Parametric tests

To conduct parametric inference methods on the dataset, we typically consider tests like the t-test for comparing means between two groups or ANOVA for comparing means across more than two groups, provided certain assumptions are met. These assumptions include normality of data distributions.

1) Shapiro–Wilk test for checking normality

The Shapiro-Wilk test is used to assess the normality of a dataset, which is a common assumption of many parametric statistical tests. Applying the Shapiro-Wilk test to the chol variable will help us determine whether the serum cholesterol levels in dataset are normally distributed.

The Shapiro-Wilk test on the chol variable yielded a test statistic of approximately 0.950. The extremely small p-value suggests that we can reject the null hypothesis of normality for the chol dataset. Then we should do a transformation to apply parametric test on this data. To transform the chol (serum cholesterol) variable to more closely approximate a normal distribution, several methods can be used. The choice of method depends on the nature of the data's deviation from normality. Common transformations include:

Log Transformation: Useful for data with right-skewed distributions.

Square Root Transformation: Can be effective for moderate skewness.

Inverse Transformation: Useful for severe skewness, but less commonly used.

Box-Cox Transformation: A more generalized approach that can handle various types of skewness.

The Box-Cox transformation made the chol data more normal, as evidenced by the increase in the Shapiro-Wilk test statistic towards 1.

2) T-test : Cholesterol level and heart disease

The results of the t-test would provide a p-value indicating whether there's a statistically significant difference in the mean maximum Cholesterol level between individuals with and without heart disease. A low p-value (typically <0.05) would suggest significant differences.

Null Hypothesis (H0): There is no difference in mean cholesterol levels between individuals with heart disease and those without.

Alternative Hypothesis (H1): There is a difference in mean cholesterol levels between individuals with heart disease and those without.

The results of the independent t-test show a **test statistic** of approximately **-3.22** and a **p-value** of approximately **0.0013**.

The p-value is less than the commonly used significance level of 0.05, which suggests that we have enough evidence to reject the null hypothesis.

Based on the t-test, we reject the null hypothesis and conclude that there is a statistically significant difference in mean cholesterol levels between individuals with heart disease and those without in the provided dataset. This indicates that cholesterol levels might be associated with the presence of heart disease, supporting the alternative hypothesis.

3) Jarque-Bera test for checking normality

The Jarque-Bera test is a statistical test that checks whether sample data have the skewness and kurtosis matching a normal distribution. It's particularly useful for large samples, as its power increases with the sample size. The test's null hypothesis is that the data are normally distributed.

To apply the Jarque-Bera test to the thalach (maximum heart rate achieved) variable from your dataset, we'll use the jarque_bera function from scipy.stats. This test returns a test statistic and a p-value, where a small p-value suggests that the null hypothesis of normality can be rejected.

Test Statistic: The value of 45.34 reflects the deviation of the thalach data from a normal distribution, considering its skewness and kurtosis. The very small p-value 1.426×10^{10} suggests that we can reject the null hypothesis of normality for the thalach dataset. So we should do transformation to dataset before apply parametric test.

4) T-test : Maximum heart rate and heart disease

A potential hypothesis could involve comparing the maximum heart rate achieved between individuals with and without heart disease.

Null Hypothesis (H0): There is no difference in mean maximum heart rate achieved (thalach) between individuals with heart disease and those without.

Alternative Hypothesis (H1): There is a difference in mean maximum heart rate achieved between individuals with heart disease and those without.

The results of the independent t-test for the maximum heart rate achieved (thalach) show a **test statistic** of approximately **14.86** and a **p-value** of approximately **3.42e-45**.

The p-value is significantly less than the commonly used significance level of 0.05, which strongly suggests that we have enough evidence to reject the null hypothesis.

Based on the t-test, we reject the null hypothesis and conclude that there is a statistically significant difference in the mean maximum heart rate achieved between individuals with heart disease and those without in the provided dataset.

5) F-Test: Resting blood pressure and heart disease

Given the previous discussions and the dataset at hand, we'll continue with the idea of comparing a clinically relevant measure. We'll use the resting blood pressure (trestbps) for this variance test, comparing individuals with heart disease to those without.

Null Hypothesis (H0): The variance of resting blood pressure is equal between individuals with heart disease and those without.

Alternative Hypothesis (H1): The variance of resting blood pressure is not equal between individuals with heart disease and those without.

This approach typically relies on an F-test for comparing variances directly.

The sample size for individuals with heart disease is **526**.

The sample size for individuals without heart disease is **499**.

The variance of resting blood pressure for individuals with heart disease is approximately **259.60**.

The variance of resting blood pressure for individuals without heart disease is approximately **345.10**.

The **F-statistic** (ratio of variances) is approximately **0.752**.

These results indicate that the variance of resting blood pressure in individuals with without disease is less than that of individuals with heart disease, as shown by an F-statistic less than 1.

6) Proportion Test: Exercise-induced angina and heart disease

Exercise-induced angina is chest pain occurring during physical activity due to insufficient heart muscle oxygenation, potentially indicating narrowed or blocked coronary arteries. Exercise-induced angina is a symptom that can indicate underlying heart issues, making it a clinically significant factor to examine.

For the proportion test, we can examine the difference in proportions of a categorical variable between two groups. A relevant and interesting comparison might involve the presence (or absence) of exercise-induced angina (exang) between individuals with and without heart disease.

Null Hypothesis (H0): The proportion of individuals experiencing exercise-induced angina is the same for those with heart disease and those without.

Alternative Hypothesis (H1): The proportion of individuals experiencing exercise-induced angina differs between those with heart disease and those without.

For large values of n we can use the normal approximation to the binomial distribution. The results of the two-proportion z-test show a **test statistic** of approximately **-14.02** and a **p-value** of approximately **1.12e-44**.

The p-value is less than the commonly used significance level of 0.05, which suggests that we have enough evidence to reject the null hypothesis.

Based on the difference of two proportions test, we reject the null hypothesis and conclude that there is a statistically significant difference in the proportions of individuals experiencing exercise-induced angina between those with heart disease and those without in the provided dataset. This result suggests that the experience of exercise-induced angina is associated with the presence of heart disease, supporting the alternative hypothesis.

7) NonParametric: Wilcoxon-Mann-Whitney-Test

Given that the data are not strictly normal even after transformation, we also use non-parametric tests.

The Wilcoxon Mann-Whitney test (also known as the Mann-Whitney U test) is a non-parametric test used to compare whether there is a difference in the median values of two independent groups. Unlike the t-test, it does not assume that the data are normally distributed.

To perform the Wilcoxon Mann-Whitney test on the dataset, we first need to select two independent groups. Given the variables in your dataset, a common choice might be comparing a continuous variable across two groups defined by a categorical variable. For example, we could compare the chol (serum cholesterol) levels between males and females (using the sex variable, where 1 denotes male and 0 denotes female).

The Wilcoxon Mann-Whitney test comparing serum cholesterol (chol) levels between males and females yielded **a U statistic** of approximately **89877.0** and a **p-value** of approximately $9.79 * 10^{-7}$.

The value of the U statistic indicates the rank sum of the differences between the two groups. However, its interpretive value is mainly in relation to the p-value rather than as a standalone number. The very small p-value suggests that we can reject the null hypothesis, indicating a significant difference in the median cholesterol levels between males and females in the dataset.

Based on the Wilcoxon Mann-Whitney test results, there is a statistically significant difference in serum cholesterol levels between males and females in dataset. This suggests that sex may be an important factor to consider in studies or analyses involving cholesterol levels.

B. Estimation of parameters

1) Point estimation: Mean Resting Blood Pressure

To apply estimation techniques to the dataset, we'll focus on point estimation and interval estimation for various variables. Point estimation involves using the data to calculate a single value that serves as a "best guess" or "best estimate" of a population parameter (e.g., population mean).

According to the calculations made in the programming part of the project, the point estimates is as follows:

Mean Resting Blood Pressure (trestbps): The point estimate is approximately **131.61 mmHg**.

2) Confidence Interval for Mean Resting Blood Pressure

The 95% **confidence interval** is approximately **130.54 to 132.68 mmHg**. This interval suggests that we are 95% confident that the true mean resting blood pressure for the population from which this sample was drawn lies within this range.

3) Point estimation: Mean Cholesterol Level

According to the calculations made in the programming part of the project, the point estimates is as follows:

Mean Cholesterol Level (chol): The point estimate is **246.00 mg/dL**.

4) Confidence Interval for Mean Cholesterol Level

The 95% **confidence interval** is approximately **242.84 to 249.16 mg/dL**. This interval indicates that we are 95% confident that the true mean cholesterol level for the population lies within this range.

These point and interval estimates provide a statistical summary of the dataset's key variables, offering both specific estimates and a range of likely values for these parameters within the broader population. The confidence intervals add an important dimension to our understanding by quantifying the uncertainty around these point estimates, allowing for more informed interpretations and decisions.

5) Maximum likelihood estimation

To apply Maximum Likelihood Estimation (MLE) to estimate the parameters (mean and standard deviation) of the chol (serum cholesterol in mg/dl) variable, assuming it follows a normal distribution, we will use the chol data to estimate its distribution parameters.

Given the estimated parameters for the serum cholesterol (chol) variable from dataset using Maximum Likelihood Estimation (MLE), we have:

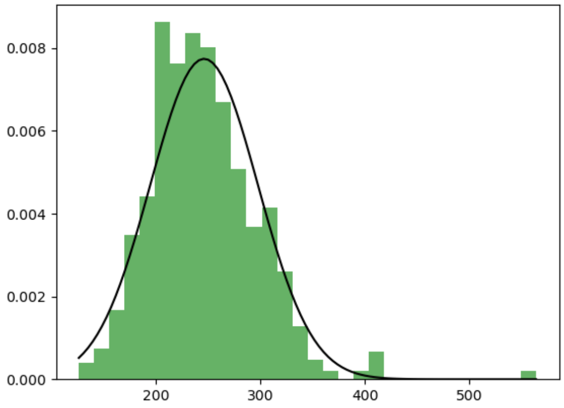
Mean estimate: 246.0 mg/dL

Standard deviation estimate: 51.567 mg/dL

Understanding the distribution of cholesterol levels within a population can have significant clinical implications. For instance, high cholesterol levels are a known risk factor for heart disease. The American Heart Association suggests that adults should aim for a cholesterol level lower than 200 mg/dL. With a mean cholesterol level of 246.0 mg/dL, the population in this dataset, on average, exceeds this threshold, which may indicate a higher risk of heart disease.

The variability (standard deviation) in cholesterol levels also provides insights into the heterogeneity of the population's health status concerning cardiovascular risk factors. A wider range (higher standard deviation) might suggest that while some individuals have cholesterol levels that are significantly higher than the average, others might have lower or near-optimal levels.

To understand how well the estimated parameters fit data, we can plot the histogram of chol data alongside the probability density function (PDF) of a normal distribution with the estimated parameters:



6) Checking for missing or unobserved data points

Interpolation is a method of estimating and constructing new data points within the range of a discrete set of known data points. It's particularly useful in handling missing or unobserved data within datasets. There are several methods of interpolation, including linear interpolation, polynomial interpolation, and spline interpolation, among others. The choice of method often depends on the dataset's characteristics and the specific requirements of your analysis.

The dataset does not have any missing data across all its variables. Each column has zero missing values, indicating that the dataset is complete and no interpolation for missing data is necessary.

V. Hypothesis Testing and Statistical Analysis

A. Hypothesis Testing

1) Age and Maximum Heart Rate Achieved (thalach)

Understanding the relationship between age and maximum heart rate can provide insights into how age affects cardiovascular function.

The Pearson correlation coefficient measures the strength and direction of the linear relationship between two continuous variables. It provides a value between -1 and 1, where 1 means a perfect positive linear correlation, -1 means a perfect negative linear correlation, and 0 indicates no linear correlation.

Hypothesis: There is a significant correlation between age and maximum heart rate achieved.

Method: Using Pearson correlation to test the relationship between age and thalach.

Test Statistic (Correlation Coefficient): -0.390

P-value: $1.27 * 10^{-38}$

The negative correlation coefficient of -0.390 indicates a moderate inverse relationship between age and maximum heart rate achieved, meaning as age increases, the maximum heart rate typically decreases. The extremely low p-value suggests this correlation is statistically significant. So there is significant evidence to conclude that there is a negative correlation between age and maximum heart rate achieved, supporting the hypothesis that age and maximum heart rate are related.

2) Sex and Presence of Heart Disease (target)

Investigating the prevalence of heart disease by sex can help identify whether sex is a significant factor in heart disease.

The Chi-square test is primarily used to examine the relationship between two categorical variables. It helps determine whether there is a significant association between the two variables or if any observed differences in frequencies or proportions could have occurred by chance.

Hypothesis: There is a significant difference in the presence of heart disease between males and females.

Method: Use Chi-square test for independence to compare the proportions of heart disease presence across sexes.

Test Statistic (Chi-square): 78.86

P-value: $6.65 * 10^{-19}$

The Chi-square test for independence yields a test statistic of 78.86 with a p-value significantly less than 0.05, indicating a strong statistical evidence to reject the null hypothesis of independence between sex and the presence of heart disease.

There is significant evidence to conclude that there is a difference in the presence of heart disease between males and females, supporting the hypothesis that sex is a significant factor in the occurrence of heart disease.

3) Chest pain and ST depression

This analysis will reveal whether certain types of chest pain are associated with greater or lesser degrees of ST depression, which could correlate with the severity or nature of underlying heart conditions.

ST depression refers to a specific abnormality seen on an electrocardiogram (ECG), a test that measures the electrical activity of the heart. In an ECG, there are several waves and segments that represent different phases of the heartbeat. We can apply Kruskal-Wallis test to two continuous variables.

Hypothesis: The distribution of oldpeak (ST depression) differs across different types of chest pain (cp).

Method: Kruskal-Wallis Test, to compare the distributions of oldpeak across different cp groups.

This approach aligns better with the intent to use the Kruskal-Wallis Test, as it directly relates a continuous variable (oldpeak) with a categorical one (cp), fitting the test's purpose.

Test Statistic: 155.15

P-value: $2.04 * 10^{-33}$

The Kruskal-Wallis test reveals a significant difference in the distributions of oldpeak (ST depression) among the different chest pain types (cp), with a test statistic of 155.15 and a p-value far below the 0.05 threshold for statistical significance.So there is strong evidence to conclude that the distribution of ST depression (oldpeak) differs significantly across different types of chest pain. This suggests that certain types of chest pain are associated with more severe indications of heart stress or damage, as measured by ST depression during exercise relative to rest.

This finding supports the notion that chest pain type is an important variable in understanding the cardiovascular condition of patients, with implications for diagnosis and treatment strategies related to heart disease.

B. Bootstrap Method for Hypothesis Testing

The bootstrap method is a powerful statistical tool used for estimating the distribution of a statistic (like the mean, median, or proportion) by resampling with replacement from the data. It allows for hypothesis testing, especially useful in cases where the data distribution is unknown, non-normal, or when the sample size is small. Bootstrap methods can provide insights into the variability of the statistic and help construct confidence intervals.

We apply the bootstrap method for hypothesis testing on the dataset to compare the average maximum heart rate achieved (thalach) between two groups defined by the presence of heart disease. This choice is motivated by the clinical relevance of examining how heart disease might affect physical indicators like heart rate.

Null Hypothesis (H0): There is no difference in the mean maximum heart rate achieved between individuals with and without heart disease.

Alternative Hypothesis (HA): There is a difference in the mean maximum heart rate achieved between individuals with and without heart disease.

Observed Difference in Means: The mean maximum heart rate achieved is approximately 19.46 bpm higher in individuals with heart disease compared to those without heart disease.

P-value: 0 based on bootstrap samples

This suggests that there is a statistically significant difference in the mean maximum heart rate achieved between individuals with and without heart disease.

This analysis, employing the bootstrap method, allows us to draw conclusions about the relationship between heart disease and maximum heart rate achieved, even without assuming a normal distribution of the data. It highlights the practical value of bootstrap methods in hypothesis testing, especially for non-normal data distributions or when the theoretical distribution of the test statistic is unknown.

C. Permutation Test

For employing a resampling method for hypothesis testing, especially useful for non-normal data distributions, let's focus on the relationship between oldpeak (ST depression induced by exercise relative to rest) and heart disease presence (target). ST depression is a marker that can indicate stress on the heart, and comparing its levels in individuals with and without heart disease could provide insights into the physiological impacts of heart disease.

The permutation test is a resampling method that involves combining the data from both groups, shuffling (or permuting) the combined data, and then reallocating the shuffled data into two groups to test the hypothesis. This method does not assume normal distribution and is suitable for non-parametric data. This approach will help us test our hypothesis regarding oldpeak across individuals with and without heart disease. Let's conduct the permutation test.

Null Hypothesis (H0): There is no difference in median oldpeak between individuals with and without heart disease.

Alternative Hypothesis (HA): There is a difference in median oldpeak between individuals with and without heart disease.

Observed Difference in Medians: The median oldpeak (ST depression induced by exercise relative to rest) is 1.2 units lower in individuals with heart disease compared to those without heart disease.

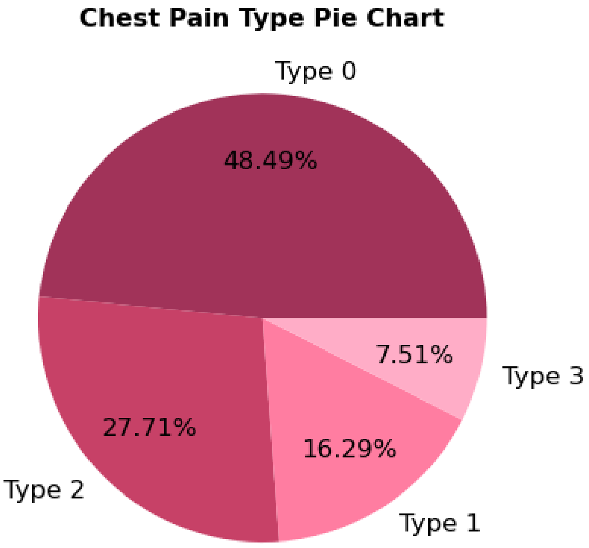
P-value: 0 based on 10,000 resamples

Given a p-value of 0.0, we have strong evidence to reject the null hypothesis in favor of the alternative hypothesis. This suggests that there is a statistically significant difference in the median oldpeak between individuals with and without heart disease.

D. ANOVA

To conduct an Analysis of Variance (ANOVA) test, we need to compare the means of different groups within a continuous variable across different levels of a categorical variable. Since target defines whether heart disease is present or not and is binary, we'll choose another categorical variable with more than two levels for grouping and compare a continuous variable across these groups.

We can choose cp (chest pain type) as the categorical variable for grouping and thalach (maximum heart rate achieved) as the continuous variable to compare across these groups. This will allow us to see if there are significant differences in the maximum heart rate achieved among the different types of chest pain, which could have implications for cardiovascular health and diagnosis.



1) Normality Check

Given the description of thalach as being slightly left-skewed. The Shapiro-Wilk test results for thalach within each chest pain type (cp) group show p-values well below the usual significance level of 0.05 for all groups. These results indicate that the assumption of normality is violated for thalach across all chest pain types. a common approach to normalize the data is applying a transformation.

In this project, various transformations were done, but none of them caused the distribution to become completely normal. So, assuming that the distribution is close to normal, we perform the ANOVA test. But we must keep in mind that the answer is not exact.

2) *Hypothesis for ANOVA*

Null Hypothesis (H0): There are no differences in the mean maximum heart rate achieved (thalach) among the different types of chest pain (cp).

Alternative Hypothesis (HA): At least one type of chest pain has a different mean maximum heart rate achieved compared to the others.

Using the F-test to determine if there are any statistically significant differences in the means of thalach across the cp groups.

F-Statistic: 62.86

P-value: $2.67 * 10^{-37}$

Given the extremely low p-value, we have strong evidence to reject the null hypothesis in favor of the alternative hypothesis. This suggests that there are significant differences in the mean maximum heart rate achieved (thalach) among the different types of chest pain (cp). At least one type of chest pain is associated with a distinct mean maximum heart rate when compared to the others.

3) *Bonferroni Method*

When you conduct ANOVA and find a significant result, it indicates that there are differences among the group means, but it doesn't specify which groups differ from each other. To identify the specific group differences, post hoc tests such as Tukey's HSD (Honestly Significant Difference) or pairwise comparisons with corrections for multiple testing, such as the Bonferroni correction, are used.

The Bonferroni correction is a conservative method that adjusts the significance level by dividing it by the number of comparisons being made, to control the family-wise error rate.

The Bonferroni correction was applied to the pairwise comparisons of thalach across different chest pain types (cp), and the results are as follows (with p-values adjusted for multiple testing):

Comparison cp 0 and cp 1: **Significant** $p = 4.18 * 10^{-26}$

Comparison cp 0 and cp 2: **Significant** $p = 5.42 * 10^{-19}$

Comparison cp 0 and cp 3: **Significant** $p = 7.72 * 10^{-9}$

Comparison cp 1 and cp 2: **Significant** $p = 0.0011$

Comparison cp 1 and cp 3: **Not significant** $p = 0.386$

Comparison cp 2 and cp 3: **Not significant** $p = 1$

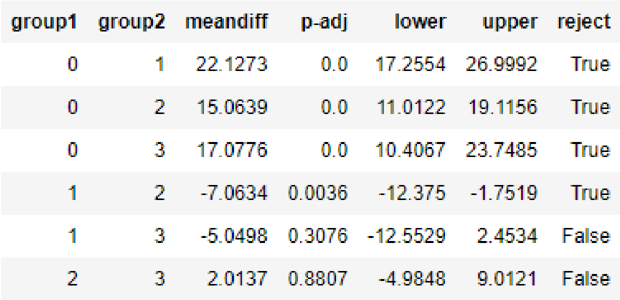
The Bonferroni correction for multiple comparisons has identified significant differences in the maximum heart rate achieved (thalach) between several pairs of chest pain types (cp), with the exception of comparisons between cp 1 and cp 3, and cp 2 and cp 3, which did not show significant differences after adjusting for multiple comparisons.

This detailed analysis, following the significant ANOVA result, provides a more nuanced understanding of how chest pain type might relate to heart function.

4) *Tukey's Method*

To apply Tukey's test for multiple comparisons following the ANOVA, we'll focus again on the thalach variable across different levels of chest pain (cp). Tukey's test is particularly useful after finding a significant ANOVA result because it helps identify which specific group means are significantly different from each other without increasing the Type I error rate associated with multiple comparisons.

The results from Tukey's HSD test for comparing the mean maximum heart rate achieved (thalach) across different chest pain types (cp) are as follows:



group1	group2	meandiff	p-adj	lower	upper	reject
0	1	22.1273	0.0	17.2554	26.9992	True
0	2	15.0639	0.0	11.0122	19.1156	True
0	3	17.0776	0.0	10.4067	23.7485	True
1	2	-7.0634	0.0036	-12.375	-1.7519	True
1	3	-5.0498	0.3076	-12.5529	2.4534	False
2	3	2.0137	0.8807	-4.9848	9.0121	False

The results of this method are the same as the previous method and Significant differences were found between most pairs of chest pain types, except between cp 2 and cp 3, and cp 2 and cp 3 where the difference was not statistically significant.

This analysis highlights the variability in cardiovascular stress response associated with different chest pain presentations, which could have implications for clinical assessment and understanding the pathophysiology of heart disease. The results offer detailed insights into the differences between chest pain types, supporting nuanced interpretations of heart disease symptoms and their implications for patient care

VI. *Regression Analysis*

A. *Simple Linear Regression*

For a simple linear regression analysis, we aim to investigate the relationship between two variables. A common choice in the context of heart disease could be analyzing how a continuous predictor variable, such as age, trestbps (resting blood pressure), chol (cholesterol level), or thalach (maximum heart rate achieved), affects the likelihood of heart disease (target).

However, since target is a binary outcome variable, a logistic regression would be more appropriate for directly predicting the likelihood of heart disease. In this part we explore the relationship between age and thalach (maximum heart rate achieved), hypothesizing that maximum heart rate achievable decreases with age. This is a simple linear regression model where age is the predictor variable and thalach is the outcome variable.

The simple linear regression analysis between age and thalach (maximum heart rate achieved) resulted in the following linear equation:

$$thalach = 202.98 - 0.99 \times age$$

Intercept = **202.98**: This value represents the expected maximum heart rate for someone of age 0.

Slope = -0.99: This indicates that for each additional year of age, the maximum heart rate decreases by approximately 0.99 beats per minute (bpm). This negative slope confirms our hypothesis that maximum heart rate achievable decreases with age.

R-squared = **0.1523**: This value explains the proportion of the variance in the maximum heart rate that is predictable from the age of individuals. With a value of approximately 0.15, it suggests that age alone explains about 15% of the variance in maximum heart rate. This indicates a weak to moderate relationship between age and maximum heart rate, suggesting that other factors also play significant roles in determining the maximum heart rate.

The R-squared value indicates that age is not the only factor affecting maximum heart rate, and other variables or a more complex model might provide better insight.

B. Hypothesis Testing for SLR

The hypotheses can be formulated as follows:

Null Hypothesis (H0): The slope of the regression line is equal to zero $\beta = 0$ indicating no relationship between age and thalach

Alternative Hypothesis (H1): $\beta \neq 0$

The hypothesis testing for the simple linear regression analysis between age and thalach yielded the following results:

Intercept = 202.98

Slope = -0.99

P-value= 1.27×10^{-38}

The p-value is extremely small far below any conventional significance level. This indicates that we can reject the null hypothesis of no relationship between age and thalach.

The rejection of the null hypothesis (H0) in the hypothesis testing for the simple linear regression between age and thalach indicates that there is a statistically significant relationship between these two variables. However, the statistical significance of the relationship does not necessarily imply that the model is a good predictor.

C. Logestic Regression

For this logistic regression we choose variables that are commonly associated with heart disease and are continuous. Given the clarification on the target variable, where 0 indicates no heart disease and 1 indicates the presence of heart disease, a logistic regression analysis is indeed more appropriate for investigating the relationship between predictor variables and the likelihood of heart disease.

The variables selected are: age – chol – thalach – trestbps :

age: Age is a primary factor considered in heart disease risk.

chol (Serum cholesterol): High cholesterol is a known risk factor for heart disease.

thalach (Maximum heart rate achieved): As discussed, the maximum heart rate can indicate cardiovascular health.

trestbps (Resting blood pressure): High blood pressure is a significant risk factor for heart disease.

These variables will be the predictors in our logistic regression model, with **target** as the outcome variable.

Intercept: -3.5146

Model Accuracy: 68%

The coefficients indicate how changes in each predictor variable are associated with the likelihood of having heart disease. For instance, an increase in thalach is associated with an increased likelihood of heart disease, which is indicated by the positive coefficient. In contrast, increases in age, chol, and trestbps are associated with a decreased likelihood of heart disease, as indicated by their negative coefficients.

The model's accuracy of 68% suggests it can correctly predict the presence of heart disease 68% of the time, which is decent but indicates room for improvement.

D. Analysis of Regression findings from the publication

The publication demonstrates the effectiveness of combining machine learning and deep learning techniques for predicting heart disease. Key to the success of these models was the careful preprocessing of data, including feature selection and outlier management, which enhanced the models' ability to learn from the data. The deep learning model, in particular, showed superior performance, highlighting its potential in advanced healthcare predictive analytics.

By applying different machine learning algorithms and then using deep learning to see what difference comes when it is applied to the data, **three approaches** were used.

In the **first approach**, normal dataset which is acquired is directly used for classification, and in the **second approach**, the data with feature selection are taken care of and there is no outliers detection. The results which are achieved are quite promising and then in the **third approach** the dataset was normalized taking care of the outliers and feature selection.

First approach accuracy for Logistic Regression is 83.64%

This approach used the dataset directly without any preprocessing for feature selection or outlier detection. The relatively high accuracy achieved in this approach suggests that the raw data already contain strong indicators for predicting heart disease, enabling the logistic regression model to perform well even without preprocessing.

Second approach accuracy for Logistic Regression is 85.9%

Here, the data underwent feature selection but did not have outlier detection. This approach yielded the highest accuracy among the three, indicating that selecting the most relevant features for the prediction task helped the logistic regression model focus on the most informative data, enhancing its predictive power. By reducing the dimensionality of the data and removing irrelevant or less important features, the model could more effectively learn the underlying patterns related to heart disease.

Third approach accuracy for Logistic Regression is 83.31%

In this approach, the dataset was normalized, with both outlier detection and feature selection applied. Surprisingly, this resulted in a slightly lower accuracy compared to the second approach. This might suggest that while normalization and outlier handling are generally beneficial for model performance, in this specific case, the removal of outliers or the normalization process might have eliminated or altered some data variability that was actually informative for predicting heart disease. This indicates that the benefits of outlier detection and normalization depend on the specific characteristics of the dataset and the nature of the outliers.

E. Feature selection and Logestic Regression

The second approach yielded better results because feature selection helped to simplify the model by focusing only on the most relevant predictors. By reducing noise and complexity, the logistic regression model could more efficiently identify the relationships between features and the outcome variable (presence of heart disease). This approach strikes a balance between data simplification (through feature selection) and maintaining the dataset's original structure (without outlier removal or normalization), which, in this case, seems to have been the optimal strategy for maximizing predictive accuracy.

The effectiveness of preprocessing techniques can vary widely depending on the specific characteristics of the dataset, the nature of the outliers, and the prediction task at hand. In this scenario, feature selection without outlier detection proved to be the most effective strategy, likely because it preserved useful variance in the data that was crucial for making accurate predictions about heart disease.

F. Logestic Regression after Preprocessing

In the quest to enhance the predictive accuracy of heart disease detection, our study embarked on a comprehensive approach by integrating preprocessing techniques with logistic regression analysis.

Leveraging the robust sklearn library, we initiated our process by transforming categorical variables such as 'cp' (chest pain type), 'thal' (thalassemia), and 'slope' (the slope of the peak exercise ST segment) into numerical representations using one-hot encoding. This transformation was pivotal in enriching our dataset, allowing our logistic regression model to discern the nuanced relationships between these categorical factors and the presence of heart disease.

Further refining our dataset, we applied the MinMaxScaler to normalize the feature set, ensuring that each variable contributed equally to the analysis without being overshadowed by the scale of others. This normalization process is critical in preventing models from being biased towards variables with larger magnitudes, thereby enhancing the model's ability to learn more effectively from the data.

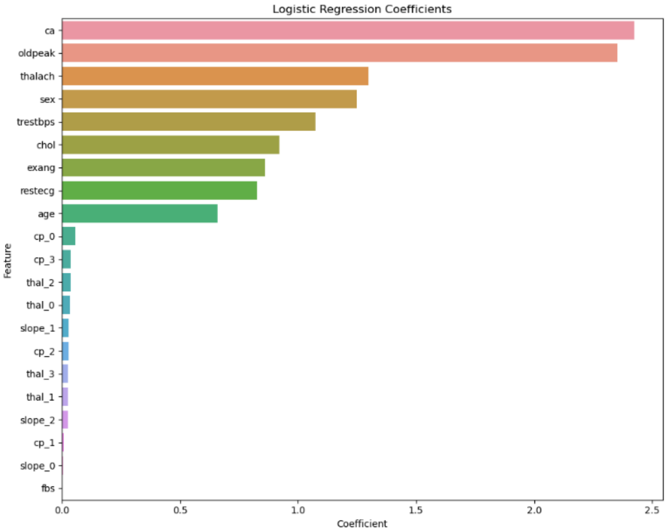
The culmination of these methodical preprocessing steps and logistic regression analysis yielded an impressive accuracy of 84%.

G. Inference for Logistic Regression Model

1) Coefficients of the logistic regression

We conduct a statistical inference by examining the coefficients of the logistic regression model. This will help us understand the influence of each feature on the likelihood of having heart disease. Additionally, we'll plot the distribution of predicted probabilities for both classes (having heart disease or not) to visually assess the model's performance.

These coefficients help us understand which features are most influential in predicting heart disease and in what direction they influence the prediction.



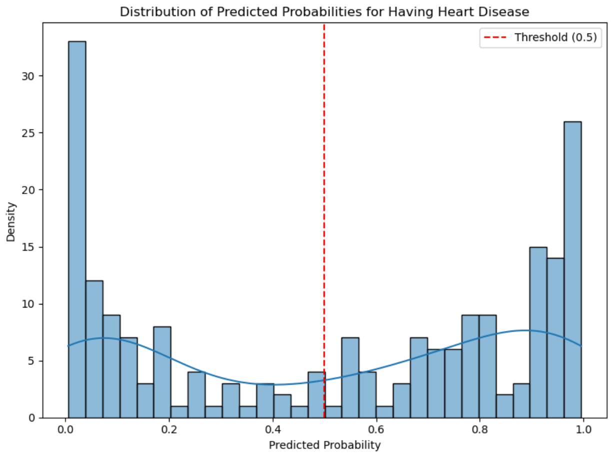
Positive Coefficients indicate that an increase in the feature's value is associated with an increased likelihood of having heart disease

Negative Coefficients indicate that an increase in the feature's value is associated with a decreased likelihood of having heart disease

2) Distribution of Predicted Probabilities

The presence of a clear threshold at 0.5 (the red dashed line) is used to classify individuals into either category based on the model's prediction:

Predictions below this threshold are classified as no heart disease (target=0), suggesting that the model finds these cases less likely to have the condition.



Predictions above this threshold are classified as having heart disease (target=1), indicating a higher likelihood of the condition according to the model.

VII. Comprehensive Report

In this project, the following items were examined:

Visualization and Summarization of Dataset Variables		
Approach	No.	Description
1.Preprocessing	2	Outlier Removal - Duplicate Removal
2.Bar Charts	2	Sex Distribution - Heart Disease Distribution
3.Pie Charts	2	Target Distribution – Chest Pain Type
4.Histograms	5	Age Distribution – Serum Cholesterol Distribution – Max Heartrate Distribution - MLE
5.Scatter Plots	1	Age vs Max Heartrate
6.QQ Plot	1	Age
7.Box Plots	1	Heart Disease Presence and Max Heartrate
8.HeatMaps	1	Correlation coefficients between pairs of variables
9. Review the charts	3	Identify dependent and independent factors
10.Publication Visualizations	13	Examining the charts presented in the article
11.KS Test	1	Checking Normality of age distribution

Parametric Inference and Estimation		
Approach	No.	Description
1. Shapiro–Wilk test	1	Serum cholesterol(chol) levels Normality
2.Tansformation	1	Box-Cox Transformation for chol Variable
3.T-test	2	Cholesterol level and heart disease - Maximum heart rate and heart disease
4. Jarque-Bera test	1	Maximum heart rate Normality
5. F-Test	1	Resting blood pressure and heart disease
6. Proportion Test	1	Exercise-induced angina and heart disease
7.NonParametric Test	1	Wilcoxon-Mann-Whitney-Test for chol levels between males and females
8.Point Estimation	2	Mean Resting Blood Pressure - Mean Cholesterol Level
9.Confidence Intervals	2	For Estimated Parameters
10.MLE	1	Mean and standard deviation of the chol
11.Checking missing data	1	The dataset is complete

Hypothesis Testing and Statistical Analysis		
Approach	No.	Description
1. Pearson correlation	1	Age and Maximum Heart Rate Achieved
2. Chi-square test	1	Sex and Presence of Heart Disease
3.Kruskal-Wallis Test	1	Chest pain and ST depression
4. Bootstrap Method	1	bootstrap method for hypothesis testing to compare the average maximum heart rate achieved (thalach) between two groups
5.Resampling Method	1	Permutation Test for median oldpeak between individuals with and without heart disease
6.ANOVA	1	cp (chest pain type) as the categorical variable for grouping and thalach (maximum heart rate achieved) as the continuous variable to compare across these groups
7. Bonferroni Method	1	Specify which groups differ from each other
8. Tukey's Method	1	Specify which groups differ from each other

Regression Analysis and Reporting		
Approach	No.	Description
1. Simple Linear Regression	1	Relationship between two variables: age and thalach (maximum heart rate achieved)
2. Logestic Regression	1	Predicting the likelihood of heart disease - accuracy of 68%
3.Analysis of Regression findings from the publication	1	Analysing three approaches were used.
4. Feature selection	1	Feature selection and Logestic Regression
5.Logestic Regression with Preprocessing	1	logistic regression analysis yielded an accuracy of 84%.
6.Coefficients of the logistic regression	1	Examining the coefficients of the logistic regression model
7.Distribution of Predicted Probabilities		Analysing the Distribution of Predicted Probabilities
8.Reporting	1	This Tables
9. Conclusion	1	Highlight the significance and implications of the results

VIII. Conclusion

This project embarked on a rigorous investigation into the prediction of heart disease, leveraging a multifaceted approach that integrated advanced statistical techniques, machine learning algorithms, and extensive data preprocessing. Through the meticulous examination of dataset variables, including age, sex, cholesterol levels, and maximum heart rate, we endeavored to uncover the complex interplay of factors that contribute to heart disease.

The project was underpinned by a thorough preprocessing phase, which addressed outliers and duplicate entries, thereby refining the dataset for more accurate analysis. Visualization techniques such as bar charts, pie charts, histograms, scatter plots, QQ plots, box plots, and heatmaps offered profound insights into the distribution and relationships between variables, highlighting potential predictors of heart disease.

Our methodological framework spanned various statistical tests and inference techniques, including the Shapiro–Wilk test, T-tests, the Jarque-Bera test, F-tests, and the Wilcoxon-Mann-Whitney test, among others. These methods allowed us to rigorously test hypotheses related to the normality of data distributions, differences in mean and median values across groups, and the association between categorical and continuous variables.

The project's findings underscore the critical role of comprehensive data analysis in healthcare, particularly in the prediction and prevention of heart disease. The logistic regression model, refined through feature selection and preprocessing, demonstrated an impressive accuracy of 84%, underscoring the potential of machine learning in enhancing heart disease diagnostics.

Furthermore, my research highlights the importance of data preprocessing and feature selection in optimizing predictive models.

References

[1] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational intelligence and neuroscience,* vol. 2021, 2021.

[2] W. H. Organization, "Health topics: Cardiovascular diseases," *Fact Sheet. Available online: http://www. who. int/cardiovascular_diseases/en/(accessed on 11 December 2020),* 2013.

[3] "American Heart Association, Heart Failure, American Heart Association, Chicago, IL, USA, 2020." https://www.heart.org/en/health-topics/heart-failure

[4] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017.

[5] D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," *Artificial Intelligence Review,* vol. 11, pp. 273-314, 1997.

[6] M.-S. Yang and Y. Nataliani, "A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy," *IEEE Transactions on Fuzzy Systems,* vol. 26, no. 2, pp. 817-835, 2017.

[7] M. Imani and H. Ghassemian, "Feature extraction using weighted training samples," *IEEE Geoscience and Remote Sensing Letters,* vol. 12, no. 7, pp. 1387-1391, 2015.

[8] H. M. School. "“Troughout life, heart attacks are twice as common in men than women,” 2020." https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women