Heart Attack Prediction: A Scientific Study of Influencing Factors

By:

Amerah Otman Alamri-2105633

Wesam Mubarak Alsulami-1908409

Badriya Saeed Agala-1915665

Layan Wail Alansari-2012364

Sara Turki Alsehli-2011635

Supervised by:

Dr. Amani Alghamdi

Ms. Doaa Bogari

Department of Statistics - Faculty of Science

King Abdulaziz University-Jeddah

1445-2024

التنبؤ بالنوبات القلبية: دراسة علمية للعوامل المؤثرة

عمل:

اميرة عثمان العمري -2105633

وسام مبارك السلمي-1908409

بدرية سعيد عقالا-1915665

ليان وائل الأنصاري-2012364

سارة تركي السهلي-2011635

تحت إشراف:

د. أماني الغامدي

أ. دعاء بوقري

قسم الإحصاء – كلية العلوم

جامعة الملك عبد العزيز - جدة

1445- 2024

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

Acknowledgments

First and foremost, we thank Allah for everything that has been provided for us, who has enabled us to present this research. We would also like to extend our appreciation, thanks, and gratitude to our research supervisors, Dr. Amani Al-Ghamdi and Ms. Doaa Abdulhafiz, for their valuable comments and constructive criticism throughout the academic semester of 2024. It has been a great honor for us to work under their supervision. We would also like to thank the Statistics Department for providing us with the opportunity to focus on students' development during our educational journey. Additionally, we thank our fellow researchers for their cooperation, enthusiasm, prompt responsiveness, patience, and excellent work in presenting a successful and distinguished research project. Finally, we express our gratitude to our families and friends for their continuous encouragement and support.

Abstract

Heart Attack is a pervasive and enormous health challenge worldwide, taking millions of lives every year and placing a heavy burden on global healthcare systems. Among various cardiovascular conditions, heart attacks, or cardiac infarctions, stand out as particularly devastating events with profound effects on patients, families, and communities. Despite progress in medical science and healthcare delivery, the ability to predict and prevent heart attacks remains a major concern. Early detection and intervention are central to mitigating the adverse effects of heart attacks, emphasizing the importance of predictive analyses in cardiovascular medicine. Finding the most important variables affecting cardiac arrest and identifying the most effective statistical model for properly diagnosing the heart attack are the objectives of this study to accurately diagnose cardiac arrest, we compared two statistical models: the naive Bayes classification, and the logistical regression. Five statistical measures have been used to evaluate the performance of both models: The confusion matrix, accuracy, error rate, sensitivity, specificity receiver operating characteristic (ROC), and area under the curve (AUC) of ROC. the logistics regression model is the best to achieve our specific goal of analyzing positive cases with an accuracy of 0.9, and error rate 0.1 and sensitivity 0.92 and specificity 0.87 followed by the naive Bayes classification with an accuracy of 0.81 and error rate 0.19 and sensitivity 0.81 and specificity 0.8. we also discovered that the main symptoms of cardiac arrest are maximum heart rate achieved, chest pain type, serum cholesterol levels, exercise-induced angina, and number of major vessels colored by fluoroscopy.

الملخص

الأمراض القلبية تشكل تحديًا صحيًا هائلًا ومنتشرًا على نطاق واسع في العالم، حيث تودي بحياة ملايين الأشخاص سنويًا وتفرض عبئًا ثقيلاً على أنظمة الرعاية الصحية العامة. بين مجموعة متنوعة من الحالات القلبية، تبرز النوبات القلبية بشكل خاص كواحدة من أكثر الحالات تدميرًا، مما يترتب عليها آثار عميقة على المصابين وأسرهم والمجتمعات المحلية. على الرغم من التقدم في العلوم الطبية وتطوير الرعاية الصحية، فإن القدرة على التنبؤ بالنوبات القلبية ومنع حدوثها لا تزال تشكل تحديًا كبيرًا.

الكشف المبكر والتدخل الفوري يعدان عنصرين أساسيين لتقليل الآثار الضارة الناتجة عن النوبات القلبية، مع التأكيد على أهمية الاستراتيجيات التنبؤية في مجال طب القلب والأوعية الدموية. تهدف هذه الدراسة إلى تحديد العوامل المؤثرة الرئيسية في النوبات القلبية وتحديد النماذج الإحصائية الأكثر فعالية للتشخيص الدقيق.

لتقديم تشخيص دقيق للنوبات القلبية، قارننا بين نموذجين إحصائيين: تصنيف البايز البسيط والانحدار اللوجستي. واستخدمنا خمس مقاييس إحصائية لتقييم أداء كل نموذج: مصفوفة الارتباك، والدقة، ومعدل الخطأ، والحساسية، والخصوصية. أظهر نموذج الانحدار اللوجستي أداءً متفوقًا في تحقيق هدفنا المحدد من تحليل الحالات، حيث بلغت دقته ٩٠٪ ومعدل الخطأ ١٠٪ والحساسية ٩٢٪ والخصوصية ٨٧٪. بينما جاء تصنيف البايز البسيط بدقة ٨١٪ ومعدل الخطأ ١٩٪ والحساسية ٨١٪ والخصوصية ٨٠٪. وكشفت الدراسة أيضًا أن الأعراض الرئيسية للنوبات القلبية تتمثل في الحد الأقصى لمعدل ضربات القلب، ونوع آلام الصدر، ومستويات الكوليسترول، وآلام الصدر الناتجة عن التمارين البدنية، وعدد الأوعية الملونة بواسطة الفلوروسكوب.

Contents

List of Figures

List of Tables

List of Symbols

Chol: Serum Cholesterol

Thalach: Maximum Heart Rate Achieved

Trestbps: Resting Blood Pressure

Cp: Chest Pain Type

Fbs: Fasting Blood Sugar

Restecg: Resting Electrocardiographic Results

Exang: Exercise-Induced Angina

Oldpeak: Oldpeak (ST Depression)

Slope: Slope of Peak Exercise ST Segment

Ca: Number of Major Vessels Colored by Fluoroscopy

Thal: Thalassemia

RUC: Receiver Operation Characteristic

AUC: Area Under the ROC Curve

VIF: Variance Inflation Factor

BLR: Binary Logistic Regression

GNB: Gaussian Naïve Bayes

CNB: Categorical Naïve Bayes

Chapter 1: Introduction

1.1 Introduction

Heart disease is a pervasive and formidable health challenge worldwide, claiming millions of lives annually and presenting a significant burden on healthcare systems globally. Among the various cardiovascular conditions, heart attacks, or myocardial infarctions, stand out as particularly devastating events with profound implications for patients, families, and societies. Despite advancements in medical science and healthcare delivery, the ability to predict and prevent heart attacks remains a paramount concern. Early detection and intervention are pivotal in mitigating the adverse effects of heart attacks, underscoring the importance of predictive analytics in cardiovascular medicine. This chapter serves as a gateway to exploring the intricacies of heart attack prediction, delving into the research problem, its significance, objectives, and the underlying concepts and methodologies employed in this endeavor.

1.2 Research Problem

The research problem addressed in this study revolves around the prediction of heart attacks using statistical modeling techniques. This involves analyzing various factors and their impact on the likelihood of heart attack occurrence, aiming to develop effective predictive models for early detection and intervention.

1.3 Research Importance

The importance of predicting heart attacks lies in its potential to save lives by identifying individuals at high risk and providing timely medical intervention. Early detection can significantly improve patient outcomes and reduce the burden on healthcare systems.

1.4 Research Objectives

The main objectives of this research are:

To develop predictive models for heart attack occurrence using logistic regression and naive Bayes methods.

To compare and evaluate the efficacy of different models in terms of accuracy, error rate, sensitivity, specificity, receiver operating characteristic (ROC) curve, and area under the curve (AUC) of ROC, with the goal of identifying the most effective model for prediction or classification purposes.

To assess the significance of various risk factors in predicting heart attacks and understand their interrelationships.

1.5 Definitions and Concepts

1.5.1 Heart Attacks

Heart attacks, also known as myocardial infarctions, are acute medical conditions resulting from partial or complete blockage of the coronary arteries, leading to damage to the heart muscle and potentially life-threatening symptoms such as chest pain and shortness of breath.

1.5.2 Factors Influencing Heart Attack Prediction

Several factors influence the prediction of heart attacks, including biological factors such as age, gender, and medical history, as well as behavioral and environmental factors like lifestyle, diet, and exercise. Accurate analysis of these factors is essential for understanding the relationships with heart attack occurrence and developing effective predictive models.

1.5.3 Statistical Modeling for Heart Attack Prediction

Statistical modeling involves using various methods and techniques to analyze data and predict heart attacks. These methods include logistic regression and both simple and advanced Bayesian models, which estimate the relationships between predictive variables and the likelihood of heart attack occurrence.

1.6 Outline of the Research

This research is structured as follows:

Chapter 2: Literature Review: Discusses previous relevant studies in the field of heart attack prediction.

Chapter 3: Methodology: Explains the research methodology, tools, and study population for the applied, theoretical, and simulation studies.

Chapter 4: Results and Discussion: Presents descriptive, inferential, and theoretical analyses of the study's findings.

Chapter 5: Summary and Future Work: Summarizes the study's outcomes, scientific recommendations, and suggests avenues for future research.

Chapter 2: Literature Review

2.1 Literature Review

In this section, we highlight contributions made in the statistical analysis of heart attack.

A study by Sonam Nikhar (2006), proposed explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of heart disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier. [1]

Another study by hassan and mamun (2018), Their research provides a detailed explanation of Naïve bayes, decision trees, and logistic regression. They used ROC and specificity to measure performance, and the decision was that the performance of logistic regression It was the best and its classification accuracy was 92.76%. [2]

Another study by Gholam Hossein Alishiri (2008) explains Logistic Regression Models for Predicting Health-Related Quality of Life (HRQOL) in Rheumatoid Arthritis (RA) Patients. The study developed logistic regression models to predict both the physical and mental aspects of Health-Related Quality of Life among 411 Rheumatoid Arthritis patients. Using SF-36 measurements, the study found that poor Health-Related Quality of Life was associated with factors such as pain severity, duration of the disease, monthly family income below $300, presence of comorbidities, patient's global assessment of disease activity, and depression. The models demonstrated optimal sensitivity and specificity, indicating their potential for effective prediction and clinical decision-making in managing Rheumatoid Arthritis. [3]

The researcher Manikandan, S. conducted this study in (2017). The aim of the research was to simplify and speed up diagnosing heart failure by introducing automation through a binary classifier for risk prediction. The study presented a prototype implementation of such a system, complete with a user-friendly web-based graphical interface. The classifier utilized the Naïve Bayes algorithm, achieving an accuracy score of 81.25%. This approach has the potential to enhance efficiency and accuracy in diagnosing heart failure, ultimately improving patient care in the healthcare industry. [4]

Another study by Feng Xiao & QiZhou He (2021), that statistics indicate that cardiovascular diseases are prominent among the leading causes of death. Concurrently, the incidence of coronary heart disease (CHD) was 7.2% according to global demographic disease statistics from 2015 to 2018, with the number of deaths due to CHD reaching 360,900 in 2019 alone. Risk assessment and early diagnosis of CHD are significant for improving safety and quality of life. All patients admitted to the Traditional Chinese Medicine Hospital at Southwest Medical University from January 2019 to March 2022, suspected of having CHD and undergoing CCTA examination, were retrospectively selected. The degree of coronary artery calcification (CAC) was quantified based on the Agatston score. To compare the relationship between coronary artery calcification score (CACS) and clinical factors, 31 variables were collected, including hypertension, diabetes, smoking, and hyperlipidemia, among others. Machine learning (ML) models containing the random forest (RF), radial basis function neural network (RBFNN), support vector machine (SVM), K-Nearest Neighbor algorithm (KNN), and kernel ridge regression (KRR) were employed to assess the risk of CHD based on CACS and clinical factors. [5]

Chapter 3: Methodology

3.1 Introduction

In this chapter, we delve into the Heart attack prediction dataset and explore various aspects of it. The dataset's pre-processing is crucial for ensuring the accuracy and reliability of the analysis, and we detail this process in Section 2.3 Moving on to Section 2.4, we present an explanation of the logistic regression and naive Bayes models.

Section 2.4.1, we present the logistic regression model that will be used to predict heart attacks based on the dataset’s variables.

In Section 2.4.2, we delve into decision naive Bayes, another method that can be utilized for heart attack prediction. By comparing and contrasting the logistic regression model with decision naive Bayes, we aim to provide comprehensive insights into the efficiency and effectiveness of these two approaches.

Lastly, in Section 2.5, we present Performance evaluation and involves three points: confusion matrix, receiver operating characteristic (ROC), and area under the curve (AUC) of ROC.

3.2 The Heart attack prediction Data

A sample size of 1026. There are fourteen attributes in the dataset; we used all the data, using thirteen for analysis and one as the target attribute. This target attribute, (target 0=No disease; 1=Disease) was selected in order to predict Heart Attack . The source of it was the platform for data science competitions (Kaggle) Heart Disease Dataset (kaggle.com). We used the R software to clean and examine the data. Several statistical. packages from the R software were employed in this investigation. The variables taken into account in the study are listed in table 3.1 along with their names and types.

Table 3.1 The variables in the Heart attack dataset

Variable	type
Sex	Qualitative (Categorical, Two levels)
Age	Quantitative (Numeric)
Serum Cholesterol (chol)	Quantitative (Numeric)
Maximum Heart Rate Achieved (thalach)	Quantitative (Numeric)
Resting Blood Pressure (trestbps)	Quantitative (Numeric)
Chest Pain Type (cp)	Qualitative (Categorical, Four levels)
Fasting Blood Sugar (fbs)	Qualitative (Categorical, Two levels)
Resting Electrocardiographic Results (restecg)	Qualitative (Categorical, Three levels)
Exercise-Induced Angina (Exang)	Qualitative (Categorical, Two levels)

Oldpeak (ST Depression)	Quantitative (Numeric)
Slope of Peak Exercise ST Segment (slope)	Qualitative (Categorical, Three levels)
Number of Major Vessels Colored by Fluoroscopy(ca)	Quantitative (Numeric)
Thalassemia (thal)	Qualitative (Categorical, Three levels)
Target	Qualitative (Categorical, Two levels)

3.3 Pre-processing tools

Data preprocessing is the process of preparing data for analysis. Data preprocessing is an important phase in the data mining process, as it helps to ensure that the data is accurate, complete and suitable for analysis.

3.3.1 Missing Value

Missing values are missing data that was not written by the entry, whether intentionally, forgotten, or the connection to it was lost. In the case of missing values, we delete the row or variable that contains many missing values or fill it with the mean of the numerical variable or the most frequent value for the categorical variables.

3.3.2 Outliers

They are the values that are too large or too small relative to the average of the data, which are outside the boundaries of the data and affect the accuracy of the data analysis. The problem is solved by trimming estimators, which omit the beginning and end of the data to reduce outliers.

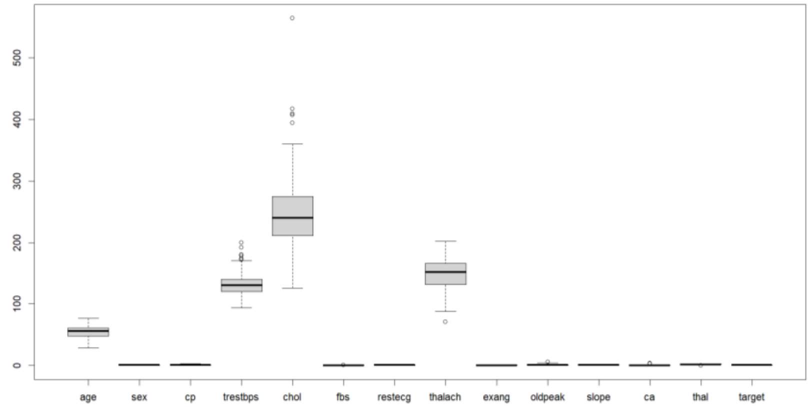


Figure 3.1 Box-plots to Check outliers

Figure 3.1 shows that there is extreme value. The percentage of the amount of outliers is normal.

3.3.3 Splitting Data

The study divided the data into 70% training and 30% testing sets, ensuring a comparable training and testing set. The majority of the data is used for training, with a small portion for testing. Random sampling ensures comparable data. Testing assesses model performance, while training builds models.

3.4 Naive Bayes and Logistic regression

Naive Bayes and logistic regression are two popular supervised machine learning algorithms that learn from labeled data to make predictions. They are used for classification problems where the goal is to assign categories or labels to input based on features. In addition, Naive Bayes is a probability based on Bayes theory, while logistic regression is a discriminant model that estimates conditional probabilities. Naive Bayes can also be used for regression tasks. [6]

3.4.1 Logistic Regression

Logistic regression is a powerful analytic tool used to predict binary outcomes, such as the occurrence of heart attacks. Utilizing logistic regression in studying heart attack prediction offers numerous advantages, including:

- Relationship Analysis: Logistic regression provides a method to analyze the relationship between various variables, such as age, blood pressure, cholesterol levels, and the likelihood of a heart attack occurrence.
- Estimating Event Probabilities: Logistic regression provides accurate estimates of the probability of a heart attack based on different variable values in the model.
- Case Prediction: Through a logistic regression model, data can be used to predict the likelihood of a heart attack occurrence in a new set of units, enabling healthcare professionals and researchers to take necessary preventive measures.

$$P = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} \;,$$

[7]

where:

- P is the probability of an event occurring.
- β_0 intercept.

- β_1 the regression coefficient.
- $x_i$is an independent variable.

Assumptions

- The independent variables are continuous or categorical.
- Observations are independent of each other.
- The dependent variable is binary or dichotomous.

Binary Logistic Regression (BLR)

Binary logistic regression is a statistical technique used to model the relationship between a binary outcome variable and one or more predictor variables. In binary logistic regression, the outcome variable is categorical and has only two possible outcomes, typically coded as 0 and 1. The goal of binary logistic regression is to estimate the probability that the outcome variable is equal to one of the two existing categories based on the predictor variables. It predicts the log-odds of the dependent variable based on the independent variables. [7]

In logistic regression, the logistic transformation of the odds (referred to as logit) is used as the dependent variable. This transformation, also known as the logarithm of P or logit of P, establishes a relationship with the standard regression equation. The logit of P is a crucial component in statistical modeling, aiding in predicting the outcomes of experiments and improving the accuracy of these predictions by establishing strong associations between various variables and the expected outcomes.

$$\log(odds) = logit(P) = \ln\left(\frac{P}{1-P}\right)$$

If we take the dependent variable above and add the regression equation for the independent variables, we get logistic regression:

$$logit(p) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots$$

The relationship between logit(P) and X is assumed to be linear.

The binary logistic regression equation can be expressed as:

$$P = \frac{exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots)}{1 + exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots)} \ ,$$

[7]

where:

- P The probability of a case being in a specific category.
- exp is the exponential function.
- b_0 intercept.
- b is the coefficient of predictor variables.

3.4.2 Naive Bayes Classifier

The Naive Bayes classifier is a probabilistic machine learning model based on Bayes' theorem. It assumes independence between features, which means the presence of a specific feature in a class is independent of the presence of another feature, and calculates the probability that a given input belongs to a specific class.

The Naïve Bayes classifier helps predict the probability of developing a heart attack based on symptoms. and provides many advantages, including:

- Simplicity: Naive Bayes is a simple and easy-to-understand algorithm.
- It is computationally efficient and scales well to large datasets.
- It is fast and making predictions is easy with high dimension of data.
- It works well when categorical features are present.

One disadvantage is the assumption that features are independent. Therefore, Naive Bayes may lead to suboptimal results if features are highly correlated. [6]

Assumption:

- Naive Bayes assumes that all features are conditionally independent of each other given the class variable.
- Features are equally important: All features are assumed to contribute equally to predicting class names.
- No missing data: The data should not contain any missing values.

[6]

The Naive Bayes classifier is a popular classification model based on Bayes' theorem. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (C_i) is independent of the values of other predictors. Applying Bayes' theorem, and simplifying the notation a little, we obtain:

$$p(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)} \quad ,$$

where:

$p(C_i|X)$is the posterior probability of *class* given *predictor*.

$P(X|C_i)$ is Likelihood probability.

$P(C_i)$ is the prior probability of class.

$P(X)$ is the evidence.

That is, the instance X belongs to class C_i if and only if

$p(C_i|X)$> $p(C_j|X)$ for $1 \le j \le m, j \neq i$.

In this research, we will apply two types of Naive Bayes model:

1- Gaussian Naive Bayes (GNB):

Gaussian Naive Bayes assumes that the continuous values associated with each feature are distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. conditional probability is given by: [6]

$$P(X_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_{i-}\mu)^2}{2\sigma^2}\right)$$

where μC_i and σC_i are the mean and standard deviation of X over the class C_i .

2- Categorical Naive Bayes (CNB):

Categorical Naive Bayes is useful when the features are categorical distributions. To use this algorithm, we need to encode the categorical variables in numeric format using an ordinal encoder. The conditional probability is computed from training data: $X = \{a_1, a_2, a_3\}$ [8]

$$p(X = a_k|C_i) = \frac{|a_k, C_i|}{|C_i|}$$

where:

- $|a_1, C_i|$ is the number of training examples corresponding to feature variable a_k and class C_i.
- $|C_i|$ is the number of training examples in class C_i .

3.5 Performance Evaluation

1- Confusion Matrix: is a table that is frequently used to summarize a classification algorithm’s performance on a set of test data. A confusion matrix displays the number of correct and incorrect predictions made by the classification model in relation to the data’s actual outcomes (target value) and can help you understand what your classification model gets right and what types of errors it makes. [12]

Table 3.2: Confusion Matrix Table.

	Predicted class		
Actual Class		Yes	No
	Yes	TP	FN
	No	FP	TN

- True positive (TP): The number of positive instances that were correctly predicted by the classifier.
- True negative (TN): The number of negative instances that were correctly predicted by the classifier.

- False positive (FP): The number of negative instances that were incorrectly predicted as positive. This is also known as the type 1 error, **α**.

- False negative (FN): The number of positive instances that were predicted as negative. This is also known as the type 2 error, β.

The confusion matrix allows us to compute the following:

- **Accuracy:**

It is the percentage of test instances that are correctly classified by the classifier. It is defined by:

$$Accuracy = \frac{TP+TN}{P+N}$$

- **Error rate (misclassification rate):**

It is the percentage of test instances that are incorrectly classified by the classifier. It is defined by:

$$Error\,rate\ = \frac{FP + FN}{P + N}$$

- **Sensitivity (Recall):**

It is the proportion of positive instances that are correctly classified (true positive rate). It is defined by:

$$Sensivity = \frac{TP}{P}$$

- **Specificity:**

It is the proportion of negative instances that are correctly classified (true negative rate). It is defined by:

$$Specificity = \frac{TN}{N}$$

2- Receiver Operation Characteristic (ROC): Using the ROC curve to evaluate the performance of your binary classifier involves several steps and considerations. Ensure your classifier can generate probabilistic predictions or scores indicating the likelihood of each instance belonging to the positive class. These scores will be used to vary the threshold and plot the ROC curve. Ensure your classifier can generate probabilistic predictions or scores indicating the likelihood of each instance belonging to the positive class. These scores will be used to vary the threshold and plot the ROC curve. An AUC of 1.0 indicates a perfect classifier, while an AUC of 0.5 indicates no discriminative power. A higher AUC value indicates a better classifier. The ROC curve can help identify a suitable threshold that balances these according to your requirements. The one with a curve closer to the top-left, or with a higher AUC, generally indicates a better classifier. It provides a mechanism to evaluate classifier performance in a way that is independent of the class distribution or specific costs. In such cases, other metrics such as the Precision-Recall curve might provide a more informative performance assessment.

By following these steps and considerations, you can effectively use the ROC curve to evaluate, compare, and improve your binary classifiers, making informed decisions on their deployment based on your specific performance criteria. [11]

3- Area Under the ROC Curve (AUC): AUC stands for "area under the ROC curve." That is, AUC measures the entire two-dimensional area under the entire ROC curve from (0,0) to (1,1). A model that predicts 100% of the time wrong has an AUC of 0.0; a model that predicts 100% of the time correctly has an AUC of 1.0. For a diagnostic technique to be useful, the AUC must be greater than 0.5, and generally must be greater than 0.8 to be considered acceptable. Furthermore, when comparing the performance of two or more diagnostic tests, the ROC curve with the largest AUC is considered to have better diagnostic performance. [11]

Chapter 4: Results and Discussion

4.1 Introduction

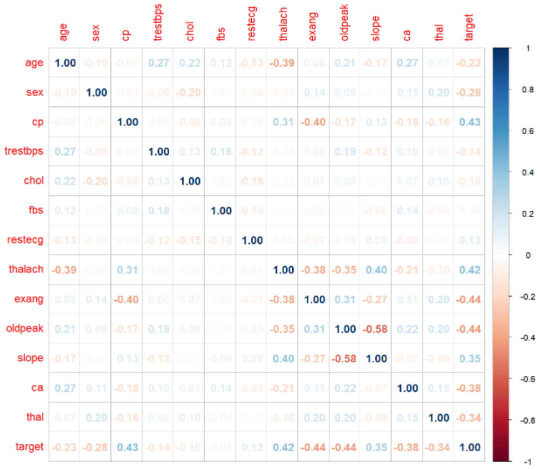
In this section, we will provide the results of our research, summarize the dataset, review and evaluate predictive models of heart attack.

4.2 Exploratory Data Analysis and Visualization

Table 4.1: Descriptive Analysis of Numeric Variables.

Variables	Min	1st Qu	Median	Mean	3rd Qu	Max
Age	29.00	48.00	56.00	54.43	61.00	77.00
Resting Blood Pressure (Trestbps)	94.00	120.0	130.0	131.6	140.0	200.0
Serum Cholesterol (Chol)	126.0	211.0	240.0	246.0	275.0	564.0
Maximum Heart Rate Achieved (Thalach)	71.0	132.0	152.0	149.1	166.0	202.0

4.3 Correlation Matrix



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-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	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.20	-0.28
cp	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.04	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1.00	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.10	0.06	-0.14
chol	0.22	-0.20	-0.08	0.13	1.00	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.10	-0.10
fbs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	[illegible]	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10	1.00	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	[illegible]	0.05	1.00	-0.38	-0.35	0.40	-0.21	-0.10	0.42
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38	1.00	0.31	-0.27	0.11	0.20	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1.00	-0.58	0.22	0.20	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58	1.00	-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1.00	0.15	-0.38
thal	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1.00

Figure 4.1: Correlation Between Variables

From Figure 4.1, we can observe that there is no strong correlation between variables. This indicates that the assumption of no collinearity between independent variables has been achieved, and there is no need to remove any variables.

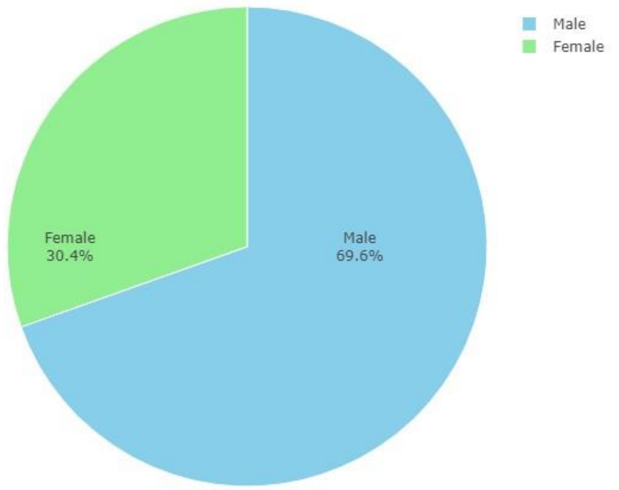


Figure 4.2: Pie chart for sex

Figure 4.2 shows that most of the people in the dataset belong to the male category.

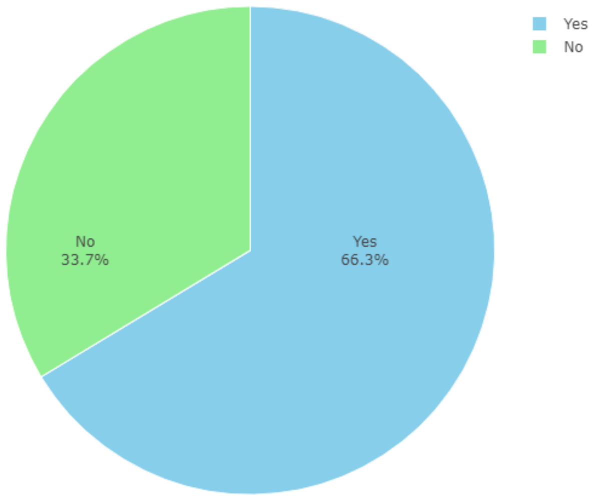


Figure 4.3: Pie chart of Exercise-Induce Angina (Exang)

Figure 4.3 shows that the highest percentage is for people who suffer from angina pain resulting from exercise.

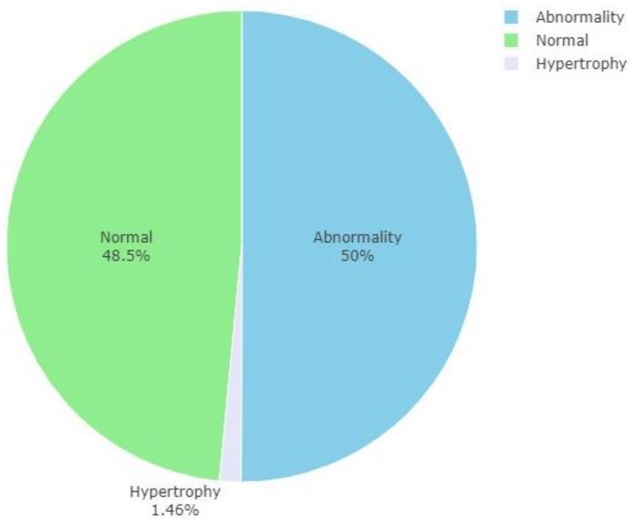


Figure 4.4: Pie chart of Resting Electrocardiographic Results (Restecg)

Figure 4.4 shows that the highest percentage of resting electrocardiographic results is the Abnormality percentage, followed by the Normal percentage, and the lowest percentage is the Hypertrophy.

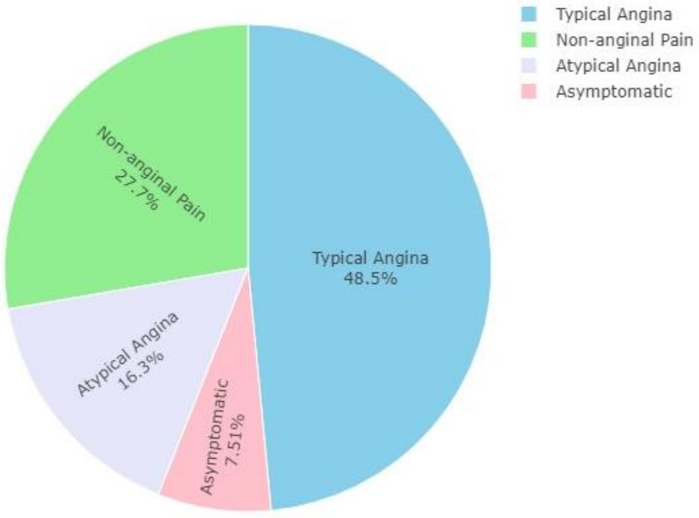


Figure 4.5: Pie chart of Chest Pain Type (Cp)

Figure 4.5 shows that the highest rate in the type of chest pain is Typical angina, followed by non-anginal with a lower rate.

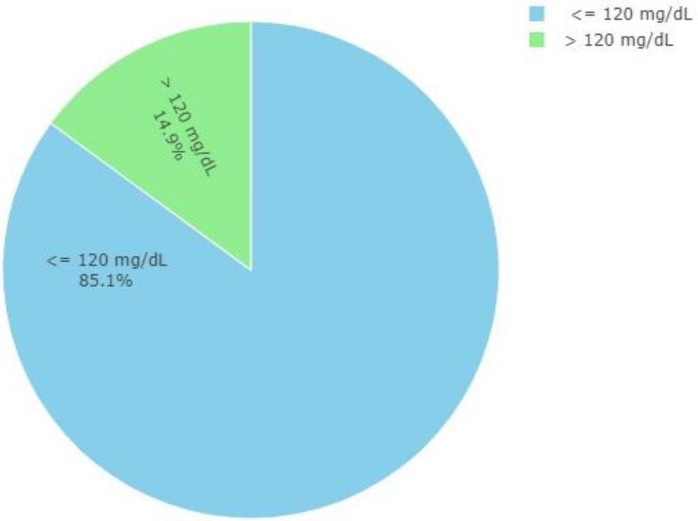


Figure 4.6 Pie chart of Fasting Blood Sugar (Fbs)

Figure 4.6 shows that the highest level of fasting blood sugar is <=120mr/dL.

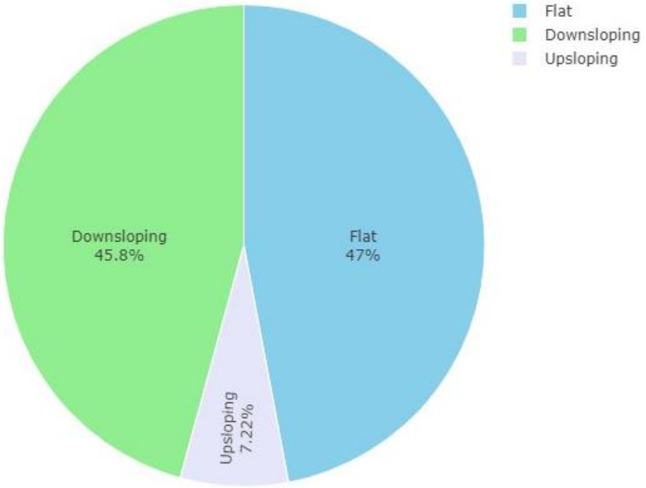


Figure 4.7: Pie chart of Slope of Peak Exercise ST Segment (slope)

Figure 4.7 shows that the highest percentage of peak exercise slope is Flat, followed by Downsloping, and the lowest percentage is Upsloping.

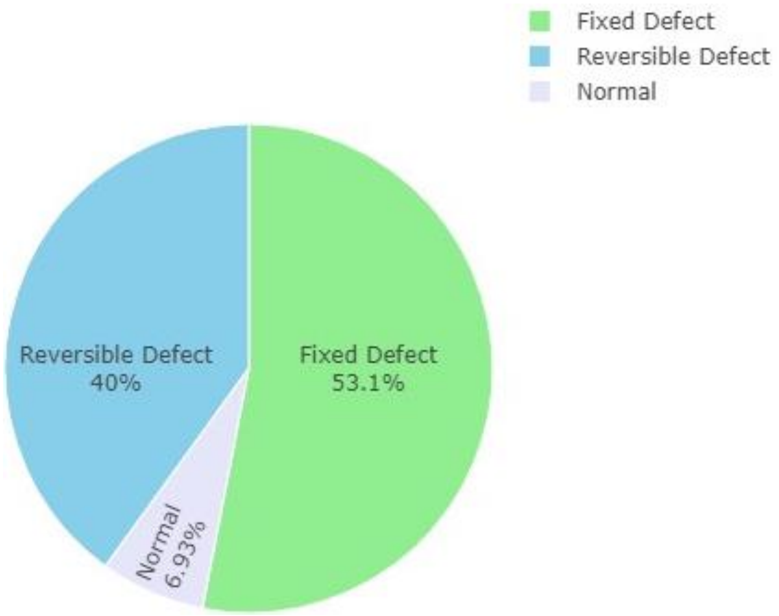


Figure 4.8 Pie chart of Thalassemia (Thal)

Figure 4.8 shows that the highest percentage of thalassemia is Reversible Defect, and the lowest percentage is Normal.

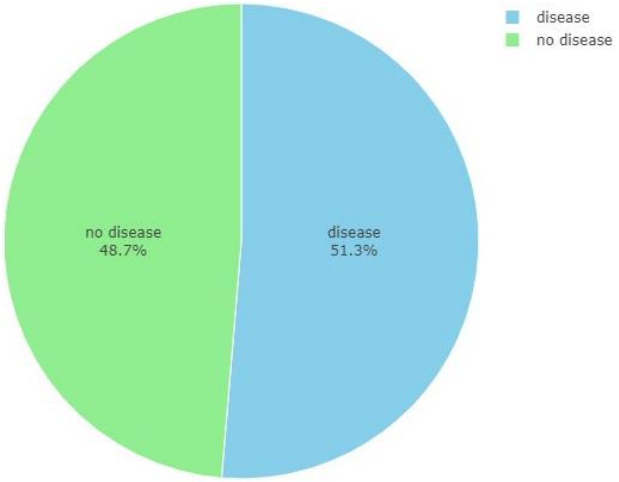


Figure 4.9: Pie chart of Target

Figure 4.9 shows that the incidence of heart attacks is the highest.

4.3 Classification Analysis

In this section, we delve into classification analysis as a distinctive tool for understanding data and identifying influential factors in predictive models. The aim of classification analysis is to segregate data into different groups based on a set of criteria or features, facilitating the understanding of relationships between them and the anticipated outcomes.

4.3.1 Logistic Regression

In this section, we delve into binary logistic regression as a method to identify the optimal variables for our model. Known for its flexibility, logistic regression is favored due to its ability to work without assumptions of normality, making it a versatile tool for predictive modeling.

Variance Inflation Factor (VIF) for the model:

Table 4.2: Multicollinearity in the Mode

Variable	**VIF**	**Interpretation**
Age	1.548982	Moderate
Sex	1.811508	Moderate
Cp	1.990425	Moderate
Trestbps	1.329048	Moderate
Chol	1.346087	Moderate
Fbs	1.179249	Moderate
Restecg	1.167716	Moderate
Thalach	1.530588	Moderate
Exang	1.181619	Moderate
Oldpeak	1.595667	Moderate
Slope	1.948096	Moderate
Ca	2.204531	Moderate
Thal	1.612701	Moderate

Table 4.2 Based on the results, moderate levels of multicollinearity are evident among some variables in the dataset. However, it appears that these issues do not significantly impact the accuracy of the model.

Coefficients Estimate of Logistic Regression (Full Model):

Table 4.3: The coefficients estimate of the logistic regression” Full model.”

Coefficient	**Estimate**	**P-value**
Intercept	1.307375	0.658908
Sex	-2.079658	4.02e-08
Age	0.020154	0.233598
Chol ★	0.005870	0.024778
Thalach ★	0.021131	0.005287
Trestbps	-0.028219	0.0000259
Cp1 ★	0.744279	0.044493
Cp2 ★	2.053837	2.47e-09
Cp3 ★	2.448052	1.75e-07
Fbs	0.226466	0.531451
Restecg1	0.311143	0.231697
Restecg2	-0.692812	0.718926
Exang ★	0.794834	0.006645
Oldpeak	-0.270470	0.080576
Slope1	-0.530112	0.303861
Slope2	1.058052	0.055185
Ca ★	2.423695	7.68e-08
Thal1	2.180048	0.380083
Thal2	1.805452	0.460943
Thal3	0.280387	0.908886

Table 4.3 The findings suggest that variables marked with asterisks in the table exhibit a notable impact on the probability of experiencing a heart attack. The estimated coefficients for each variable indicate a positive association with an increased likelihood of occurrence. This implies that higher values of these variables correspond to a heightened probability.

Predicted Probability:

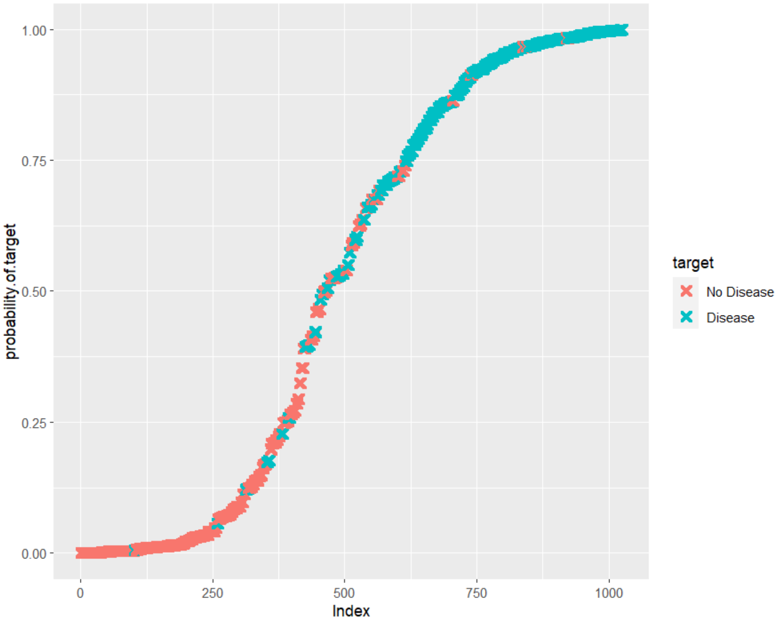


Figure 4.10 Predicted Probability of Heart Disease

This Figure 4.10 illustrates the predicted probability of heart disease, with data distributed based on these probabilities. Blue color denotes the presence of disease, while red represents its absence. The curve initiates from the bottom in red (indicating the absence of disease) and exhibits an overlap with blue (indicating disease) just before the midpoint, suggesting a low likelihood of disease occurrence. Subsequently, the curve ascends and tightens upwards, signifying an increased likelihood of disease. There's a slight overlap between the red and blue colors beyond the midpoint, indicating enhanced confirmation of disease probability over time.

Model Fit Assessment:

Table4.4: Model Goodness of Fit.

Test	**Value**
R^2	0.5680

Table 4.4: The R-squared value of 0.5680 suggests that the model explains approximately 56.80% of the variance in the response variable. While this indicates a moderate level of explanatory power, it's important to consider additional evaluation metrics to comprehensively assess the model's goodness of fit.

Confusion Matrix :

Table 4.5: Confusion Matrix- Logistic Regression

Prediction	Reference	
	No	Yes
No	127	18
Yes	13	150

The confusion matrix in Table 4.5 illustrates the model's performance in classifying cases using test data. The model accurately predicted 127 cases of no disease and 150 cases of disease. However, it misclassified 13 cases of disease as no disease and 18 cases of no disease as disease. These findings affirm the model's effectiveness in discerning between positive and negative cases effectively, demonstrating its ability to distinguish between positive and negative cases using test data.

4.3.2 Naive Bayes Classifier

In this subsection, we discuss the Naive Bayes classifier as a fundamental tool for data classification. The Naive Bayes classifier relies on the application of the principle of simple probabilities and the assumption of independence among variables, making it an intriguing option for analyzing data with multiple attributes.

- **Prior Probability**

Table 4.6: Prior Probability for Naïve Bayes

Prior Probability	
0.5244073	Disease
0.4755927	No Disease

As a result of Table 4.6, in the training data, the probability for class No Disease is 48%, implying that 48% of them have no heart disease, while the probability for class Disease is 52%, implying that 52% of them have heart disease.

- **Categorical Attributes Plot**

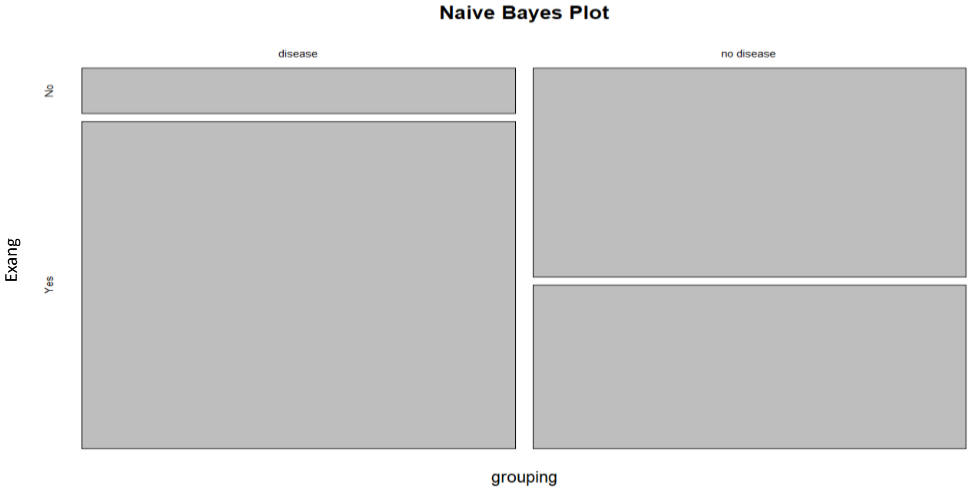


Figure 4.11 Class-Conditional Probability Plots of Exang

Figure 4.11 shows the pain resulting from exercising and its relationship to angina. We conclude that the incidence rate for people who practice sports and feel heart pain is greater than the incidence rate for people who do not feel heart pain when exercising. So, feeling pain during exercise is indicate angina.

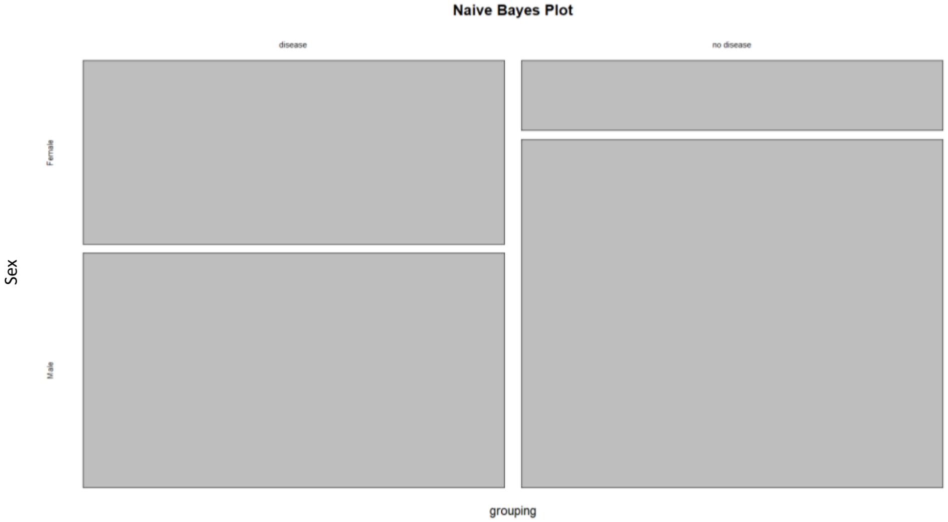


Figure 4.12 Class-Conditional Probability Plots of Sex

As seen in Figure 4.12, which compares disease and non-disease between males and females, we conclude that women are more affected by the disease than men.

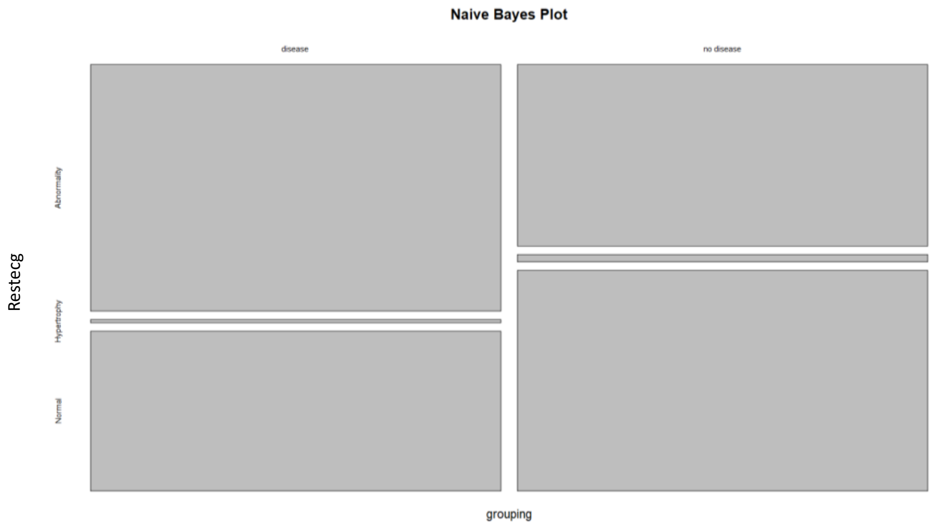


Figure 4.13 Class-Conditional Probabilit Plots of Restecg

As seen in Figure 4.13, which compares normal, abnormality, and hypertrophy, we conclude that abnormality has the highest incidence of heart disease.

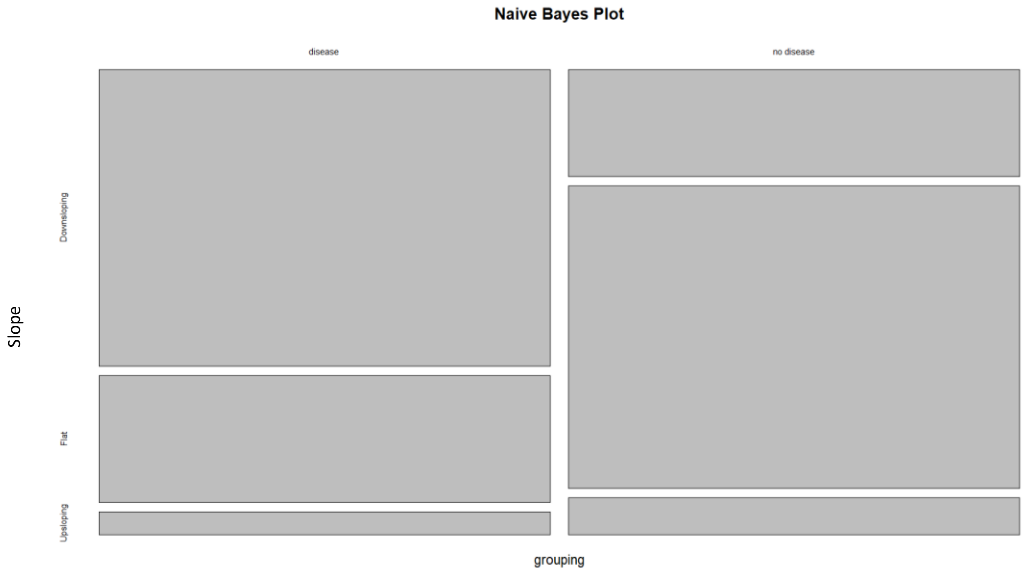


Figure 4.14 Class-Conditional Probability Plots of Slope

As seen in Figure 4.14, comparing upsloping, flat, and downsloping, we conclude that downsloping has the highest risk of heart disease.

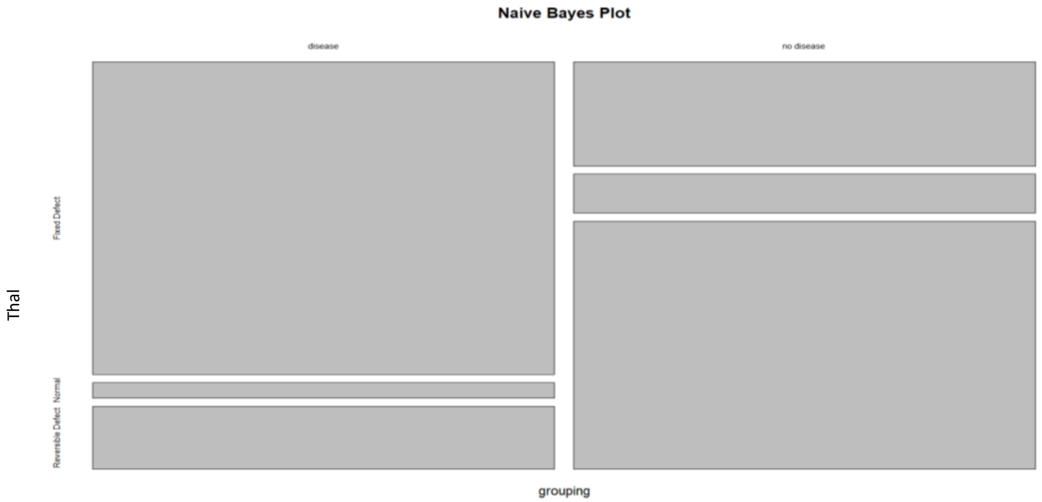


Figure 4.15 Class-Conditional Probabilit Plots of Thal

As seen in Figure 4.15, which compares fixed defect, reversible defect, and normal, we conclude that fixed defect has the highest incidence of heart disease.

- **Continues Variables Plot**

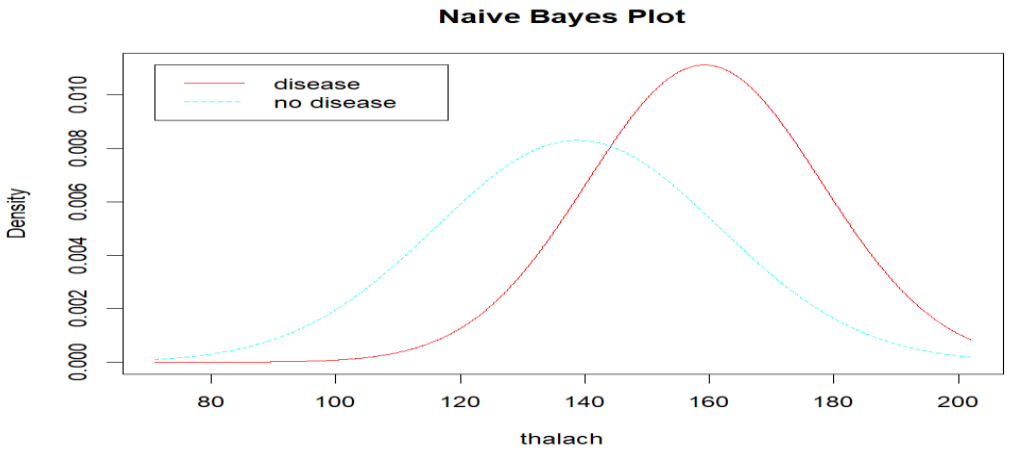


Figure 4.16 Calss-conditional Probability Density plot of Thalach

Figure 4.16 shows a density plot of the maximum heart rate. The average maximum heart rate for patients with heart disease is higher than for those without (159 compared to 139, respectively). However, this does not necessarily mean that a high heart rate is a sign of a heart attack, as there can be other factors that contribute to heart disease.

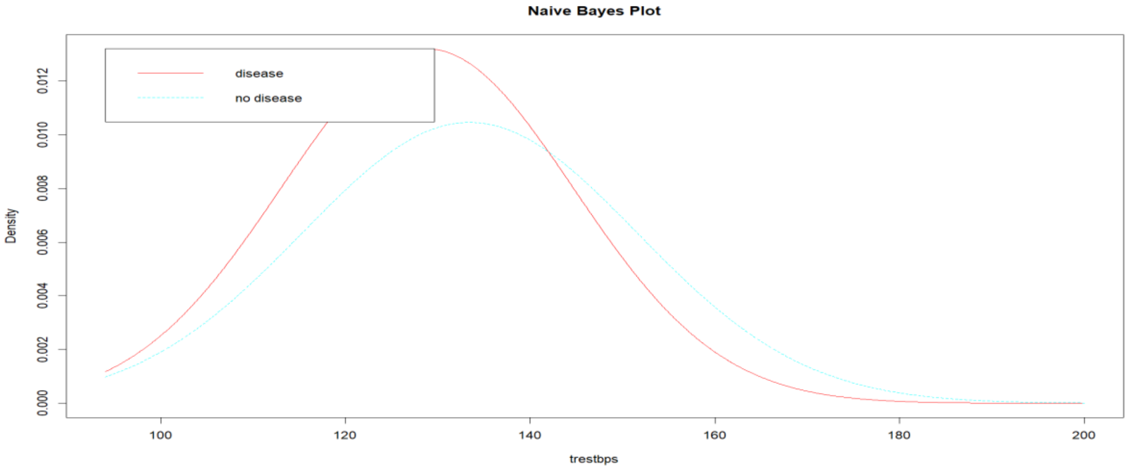


Figure 4.17 Calss-conditional Probability Density plot of Trestbps

Figure 4.17 shows that the Resting Blood Pressure Index is 128.84 for Disease, for No Disease 133.46.

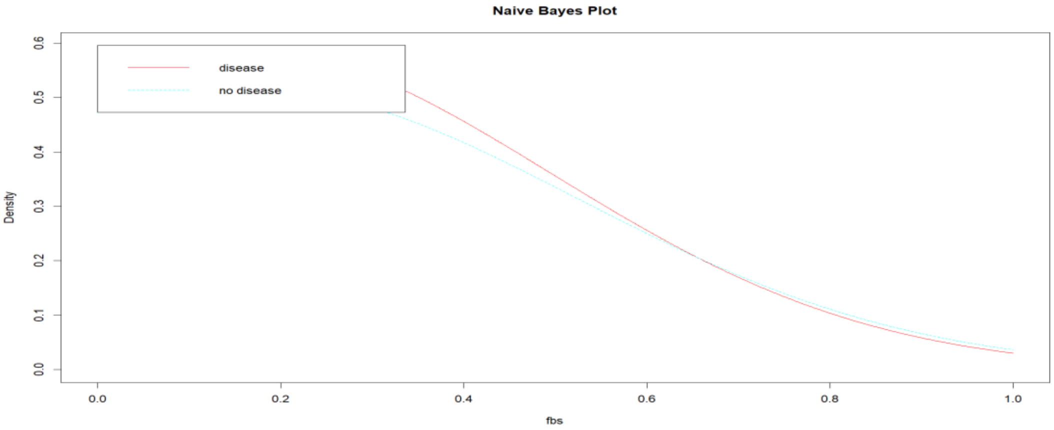


Figure 4.18 Calss-conditional Probability Density plot of Fbs

Figure 4.18 shows the mean of Fasting Blood Sugar are similarly the same for both classes. Therefore, we cannot determine the class of disease and class of no disease.

- We note from the results above that the two variables that most influence the incidence of a heart attack are: Maximum Heart Rate Achieved and Exercise-Induced Angina, because their graphics show a clear classification of disease and no disease.

- **Confusion Matrix**

Table 4.7: Confusion Matrix-Naïve Bayes Model

	Disease	No disease
Disease	120	30
No disease	30	128

The confusion matrix of the Naïve Bayes model is displayed in Table 4.7. The diagonal elements represent the correct classifications made by the model. Among the 150 individuals with heart disease, the model correctly classified 120, but incorrectly classified 30 as not having it. Also, 128 people were correctly classified as not having heart disease, while 30 people were incorrectly classified as having heart disease.

4.4 Classifiers Comparison

After analyzed the data and obtained all the result lastly, we will compare between the models after that, we will select the best models.

Table 4.8: Comparison table between the classifiers

	Logistic Regression	Naïve Bayes
Accuracy	0.9	0.81
Error rate	0.1	0.19
Sensitivity	0.92	0.81
Specificity	0.87	0.8

The Logistic Regression model outperformed the Naïve Bayes model in terms of accuracy, error rate, sensitivity, and specificity. The Logistic Regression achieved an accuracy of 90%, while Naïve Bayes achieved 81%. The error rate for Logistic Regression was 10%, compared to 19% for Naïve Bayes. The sensitivity of the Logistic Regression model was 92%, while Naïve Bayes had a sensitivity of 81%. Furthermore, the specificity of Logistic Regression was 87%, whereas Naïve Bayes demonstrated a specificity of 80%. These metrics indicate that the Logistic Regression model is a superior classification method for this balanced dataset.

In light of the balanced dataset, it's essential to consider both sensitivity and specificity. Sensitivity measures the ability to correctly identify positive events, while specificity measures the ability to correctly identify negative events.

The logistic regression model excels in both sensitivity and specificity, with values of 0.92 and 0.87, respectively. This indicates the model's robust capability to identify individuals experiencing a heart attack and accurately classify those not experiencing a heart attack. Although the Naïve Bayes method also exhibits good sensitivity (0.81), its specificity (0.80) is slightly lower compared to the logistic regression model. Nonetheless, both models effectively identify cases of heart attacks and non-heart attacks in this balanced dataset.

ROC curve and AUC

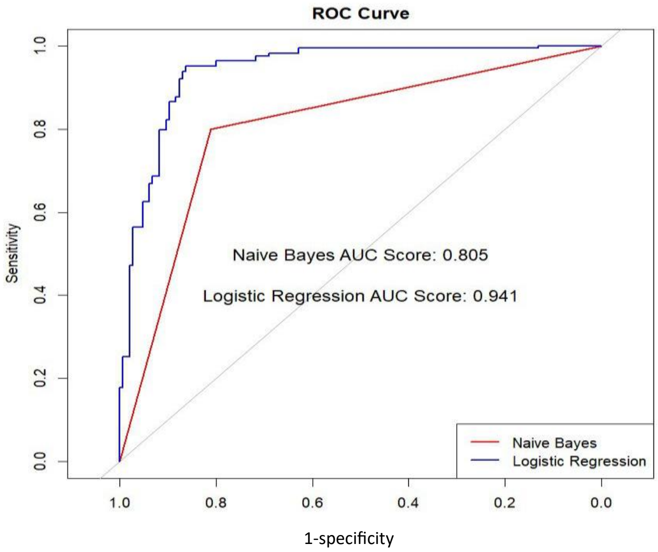


Figure 4.19 Comparison between the classifiers by ROC curve

The high AUC values, such as that achieved by logistic regression (0.94), indicate its strong ability to effectively distinguish between positive and negative cases. Conversely, the lower AUC value for Naïve Bayes (0.81) suggests a weaker discriminatory ability. This difference underscores the importance of selecting the appropriate model, as logistic regression can provide superior predictive power for tasks such as predicting heart attacks.

Chapter 5: Summary and Future Work

5.1 Introduction

This chapter provides a comprehensive overview of the scientific study conducted on predicting myocardial infarction using logistic regression and naive Bayes models. It summarizes the main results, discusses the implications of the study's findings, and proposes areas for future research in the field of cardiology.

5.2 Summary and Conclusion

In conclusion, the aim of this study was to identify the most impactful symptoms of heart attack and determine the best statistical models for its accurate diagnosis. We compared two statistical models, logistic regression, and Naive Bayes, using data related to predicting heart attacks. The performance of these two models was evaluated using five statistical measures, including classification accuracy, error rate, sensitivity, specificity, and the area under the curve (AUC). The results indicate that logistic regression outperformed Naive Bayes. Common variables contributing to predicting heart attacks were identified as maximum heart rate achieved, chest pain type, serum cholesterol levels, and exercise-induced chest pain, number of major vessels colored by fluoroscopy.

5.3 Recommendations and Future Studies

In the future, we may evaluate the findings using other methods such as support decision tree model, random forests, etc., or we could add new variables to the model that could affect the responses. This research has shown us that studies on this topic are limited to specific nations; our community is not represented in them. We can specifically focus on the Saudi community in the study. In order to include a wider swath of society, we might potentially augment the quantity of data.

Reference

.

[1] Khan, A., Zaidi, S., & Ahmad, T. (2019). Risk stratification using decision trees for liver disease patients. Clinical Epidemiology and Global Health, 7(3), 205-21

[2] Hasan, S. M. M., Mamun, M. A., Uddin, M. P., & Hossain, M. A. (2018). International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 1-4, 2018.

[3] Alishiri, G. H., Bayat, N., Ashtiani, A. F., Tavallaii, S. A., Assari, S., & Moharamzad, Y. (2008). Logistic regression models for predicting physical and mental health-related quality of life in rheumatoid arthritis patients. Modern Rheumatology, 18(6), 601–608.

[4] Manikandan, S. (2017, August). *Heart attack prediction system*.

[5] Song, X., Liu, X., Liu, F., & Wang, C. (2021). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. International Journal of Medical Informatics, 151, 104484.

[6] Kamber, M., & Han, P. J. (2012). *Data Mining Concepts and Techniques* (3rd ed.).

[7] Stoltzfus, J. C. (2011). Logistic Regression: A brief primer. Academic Emergency Medicine, 18(10), 1099–1104.

[8] Aggarwal, C. C. (2015). Data mining: the textbook. Springer, New York, USA.

[9] Webb, G. I. (2016). Naïve Bayes.

[10] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning with Applications in R. Springer New York Heidelberg Dordrecht London

[11] Krzanowski, W. J. and Hand, D. J. (2009). ROC curves for continuous data. Crc Press

[12] sai Liang. Confusion matrix: Machine learning. POGIL Activity Clear⍰inghouse, 3(4), 2022.