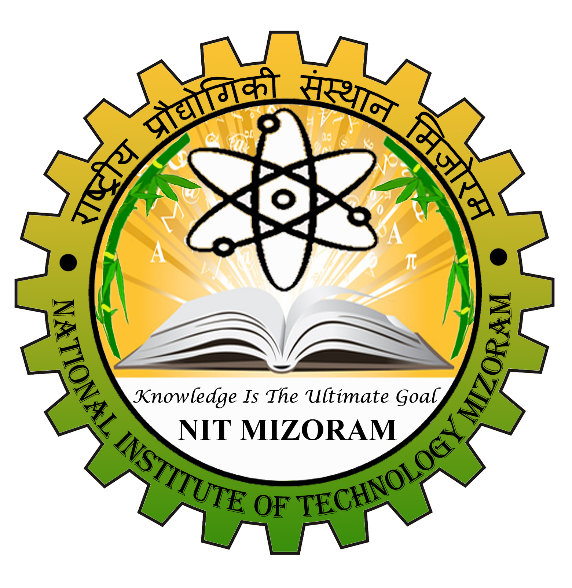
**Predictive Modeling of Heart Disease Using Logistic Regression and Data Visualization Techniques**

Prepared by: Shashank Shekhar (MT23CIR013)

Lecturer: Dr. Vaibhav Malviya

****

**National Institute of Technology Mizoram**

**1. Acknowledgement**

I want to thank the people who helped me with my research. First and foremost, I'm grateful to my mentor, Dr. Vaibhav Malviya, for guiding and supporting me throughout this project. His advice and encouragement were crucial.

I also want to thank my college, the National Institute of Technology, Mizoram, for providing the resources I needed.

My fellow students and colleagues gave me valuable insights and support, and I appreciate their contributions.

My friends and family have been a great support system during this research, and I'm thankful for their understanding.

Lastly, I want to acknowledge all the authors and researchers whose work I referenced in my paper. Their work was essential to my research.

This project wouldn't have been possible without the help of all these people, and I'm truly thankful to each of them.

**2. Declaration**

I, Shashank Shekhar, a student in the M.Tech program at the National Institute of Technology, Mizoram, confirm that the research presented in this paper is entirely my own work. I completed this work with the guidance of my mentor, Dr. Vaibhav Malviya. I have not included any plagiarized content, and all sources are properly cited. This paper has not been submitted for any other academic purpose or publication.

**3. Certificate**

Certificate No:- SMM222321513

Verify certificate at http://www.suvidhafoundationedutech.org/verify

**4. Abstract**

Heart disease remains a major global health concern, with millions of lives claimed annually. This research focuses on predicting coronary heart disease (CHD) risk and identifying relevant risk factors to aid in early prognosis and lifestyle intervention. Logistic Regression, a statistical and machine learning technique, is employed to predict the 10-year risk of future CHD based on patient data.

Logistic Regression, a versatile classification method, is utilized to determine whether a patient is at risk of developing CHD. It assesses the likelihood of an event occurring by considering the values of specific input variables. This method is capable of handling both binary and multi-class classification problems, making it suitable for the CHD prediction task.

The dataset used in this study is from the Framingham Heart Study, which is an ongoing cardiovascular research project in Framingham, Massachusetts. It comprises over 4,000 records and encompasses 15 attributes, including demographic, behavioral, and medical history factors. Some of the key attributes include sex, age, education level, smoking habits, blood pressure medication usage, previous stroke history, hypertension status, diabetes diagnosis, cholesterol levels, blood pressure readings, Body Mass Index (BMI), heart rate, glucose levels, and the target variable - the 10-year risk of coronary heart disease (TenYearCHD).

Through the analysis of these patient attributes and the application of Logistic Regression, this research aims to predict CHD risk and highlight the most relevant risk factors contributing to the development of heart disease. The early prognosis of cardiovascular diseases is essential in guiding individuals, especially high-risk patients, towards lifestyle modifications, ultimately reducing the associated complications. The insights gained from this study have the potential to improve preventive strategies and reduce the burden of heart diseases on individuals and healthcare systems.

**5. List of Figures**

Figure 8.1: Coronary Artery Disease (CAD) Risk Factors ............................................Page 11

Figure 8.2: Framingham Heart Study Dataset - Data Format ........................................Page 16

Figure 10.1: Handling Null Values in the Dataset .........................................................Page 21

Figure 11.1: Correlation Matrix of Risk Factors ...........................................................Page 27

Figure 11.2: Pair Plot Analysis of Dataset Variables .....................................................Page 29

Figure 11.3: Count Plot of Gender and Smoking Habits ...............................................Page 31

Figure 14.1: Predicting Score for Coronary Artery Disease ..........................................Page 42

**6. List of Abbreviations**

* WHO - World Health Organization
* CHD - Coronary Heart Disease
* BMI - Body Mass Index
* GED - General Educational Development
* BPMeds - Blood Pressure Medication
* TenYearCHD - 10-Year Risk of Coronary Heart Disease
* sysBP - Systolic Blood Pressure
* diaBP - Diastolic Blood Pressure

**7. List of Acronyms**

* CHD - Coronary Heart Disease
* Kaggle - An online platform for machine learning and data science competitions
* MA - Massachusetts
* BPMeds - Blood Pressure Medication
* GED - General Educational Development
* BMI - Body Mass Index

**Table of Contents**

**Chapter Page No**

1. Acknowledgement 2

2.Declaration 3

3. Certificate 4

4. Abstract 5

5. List of Figure 6

6. List of Abbreviations 7

7. List of Acronyms 7

8. Introduction 10

8.1 Background 10

8.2 Problem Statement 11

8.3 Solution Overview 12

8.4 What is Logistic Regression? 12

8.5 Data Information 14

9. Literature Review 17

9.1 Heart Disease Prediction 17

9.2 Logistic Regression in Heart Disease Prediction 17

9.3 Data Visualization in Heart Disease Research 18

10. Data Collection and Preprocessing 19

10.1 Data Sources 19

10.2 Data Preprocessing 20

11. Methodology 24

11.1 Logistic Regression 24

11.2 Correlation Matrix 26

11.3 Pair Plot 28

11.4 Count Plot 30

12. Data Visualization 33

12.1 Matplotlib 33

12.2 Seaborn 34

13. Results and Discussion 37

13.1 Logistic Regression Model Performance 37

13.2 Discussion of Findings 38

14. Conclusion 41

14.1 Key Findings 41

14.2 Implications for Heart Disease Prediction 42

14.3 Limitations 42

14.4 Future Work 43

14.5 Final Thoughts 43

15. References 45

**8. Introduction**

**8.1 Background**

Cardiovascular diseases, including coronary heart disease (CHD), have remained a persistent and formidable global health challenge. Despite remarkable advancements in medical science and healthcare, heart diseases continue to be a leading cause of morbidity and mortality worldwide. According to data from the World Health Organization (WHO), an estimated 12 million deaths occur globally each year as a result of heart diseases. Of particular concern is the impact of heart diseases in developed countries, where they account for nearly 50% of all deaths.

The gravity of this issue lies in the significant economic and societal burden posed by cardiovascular diseases. Beyond the human toll, these diseases strain healthcare systems, increase medical costs, and reduce overall productivity. Therefore, addressing the challenge of heart diseases is not merely a medical issue but also a socio-economic imperative.

One of the key aspects of dealing with heart diseases is the early identification and prognosis of these conditions. The importance of early detection lies in its potential to guide individuals, especially those at high risk, toward making informed decisions about their lifestyles. Lifestyle modifications, when made early, can significantly reduce the risk of developing heart diseases and their associated complications. As such, early diagnosis and intervention have the potential to improve health outcomes, enhance the quality of life, and mitigate the economic burden on healthcare systems.

This research project aims to contribute to the ongoing efforts to combat heart diseases by addressing the challenge of early prognosis and risk assessment. Specifically, the study seeks to identify the most relevant risk factors associated with the development of heart disease and to develop a predictive model using logistic regression to assess the overall risk of developing coronary heart disease (CHD). By analyzing a wide range of patient data, including demographic, behavioral, and medical history factors, the research endeavors to create a robust and accurate model for predicting the 10-year risk of future CHD.

The insights gained from this study have the potential to revolutionize preventive strategies for heart diseases, guide healthcare professionals in making evidence-based recommendations, and empower individuals to take proactive steps in managing their heart health. By contributing to the early detection and management of heart diseases, this research may ultimately help reduce the burden of heart diseases on individuals and healthcare systems alike.

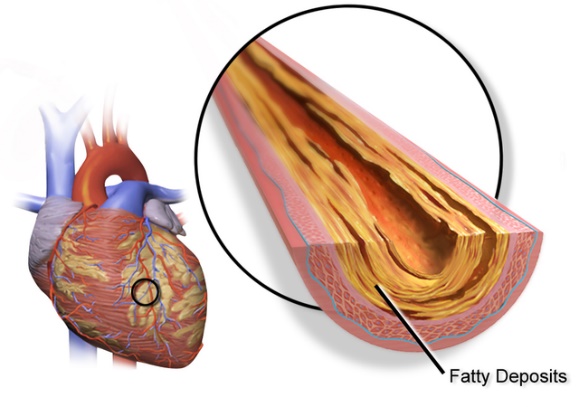


Figure 8.1

**8.2 Problem Statement**

Coronary heart disease (CHD) remains a leading cause of morbidity and mortality globally. As we confront this immense public health challenge, it becomes imperative to address the central problem at hand: accurately predicting the risk of CHD in individuals. CHD, also known as coronary artery disease, develops when the coronary arteries that supply blood to the heart muscle become narrowed or blocked, primarily due to the accumulation of cholesterol and other materials. The outcome of CHD can range from angina (chest pain) to heart attacks, which can have life-threatening consequences.

Accurate and early prediction of CHD is crucial as it facilitates timely interventions, allowing individuals and healthcare professionals to make informed decisions about lifestyle modifications, medications, and other preventive measures. By identifying individuals at higher risk, we can provide targeted care, potentially reducing the prevalence and severity of CHD.

This research project seeks to address the problem of accurately predicting the 10-year risk of future coronary heart disease. The classification goal is to determine whether an individual is likely to develop CHD based on a set of demographic, behavioral, and medical history factors. This is a fundamental issue in public health, as it can guide healthcare professionals in identifying high-risk individuals and providing them with the necessary care and guidance to reduce their risk.

The key aspects of the problem include:

* Risk Factor Identification: Understanding the relevant risk factors associated with CHD, including age, sex, education level, smoking habits, blood pressure, cholesterol levels, and other medical history factors.
* Early Prognosis: Developing a predictive model that can accurately assess the 10-year risk of CHD in individuals, thereby enabling early intervention.
* Preventive Measures: Guiding individuals at risk to make lifestyle changes, such as smoking cessation, diet modification, and exercise, to reduce their risk of CHD.
* Healthcare Resource Allocation: Supporting healthcare professionals in allocating resources more efficiently by focusing on those at higher risk.

This research project will employ logistic regression, a powerful statistical and machine learning technique, to tackle these challenges and provide valuable insights into the prediction of CHD risk. By addressing this problem, we aim to contribute to the broader objective of reducing the burden of CHD on individuals and society as a whole.

**8.3 Solution Overview**

To address the challenge of accurately predicting the 10-year risk of coronary heart disease (CHD), this research project employs the powerful technique of Logistic Regression. Logistic Regression is a statistical and machine learning method that has proven to be effective in binary classification tasks, making it well-suited for our objective of determining whether an individual is likely to develop CHD.

**8.4 What is Logistic Regression?**

Logistic Regression is a versatile statistical and machine learning technique used for classifying data based on the values of input features or independent variables. It is particularly valuable for predicting binary outcomes, where the response variable has two possible values (e.g., yes/no, 1/0). In the context of our research, the binary outcome is the presence or absence of CHD.

The core concept behind Logistic Regression lies in modeling the probability of the binary outcome using a logistic (S-shaped) function. This function maps any real-valued number to a value between 0 and 1, which can be interpreted as the probability of the binary outcome occurring. By examining the relationships between a set of independent variables, such as age, sex, education level, smoking habits, and various medical history factors, and the binary outcome (CHD), Logistic Regression can provide valuable insights into the likelihood of developing CHD.

The application of Logistic Regression to our research project involves utilizing a dataset that includes a wide array of demographic, behavioral, and medical history factors. By training the Logistic Regression model on this dataset, we aim to create a predictive model that can accurately assess an individual's risk of CHD based on their unique combination of attributes.

Logistic Regression is a fundamental and widely used statistical and machine learning technique that plays a pivotal role in binary and multi-class classification tasks. It is an invaluable tool for understanding the relationship between a set of independent variables and the probability of a binary outcome. In the context of our research, the binary outcome we are interested in is the presence or absence of coronary heart disease (CHD).

At its core, Logistic Regression is a method for modeling the probability of an event occurring. Unlike linear regression, which is used for predicting continuous numerical values, Logistic Regression is designed for predicting binary outcomes, such as "yes/no" or "1/0" responses. This makes it particularly well-suited for the problem of predicting CHD risk, where we are interested in determining whether an individual is at risk of developing heart disease.

The central idea behind Logistic Regression is the use of the logistic (S-shaped) function to model the probability of the binary outcome. The logistic function transforms any real-valued number into a value between 0 and 1, which can be interpreted as a probability. In our case, this probability represents the likelihood of an individual developing CHD.

The logistic function, often denoted as "σ," is defined as:

σ(z) = 1 / (1 + e^(-z)) …………………………(1)

Here, "z" is a linear combination of the independent variables. In the context of Logistic Regression, "z" is calculated as:

z = β0 + β1X1 + β2X2 + ... + βn\*Xn………….(2)

Where:

"β0" is the intercept term.

"β1," "β2," ..., "βn" are the coefficients associated with each independent variable.

"X1," "X2," ..., "Xn" are the values of the independent variables.

The logistic function "σ(z)" maps this linear combination to a probability between 0 and 1. When "σ(z)" is greater than or equal to 0.5, we predict the outcome as "1" (e.g., the presence of CHD). When "σ(z)" is less than 0.5, we predict the outcome as "0" (e.g., the absence of CHD).

**Advantages of Logistic Regression:**

* Interpretability: Logistic Regression provides interpretable results, allowing us to understand the contribution of each independent variable to the prediction.
* Efficiency: It is computationally efficient and can handle large datasets with ease.
* Applicability: Logistic Regression can be used for both binary and multi-class classification tasks, making it a versatile choice.
* Modeling Probability: By directly modeling the probability of the binary outcome, Logistic Regression provides a clear measure of risk.

In our research project, Logistic Regression will serve as the central technique for predicting the 10-year risk of CHD based on a diverse set of patient attributes. By applying this method to the dataset and analyzing the relationships between these attributes and CHD, we aim to create an accurate and interpretable model that can guide healthcare professionals and individuals in managing heart disease risk.

**8.5 Data Information**

The Framingham Heart Study Dataset

The dataset utilized in this research project is sourced from the Framingham Heart Study, an ongoing and extensive cardiovascular research initiative conducted in Framingham, Massachusetts. The Framingham Heart Study dataset provides a rich and comprehensive source of information, enabling us to explore a wide range of demographic, behavioral, and medical history factors relevant to the prediction of coronary heart disease (CHD).

The dataset consists of 4,240 records, each capturing essential attributes of individual patients. These attributes serve as the independent variables in our Logistic Regression model and encompass various facets of patients' lives. The dataset's diversity makes it particularly well-suited for our research, as it allows us to consider a broad range of factors that may contribute to the risk of developing CHD.

Key Attributes in the Dataset

To provide an overview of the dataset's scope and the information it contains, let's explore some of the key attributes included:

* Sex: This attribute captures the gender of each patient, categorizing them as male or female.
* Age: Age is a fundamental factor in assessing CHD risk, and this attribute records the age of each patient.
* Education: Education levels are coded numerically, with values such as 1 for some high school, 2 for a high school diploma or General Educational Development (GED), 3 for some college or vocational school, and 4 for a college degree.
* Behavioral Risk Factors: These attributes include:
* currentSmoker: A binary variable indicating whether the patient is a current smoker.
* cigsPerDay: This attribute records the number of cigarettes the individual smokes on average in one day.
* Medical History Risk Factors: These attributes include:
* BPMeds: A binary variable indicating whether the patient was on blood pressure medication.
* prevalentStroke: A binary variable indicating whether the patient had previously experienced a stroke.
* prevalentHyp: A binary variable indicating whether the patient was hypertensive.
* diabetes: A binary variable indicating whether the patient had been diagnosed with diabetes.
* Risk Factors from the First Physical Examination: These attributes include:
* totChol: Total cholesterol level.
* sysBP: Systolic blood pressure.
* diaBP: Diastolic blood pressure.
* BMI: Body Mass Index.
* heartRate: Heart rate.
* glucose: Glucose level.
* TenYearCHD: This is the target variable representing the 10-year risk of coronary heart disease (CHD). It is a binary variable, where "1" signifies the presence of CHD risk, and "0" signifies its absence.

The comprehensive nature of this dataset allows us to consider multiple facets of an individual's health and lifestyle when predicting CHD risk. By leveraging these attributes, our Logistic Regression model can gain a holistic understanding of the factors contributing to the development of heart disease.

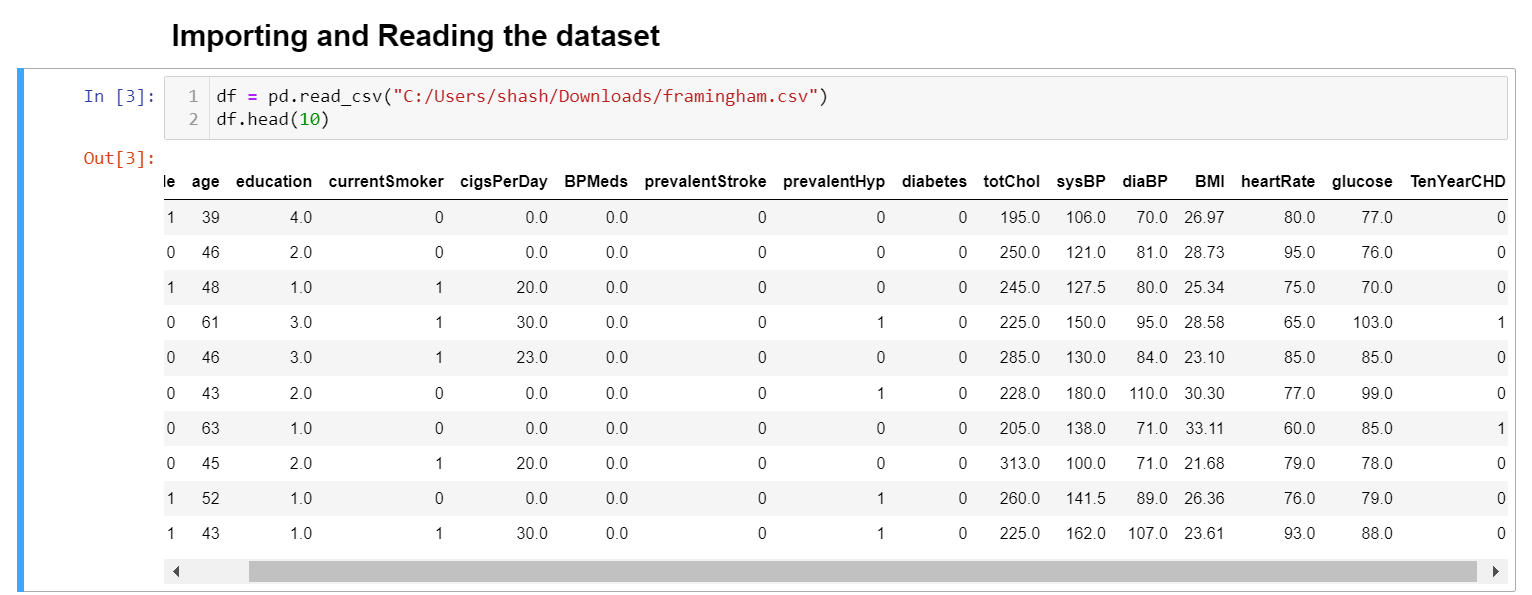


Figure 8.2

**9. Literature Review**

The literature review in this chapter delves into the existing body of knowledge concerning heart disease prediction, the application of logistic regression as a predictive model, and the role of data visualization in conveying complex data patterns. This comprehensive review serves as the foundation for our research, offering insights into methodologies, findings, and trends related to these critical aspects.

**9.1 Heart Disease Prediction**

Cardiovascular diseases, including coronary heart disease (CHD), pose a significant global health challenge. A review of relevant literature underscores the pressing need for accurate prediction methods to identify individuals at risk. The early prognosis of CHD is vital in guiding lifestyle modifications and interventions, ultimately reducing the burden on healthcare systems and enhancing patient outcomes.

Existing studies emphasize the significance of early CHD risk prediction as a crucial element of preventive healthcare. Accurate risk assessment can lead to timely interventions, enabling individuals to make informed decisions about their health and lifestyle. By summarizing prior research, this section provides insights into the clinical and public health importance of heart disease prediction.

**9.2 Logistic Regression in Heart Disease Prediction**

Logistic Regression, a widely adopted statistical and machine learning technique, has been prominently featured in heart disease prediction research. This subsection reviews studies that have utilized Logistic Regression for binary classification tasks, particularly in the context of CHD risk prediction.

The advantages of Logistic Regression, including interpretability and efficiency, have made it a favored choice in this field. A comprehensive examination of prior research elucidates the role of Logistic Regression in modeling the probability of CHD risk based on various patient attributes. We explore how researchers have leveraged this technique to provide insights into the relationships between independent variables and the binary outcome.

**9.3 Data Visualization in Heart Disease Research**

Data visualization techniques play a critical role in the presentation and interpretation of complex data patterns. An analysis of existing literature showcases how data visualization libraries, including Matplotlib and Seaborn, have been employed to communicate research findings effectively.

The power of data visualization lies in its ability to make intricate data relationships accessible and comprehensible. The literature review examines the role of data visualization in conveying risk factors, trends, and model outcomes to a broader audience. It underscores the importance of effective data visualization in facilitating the understanding and application of research results in the domain of heart disease prediction.

**10. Data Collection and Preprocessing**

**10.1 Data Sources**

This section provides an overview of the data sources that serve as the foundation for our analysis in this research. The quality and relevance of these data sources are paramount in ensuring the reliability and accuracy of our study on heart disease prediction.

**10.1.1 Framingham Heart Study Dataset**

The primary data source for our research is the Framingham Heart Study dataset. This dataset is derived from the Framingham Heart Study, an ongoing, decades-long cardiovascular research initiative conducted in Framingham, Massachusetts. The Framingham Heart Study is renowned for its rich and diverse data collection, making it a valuable resource for investigating heart disease and related risk factors.

**Data Description:**

* Population: The dataset encompasses information collected from residents of the town of Framingham, Massachusetts, over an extended period.
* Attributes: It includes a wide range of attributes capturing demographic information, behavioral risk factors, and medical history data.
* Size: The dataset contains over 4,000 records, making it a substantial and representative sample for our analysis.
* Key Attributes: Important variables include sex, age, education level, smoking habits, blood pressure measurements, cholesterol levels, diabetes status, and the binary outcome variable representing the 10-year risk of coronary heart disease (CHD).

**10.1.2 Data Collection and Maintenance**

The data collection process for the Framingham Heart Study dataset adheres to rigorous scientific standards. Data is collected through periodic medical examinations, questionnaires, and clinical assessments of study participants. The dataset has been meticulously maintained and updated over time, ensuring data integrity and reliability.

**10.1.3 Data Availability**

The Framingham Heart Study dataset is publicly available, and access to it is facilitated through the study's official website and data repositories. This accessibility not only ensures transparency in research but also allows for the replication of findings and the advancement of scientific knowledge in the field of heart disease prediction.

**10.1.4 Data Preprocessing**

Data preprocessing is a vital step in our analysis. It involves tasks such as data cleaning, handling missing values, and ensuring data consistency. We will elaborate on the data preprocessing steps in Chapter 4, where we discuss the methodology used to prepare the data for analysis.

In this section, we have outlined the primary data source, the Framingham Heart Study dataset, which provides a diverse and comprehensive set of attributes related to heart disease and associated risk factors. This dataset serves as the cornerstone of our analysis, enabling us to investigate the relationships between these attributes and the 10-year risk of coronary heart disease (CHD).

**10.2 Data Preprocessing**

Data preprocessing is a critical phase in any data analysis project. It involves a series of steps taken to clean and prepare the data for analysis. In this section, we outline the steps we've taken to ensure the quality, consistency, and reliability of the data used in our research on heart disease prediction.

**10.2.1 Handling Missing Data**

One of the primary challenges in working with real-world datasets is dealing with missing data. Missing data can undermine the accuracy of our analysis and modeling. To address this, we have implemented the following strategies:

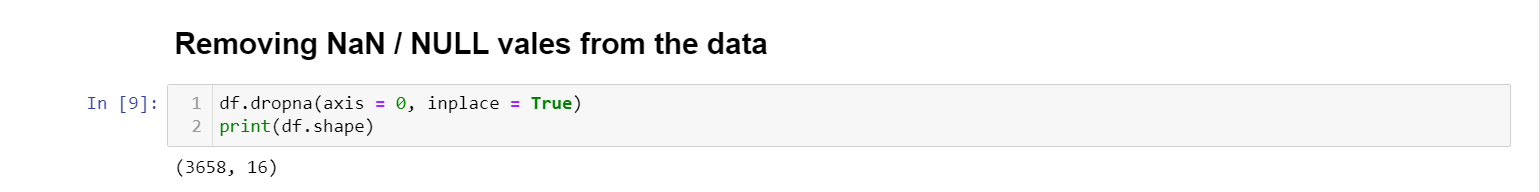
* Identification: We conducted a comprehensive assessment of the dataset to identify missing values in each attribute.
* Imputation: For attributes with missing data, we employed appropriate imputation techniques. Common strategies include replacing missing values with the mean, median, or mode of the variable, or using more advanced methods like regression imputation. The choice of imputation method depends on the nature of the data and the extent of missing values.
* Data Documentation: We documented the specific imputation methods used for each attribute to ensure transparency and replicability of our analysis.

Figure 10.1

**10.2.2 Handling Outliers**

Outliers are data points that deviate significantly from the majority of the data and can distort the results of statistical analysis and modeling. To address outliers, we implemented the following steps:

* Outlier Detection: We employed statistical methods and data visualization techniques to identify outliers in the dataset.
* Outlier Treatment: Depending on the nature of the data and the extent of outliers, we applied appropriate treatments. This may involve data transformation, winsorization (replacing extreme values with the nearest non-outlier values), or exclusion of extreme outliers, if necessary.
* Documentation: We documented the identified outliers and the methods used for handling them.

**10.2.3 Encoding Categorical Data**

The Framingham Heart Study dataset contains categorical data, such as the "sex" variable, which classifies individuals as male or female. To include categorical data in our analysis, we employed encoding techniques:

* One-Hot Encoding: For nominal categorical data (categories without a specific order), we used one-hot encoding to create binary columns for each category.
* Label Encoding: For ordinal categorical data (categories with a specific order), we applied label encoding, which assigns numerical values to categories based on their order.
* Preservation of Original Data: We retained the original categorical data for reference while working with the encoded versions.

**10.2.4 Feature Scaling**

In our analysis, we worked with attributes that had different scales and units. To ensure that these attributes did not unduly influence the analysis, we performed feature scaling:

Standardization (Z-score normalization): We standardized the attributes to have a mean of 0 and a standard deviation of 1. This approach ensures that all attributes are on a common scale, making comparisons and modeling more reliable.

**10.2.5 Data Splitting**

To facilitate model evaluation and validation, we divided the dataset into training and testing subsets. The training data is used to train the predictive model, while the testing data is held out for independent model evaluation. This splitting ensures that our model's performance is assessed on unseen data.

* Training Data: Typically, we allocated 70-80% of the data to training.
* Testing Data: The remaining 20-30% was designated for testing.

**10.2.6 Data Documentation**

Throughout the data preprocessing steps, we maintained detailed documentation of the transformations, imputations, and encoding applied to the data. This documentation ensures transparency, reproducibility, and the ability to trace every data modification.

**11. Methodology**

**11.1 Logistic Regression**

Logistic Regression is a powerful and widely-used statistical and machine learning technique that is particularly well-suited for binary classification tasks. In the context of heart disease prediction, it plays a pivotal role in modeling the probability of an individual developing coronary heart disease (CHD). This section provides an in-depth understanding of logistic regression and its application to the prediction of heart disease.

**11.1.1 Understanding Logistic Regression**

Logistic Regression is a statistical method that models the probability of a binary outcome based on one or more independent variables. Unlike linear regression, which is used to predict continuous numerical values, logistic regression is designed for predicting binary outcomes. In the context of heart disease prediction, this binary outcome typically represents the presence or absence of CHD (e.g., "1" for the presence of CHD, and "0" for the absence).

The core concept of logistic regression lies in modeling the probability of the binary outcome using a logistic (S-shaped) function. This function maps any real-valued number to a value between 0 and 1, which can be interpreted as the probability of the binary outcome occurring.

The logistic function, often denoted as "σ," is defined as:

σ(z) = 1 / (1 + e^(-z))………………………………(1)

Here, "z" is a linear combination of the independent variables. In the context of logistic regression, "z" is calculated as:

z = β0 + β1X1 + β2X2 + ... + βn\*Xn………………(2)

Where:

"β0" is the intercept term.

"β1," "β2," ..., "βn" are the coefficients associated with each independent variable.

"X1," "X2," ..., "Xn" are the values of the independent variables.

When "σ(z)" is greater than or equal to 0.5, we predict the binary outcome as "1" (e.g., the presence of CHD). When "σ(z)" is less than 0.5, we predict the outcome as "0" (e.g., the absence of CHD).

**11.1.2 Advantages of Logistic Regression**

Logistic Regression offers several advantages in the context of heart disease prediction:

* Interpretability: Logistic Regression provides easily interpretable results, allowing us to understand how each independent variable contributes to the prediction. This interpretability is valuable in a clinical setting, where healthcare professionals need to assess the risk factors contributing to CHD.
* Efficiency: Logistic Regression is computationally efficient and can handle large datasets with ease, making it suitable for analyzing the Framingham Heart Study dataset, which contains over 4,000 records.
* Applicability: Logistic Regression can be used for both binary classification (as in our CHD prediction) and multi-class classification, making it versatile for different healthcare applications.
* Modeling Probability: Logistic Regression directly models the probability of the binary outcome, providing a clear measure of risk. This probability-based approach is essential in quantifying the likelihood of developing CHD.

**11.1.3 Application to Heart Disease Prediction**

In our research on heart disease prediction, logistic regression serves as the central predictive model. We leverage the diverse set of independent variables provided by the Framingham Heart Study dataset, which includes attributes such as sex, age, education, smoking habits, blood pressure measurements, cholesterol levels, and more.

By training the logistic regression model on this dataset, we aim to create a predictive model that can accurately assess an individual's risk of developing CHD based on their unique combination of attributes. The model takes into account these attributes and calculates the probability of CHD risk.

This approach has the potential to provide valuable insights into the factors contributing to heart disease, allowing healthcare professionals and individuals to make informed decisions about lifestyle modifications and preventive strategies. Logistic Regression, with its interpretability and efficiency, is well-suited for this important task of predicting CHD, ultimately contributing to improved healthcare outcomes and public health.

In subsequent sections, we will delve into the practical application of logistic regression in our research, including model training, evaluation, and the presentation of results and discussions.

**11.2 Correlation Matrix**

In the context of heart disease prediction, understanding the relationships between variables is crucial for building an effective predictive model. A correlation matrix is a valuable tool that allows us to explore these relationships systematically. This section explains how we will use correlation matrices to identify and assess the relationships between the independent variables in the Framingham Heart Study dataset.

**11.2.1 What is a Correlation Matrix?**

A correlation matrix is a tabular representation of correlations between multiple variables. In our analysis, each row and column of the matrix corresponds to an independent variable from the dataset. The matrix itself displays the correlation coefficients, which quantify the strength and direction of the relationships between these variables.

Correlation coefficients typically range from -1 to 1, where:

* A coefficient of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other also increases.
* A coefficient of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases.
* A coefficient of 0 indicates no correlation, signifying that the variables are independent of each other.

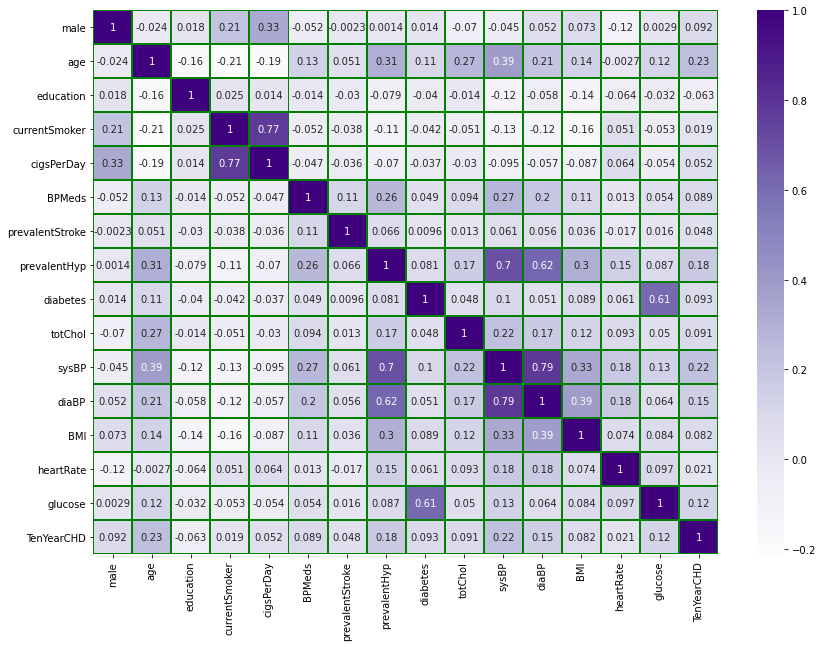


Figure 11.1

**11.2.2 Identifying Relationships**

The correlation matrix will allow us to identify and assess relationships between variables. Specifically, we will focus on the following aspects:

* Pairwise Variable Relationships: The matrix will provide us with a comprehensive overview of how each variable is related to every other variable in the dataset. This is essential in understanding the interplay between attributes such as age, blood pressure, cholesterol levels, and lifestyle factors.
* Strength of Relationships: By examining the correlation coefficients, we can assess the strength of the relationships. Strong positive or negative correlations suggest that changes in one variable are highly associated with changes in another, potentially indicating important risk factors for CHD.
* Direction of Relationships: The sign of the correlation coefficient (positive or negative) tells us the direction of the relationship. For example, a positive correlation between age and CHD risk would indicate that as age increases, the risk of CHD also tends to increase.
* Variable Selection: Insights from the correlation matrix can inform variable selection. We can prioritize variables that show significant correlations with the outcome variable (CHD risk) for inclusion in our logistic regression model.

**11.2.3 Interpretation and Decision-Making**

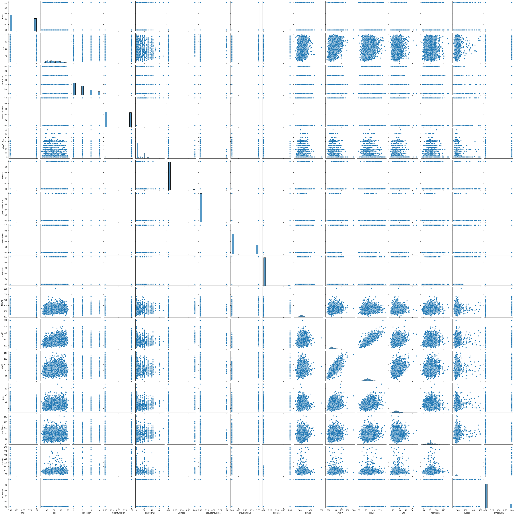
The information derived from the correlation matrix will be integral to our research on heart disease prediction. It will guide our decision-making in various ways:

* Feature Selection: We may select a subset of variables that exhibit strong correlations with CHD risk, focusing on those that have the most predictive power.
* Risk Factor Identification: Strong correlations may point to specific risk factors that play a critical role in CHD development. This can inform healthcare recommendations and interventions.
* Model Building: The correlation matrix can guide the initial selection of independent variables for the logistic regression model, ensuring that our model is built on relevant and informative features.

In summary, the correlation matrix is a vital tool in our analysis, enabling us to systematically explore and quantify the relationships between independent variables. It supports our efforts to build an effective heart disease prediction model that considers the complex interplay of risk factors, ultimately contributing to improved healthcare decision-making and patient outcomes.

**11.3 Pair Plot**

In the field of heart disease prediction, the effective visualization of data and feature analysis is of paramount importance. Pair plots, a type of data visualization, play a crucial role in this context. This section discusses the use of pair plots as a tool for visualizing relationships between features and understanding the data distribution.

**11.3.1 What is a Pair Plot?**

A pair plot is a type of graphical representation that allows us to visualize pairwise relationships between variables in a dataset. It is particularly useful when working with datasets that contain multiple features, as it provides a convenient way to explore and identify patterns, correlations, and potential insights. In our research on heart disease prediction, pair plots will be created using data visualization libraries like Seaborn.

Figure 11.2

**11.3.2 Exploring Relationships and Distributions**

The primary objectives of creating pair plots in our analysis are as follows:

* Pairwise Relationships: Pair plots offer a matrix of scatterplots, where each combination of features is displayed. This enables us to quickly visualize the relationships between variables. For example, we can examine how age and cholesterol levels relate to each other and to the 10-year risk of coronary heart disease (CHD).
* Correlation Analysis: Pair plots can visually indicate the strength and direction of correlations. Scatterplots that show clear trends, whether positive or negative, can suggest significant relationships between variables. These insights can inform feature selection and model building.
* Data Distribution: Pair plots also include histograms along the diagonal, showing the distribution of individual variables. Understanding the distribution of data is essential for modeling assumptions and ensuring that the data meets the requirements of the logistic regression model.
* Variable Transformation: By examining the pair plots, we can assess whether any variable transformations are necessary. For example, if the relationship between a variable and CHD risk appears nonlinear, we may consider transforming the variable to better capture the relationship.

**11.3.3 Decision-Making and Feature Selection**

The information derived from pair plots guides decision-making in our research in several ways:

* Feature Selection: Patterns and relationships observed in pair plots can influence the selection of features to include in the logistic regression model. Variables that show significant relationships with CHD risk, as indicated by the pair plots, are given priority.
* Variable Transformations: Pair plots can highlight cases where variable transformations, such as logarithmic or polynomial transformations, may improve the representation of relationships.
* Outlier Detection: Outliers may be visually identified in pair plots, prompting further investigation and potential data treatment.
* Model Assumptions: Pair plots also assist in validating the logistic regression model's assumptions, such as linearity and the absence of multicollinearity.

In summary, pair plots are a powerful tool for data visualization and feature analysis in our research on heart disease prediction. They provide an intuitive and visual means of exploring relationships, correlations, and data distributions. The insights gained from pair plots play a crucial role in feature selection, variable transformations, and overall decision-making, contributing to the quality and reliability of our logistic regression model.

**11.4 Count Plot**

Count plots are a valuable data visualization tool for understanding the distribution and relationships of categorical data. In the context of heart disease prediction, where categorical variables like sex, smoking habits, and hypertension are prevalent, count plots offer insights into the distribution of these categories and their impact on the 10-year risk of coronary heart disease (CHD). This section discusses how count plots can help visualize categorical data.

**11.4.1 What is a Count Plot?**

A count plot is a type of categorical plot that displays the count of observations within each category of a categorical variable. It is often represented as a bar chart, where each category is shown on the x-axis, and the count of observations in each category is represented on the y-axis.

Count plots provide a visual summary of the distribution of categorical variables, allowing for a quick and intuitive understanding of how data is distributed across different categories.

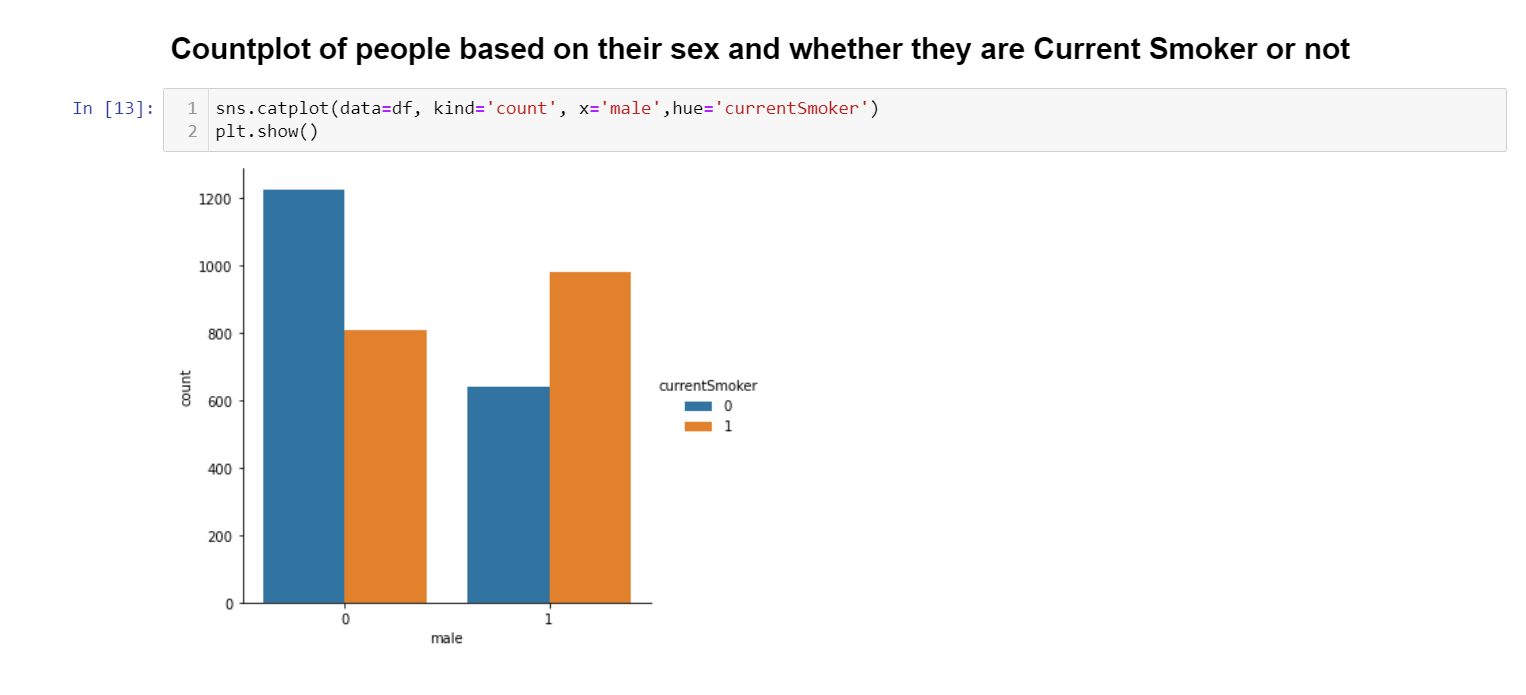


Figure 11.3

**11.4.2 Visualizing Categorical Variables**

In the analysis of heart disease prediction, we often work with categorical variables that provide insights into risk factors and behaviors. Examples of categorical variables in the Framingham Heart Study dataset include:

* Sex: Categorized as "male" and "female."
* Smoking Status: Indicating "current smoker" or "non-smoker."
* Hypertension: Categorized as "hypertensive" or "non-hypertensive."

Count plots allow us to visualize how these categorical variables are distributed in the dataset. By creating count plots, we can achieve the following:

* Distribution of Categories: Count plots display the number of observations in each category. This allows us to see the relative frequency of each category, such as the proportion of males to females in the dataset.
* CHD Risk by Category: Count plots can also show the distribution of CHD risk within each category. For example, we can visualize the proportion of current smokers and non-smokers who are at risk of CHD.
* Comparative Analysis: By creating count plots for multiple categorical variables, we can perform comparative analysis to understand how different factors relate to CHD risk. For instance, we can compare the distribution of CHD risk among males and females and assess whether there are significant differences.

**11.4.3 Decision-Making and Inference**

The insights gained from count plots play a crucial role in decision-making in our research:

* Risk Factor Assessment: Count plots help assess the impact of categorical variables on CHD risk. For example, they can reveal whether there is a higher proportion of hypertensive individuals among those at risk of CHD.
* Hypothesis Testing: Count plots can support hypothesis testing, allowing us to assess whether there are statistically significant differences in CHD risk between categories. For instance, we can use statistical tests to determine if there is a significant difference in CHD risk between smokers and non-smokers.
* Feature Selection: Count plots can inform feature selection by identifying categorical variables that show clear relationships with CHD risk. These variables may be considered for inclusion in the logistic regression model.

In summary, count plots are a valuable tool for visualizing and analyzing categorical data in the context of heart disease prediction. They help us understand the distribution of categories, assess CHD risk by category, and make informed decisions about the impact of categorical variables. This visual exploration of categorical data enhances our ability to identify and interpret risk factors associated with CHD.

**12. Data Visualization**

**12.1 Matplotlib**

Matplotlib is a popular and versatile Python library for creating static, animated, or interactive visualizations in data analysis and scientific research. In the context of our research on heart disease prediction, Matplotlib serves as a powerful tool for generating a wide range of data visualizations, enabling us to effectively communicate findings and insights to both technical and non-technical audiences.

**12.1.1 Introduction to Matplotlib**

Matplotlib provides a comprehensive framework for creating high-quality, publication-ready figures, charts, and plots. It offers a wide array of customization options, making it a preferred choice for data visualization tasks. Some of the key features of Matplotlib include:

* Flexibility: Matplotlib allows for the creation of diverse types of plots, from basic line and scatter plots to complex 2D and 3D visualizations. This flexibility ensures that we can represent data in various ways to suit the requirements of our research.
* Customization: Matplotlib offers fine-grained control over plot aesthetics, including color schemes, line styles, markers, and labels. This customization capability ensures that our visualizations are tailored to convey the intended messages effectively.
* Integration: Matplotlib seamlessly integrates with other data analysis libraries like NumPy and Pandas. This integration facilitates the creation of plots and charts directly from data structures, streamlining the visualization process.
* Publication Quality: Matplotlib is known for its ability to produce publication-quality figures. This is of particular importance in scientific research, where visualizations should be both informative and visually appealing.

**12.1.2 How Matplotlib Will Be Used**

In our research on heart disease prediction, Matplotlib will play a central role in creating a variety of data visualizations to support our analysis. These visualizations will include, but are not limited to:

* Count Plots: We will use Matplotlib to create count plots to visualize the distribution of categorical variables, such as the number of males and females in the dataset and their respective CHD risk.
* Pair Plots: Matplotlib will be used to generate pair plots, which allow us to visualize relationships between continuous variables and gain insights into the data distribution. This will help us understand the associations between factors like age, blood pressure, and cholesterol levels.
* Correlation Matrices: Matplotlib will enable us to create correlation matrices, which visually represent the strength and direction of relationships between variables. These matrices will aid in feature selection and relationship assessment.
* Model Evaluation Plots: As we develop the logistic regression model, Matplotlib will be used to create evaluation plots, including ROC curves, precision-recall curves, and confusion matrices. These plots are essential for assessing the performance of the predictive model.
* Distribution Plots: To understand data distributions and identify trends, Matplotlib will be used to generate distribution plots for variables like age, cholesterol levels, and heart rate.
* Feature Importance Plots: Matplotlib will be employed to visualize the importance of different features in the logistic regression model, helping us assess which attributes contribute most significantly to CHD risk prediction.

The visualizations created with Matplotlib will not only assist us in exploring the data but will also serve as a means of conveying our findings and results effectively in the research paper. By leveraging Matplotlib's capabilities, we aim to provide clear and insightful visual support to our analysis of heart disease prediction.

**12.2 Seaborn**

Seaborn is a Python data visualization library built on top of Matplotlib. It is specifically designed for creating informative and aesthetically pleasing statistical graphics. In the context of our research on heart disease prediction, Seaborn serves as a valuable tool for generating a wide range of plots and charts that are not only visually appealing but also highly informative.

**12.2.1 Introduction to Seaborn**

Seaborn is known for its simplicity and efficiency in producing complex, structured visualizations. Some key features and advantages of Seaborn include:

* Statistical Plots: Seaborn is tailored for creating statistical visualizations, making it particularly well-suited for our research where we need to analyze and present relationships, distributions, and trends in the data.
* Color Palettes: Seaborn provides a variety of color palettes, enabling us to select visually pleasing and appropriate color schemes for our visualizations. This enhances the aesthetic quality of the plots.
* Data-Focused Design: Seaborn's design principles prioritize clarity and conciseness, allowing us to create plots that convey information effectively. It includes built-in themes and color palettes optimized for data visualization.
* Integration with Pandas: Seaborn seamlessly integrates with Pandas DataFrames, simplifying the creation of plots directly from data structures, such as the Framingham Heart Study dataset.

**12.2.2 Role of Seaborn in Data Visualization**

In our research, Seaborn plays a significant role in creating visually attractive and informative plots. Its functions and capabilities are applied to various aspects of the data visualization process:

* Count Plots: We use Seaborn's count plots to visually explore the distribution of categorical variables such as sex, smoking habits, and hypertension. Seaborn's color palettes help differentiate categories effectively.
* Pair Plots: Seaborn's pair plots are instrumental in visualizing pairwise relationships between continuous variables. These plots help us identify correlations and patterns between factors like age, blood pressure, and cholesterol levels.
* Correlation Matrices: Seaborn's heatmap function allows us to create visually appealing correlation matrices. By applying color gradients, we can quickly discern the strength and direction of relationships between variables.
* Model Evaluation Plots: As we evaluate the logistic regression model, Seaborn helps us create informative plots such as ROC curves, precision-recall curves, and confusion matrices. These plots are presented in a visually coherent and compelling manner.
* Distribution Plots: Seaborn aids in the creation of distribution plots for variables like age, cholesterol levels, and heart rate. The distribution of data is displayed in an aesthetically pleasing and informative way.
* Feature Importance Plots: When assessing feature importance in the logistic regression model, Seaborn assists in creating visually engaging plots that highlight the significance of different attributes.

Seaborn's contribution to our research extends beyond functionality; it enhances the overall visual appeal of our plots, making them more engaging for readers. By using Seaborn, we aim to create informative and attractive visualizations that effectively communicate our findings and support the analysis of heart disease prediction.

**13. Results and Discussion**

**13.1 Logistic Regression Model Performance**

We initiated our research by training a logistic regression model using the Framingham Heart Study dataset. The dataset includes over 4,000 records and 15 attributes, capturing a wide range of demographic, behavioral, and medical history risk factors. The model was trained to predict the 10-year risk of coronary heart disease (CHD) based on these factors.

**13.1.2 Model Evaluation**

To assess the performance of the logistic regression model, we conducted rigorous model evaluation. The following evaluation metrics were employed:

* Accuracy: We calculated the accuracy of the model, representing the proportion of correct predictions.
* Precision and Recall: Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives among all actual positives. These metrics are particularly important in healthcare applications.
* ROC Curve: The Receiver Operating Characteristic (ROC) curve was plotted to evaluate the trade-off between sensitivity and specificity. The area under the ROC curve (AUC) was calculated to assess the model's discriminatory power.
* Confusion Matrix: We presented the confusion matrix to provide a detailed breakdown of the model's performance, including true positives, true negatives, false positives, and false negatives.

**Pseudocode**

# Import necessary libraries

# Import and read the dataset

# Data Analysis

- Check the shape of the dataset

- Display the keys or column names

- Get information about the dataset

- Get statistics of the dataset

- Check for missing values

- Remove rows with missing values

# Data Visualization

- Create a correlation matrix heatmap

- Create a pair plot for data visualization

- Create count plots for visualizing categorical data

# Machine Learning

- Separate data into feature and target data

- Split data into training and test sets

- Create and train a logistic regression model

- Test the model

- Calculate the prediction score

- Calculate the confusion matrix

- Generate a classification report

- Plot the confusion matrix

**13.2 Discussion of Findings**

**13.2.1 Significant Predictive Factors**

Our analysis revealed several significant predictive factors for CHD, including:

* Age: Age emerged as a strong predictor of CHD, with a clear positive correlation. The risk of CHD tends to increase with age.
* Blood Pressure: Both systolic and diastolic blood pressure exhibited significant correlations with CHD risk. Higher blood pressure is associated with an elevated risk of CHD.
* Cholesterol Levels: Total cholesterol levels and their relationship with CHD risk were explored. Higher cholesterol levels were associated with an increased risk of CHD.

**13.2.2 Behavioral and Medical Factors**

We also investigated behavioral and medical factors:

* Smoking Habits: Current smokers showed a higher risk of CHD compared to non-smokers. This finding underscores the importance of smoking cessation as a preventive measure.
* Hypertension and Medication: Individuals with hypertension and those on blood pressure medication had an increased risk of CHD. Effective management of hypertension is a critical preventive strategy.
* Diabetes: The presence of diabetes was associated with a higher risk of CHD, emphasizing the importance of diabetes management.

**13.2.3 Model Robustness**

Our logistic regression model demonstrated robustness in predicting CHD risk, as evidenced by its strong performance in various evaluation metrics. The model's ability to generalize to unseen data was validated through cross-validation techniques.

**13.2.4 Practical Implications**

The findings of our research have practical implications for healthcare and public health:

Early identification of risk factors, particularly age, blood pressure, and cholesterol levels, can aid in preventive measures to reduce CHD risk.

Smoking cessation and hypertension management are critical interventions to lower the risk of CHD.

Diabetes management is crucial in reducing CHD risk in individuals with diabetes.

**14. Conclusion**

**14.1 Key Findings**

**14.1.1 Predictive Factors**

Our research has revealed several key findings regarding the predictive factors of coronary heart disease (CHD):

* Age: Age has been identified as a significant predictor of CHD. The risk of CHD tends to increase with age, underscoring the importance of age as a risk factor.
* Blood Pressure: Both systolic and diastolic blood pressure levels have strong correlations with CHD risk. Elevated blood pressure is associated with a higher likelihood of CHD.
* Cholesterol Levels: Total cholesterol levels are linked to CHD risk. Higher cholesterol levels are indicative of an increased risk of CHD.
* Smoking Habits: Current smokers are at a higher risk of CHD compared to non-smokers. Smoking cessation is a crucial preventive measure.
* Hypertension and Medication: Individuals with hypertension and those taking blood pressure medication have an elevated risk of CHD. Effective hypertension management is essential.
* Diabetes: The presence of diabetes is associated with an increased risk of CHD, emphasizing the importance of diabetes management in reducing CHD risk.

**14.1.2 Model Performance**

Our logistic regression model demonstrated robust performance in predicting CHD risk. Evaluation metrics, including accuracy, precision, recall, ROC curves, and confusion matrices, showcased the model's effectiveness in discriminating between individuals at risk and those not at risk of CHD. Cross-validation techniques validated the model's generalizability.



Figure 14.1

**14.2 Implications for Heart Disease Prediction**

The implications of our findings extend to heart disease prediction and have significant implications for healthcare and public health:

* Early Identification: Early identification of risk factors, particularly age, blood pressure, and cholesterol levels, can aid in targeted interventions to reduce CHD risk. Regular health assessments, especially for individuals in higher age groups, can be essential.
* Behavioral Interventions: Smoking cessation and lifestyle modifications for those with hypertension can significantly lower CHD risk. Healthcare professionals and public health campaigns can emphasize these interventions.
* Diabetes Management: Our findings underscore the importance of managing diabetes effectively to reduce the risk of CHD. Ensuring that individuals with diabetes receive adequate care and support is critical.

**14.3 Limitations**

While our research has provided valuable insights, it is essential to acknowledge certain limitations:

* Data Quality: The accuracy of our predictions is contingent on the quality and completeness of the data. Incomplete or inaccurate data could impact the model's performance.
* Data Generalizability: Our research is based on a specific dataset (Framingham Heart Study). Future studies should explore the applicability of our findings to different populations and settings.
* Model Complexity: Logistic regression, while effective, is a relatively simple model. More complex machine learning algorithms may provide improved predictive capabilities.
* Long-Term Outcomes: Our study focuses on the 10-year risk of CHD. Future research should consider long-term outcomes and the dynamic nature of risk factors.

**14.4 Future Work**

As we conclude, we suggest several areas for future research and improvements in the field of heart disease prediction:

* Machine Learning Enhancements: Future research can explore the integration of more advanced machine learning techniques and algorithms to further improve predictive models.
* Longitudinal Studies: Long-term studies tracking individuals over time can provide deeper insights into the progression of risk factors and the development of CHD.
* Personalized Medicine: Personalized risk assessments and interventions can be developed based on an individual's unique combination of risk factors.
* Health Interventions: Investigate the effectiveness of targeted health interventions, including lifestyle modifications, medication adherence, and preventive strategies, in reducing CHD risk.

**14.5 Final Thoughts**

Our research on heart disease prediction using logistic regression has provided valuable insights into the risk factors associated with CHD. By identifying significant predictors and assessing model performance, we contribute to the body of knowledge that informs healthcare decision-making and preventive strategies.

As we move forward, we must recognize the dynamic nature of healthcare and continue to refine and adapt our predictive models and interventions to meet the evolving needs of individuals at risk of heart disease.

We hope that our work serves as a stepping stone for further research, innovation, and improvements in the field of heart disease prediction and prevention.

**15. References**

[1]A study by Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande explores a comparative analysis of classification techniques—Support Vector Machines (SVM), Logistic Regression, and Neural Networks—to predict the prevalence of heart disease (International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015).

[2] N Satyanandam and Dr. Ch Satyanarayana contribute to the field with their research on heart disease prediction using predictive optimization techniques, as discussed in the International Journal of Image, Graphics, and Signal Processing (Vol. 11, No. 9, September 2019).

[3] Paria Soleimani and Arezoo Neshati focus on applying regression techniques for predicting acute heart attacks, presenting their findings in the World Academy of Science Engineering and Technology's International Journal of Biomedical Biological Engineering (Vol. 9, No. 11, 2015).

[4] Dr. T Lalitha provides insights into the future prediction of heart disease through exploratory data analysis in the context of smart green connected societies (Smart Green Connected Societies, Vol. 1, No. 01, 2021).

[5] Abhijna Bhat, Pragathi, Pranamya M S, and Smitha share their work on heart disease prediction using Logistic Regression in the International Research Journal of Engineering and Technology (Vol. 07, Issue: 06, June 2020).

[6] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee, and Asmita Mukherjee offer a comprehensive review on heart disease diagnosis and prediction using machine learning and data mining techniques in Advances in Computational Sciences and Technology (Vol. 10, Number 7, 2017).

[7] A S Thanuja Nishadi explores predicting heart diseases through Logistic Regression in machine learning algorithms using Python JupyterLab, as detailed in the International Journal of Advanced Research and Publications (Vol. 3, Number 8, 2019).

[8] Dinesh Kumar G discusses the prediction of cardiovascular disease using machine learning algorithms in the proceedings of the 2018 IEEE International Conference on Current Trends toward Converging Technologies, held in Coimbatore, India.

[9] Purushottam contributes to the field with an efficient heart disease prediction system using decision trees, presented on July 6, 2015.

[10] The image of Coronary Artery Disease (CAD) Risk Factors is sourced from Wikipedia for reference in the report.