CSP 571 Project Report

HEART ATTACK ANALYSIS AND PREDICTION

Ayush Dhar (A20545212)
Ekta Shukla (A20567127)
Manpreet Kaur (A20551672)
Shivani Shrivastav (A20553589)
Vishwashree Reddy (A20556543)
Illinois Institute of Technology

Table of Contents

ABSTRACT

Heart disease continues to be one of the primary causes of death globally, thus requiring the development of robust methods for its prediction and analysis. The purpose of this report is to thoroughly analyze and predict heart attack occurrences, using a multifaceted approach to identify important risk factors and predictive markers. Our objective is to utilize sophisticated statistical techniques and machine learning algorithms to reveal noteworthy trends and patterns in heart attack data. The study utilizes the comprehensive Cleveland dataset, which includes demographic and clinical variables, allowing for a comprehensive analysis of the individual and combined factors affecting the risk of heart attacks.

Key findings highlight the pivotal role of factors such as age, cholesterol levels, blood pressure, and other factors in predicting heart attacks. The predictive models developed in this report demonstrate high accuracy, offering valuable insights for early diagnosis and preventive strategies. Furthermore, visualizations and detailed breakdowns of the data provide a clear understanding of the critical variables and their interrelations.

The primary objective of the project is to analyze and predict heart attacks using machine learning models. The goal is to develop a predictive model capable of accurately assessing the risk of heart attacks based on patient data. This project holds immense importance in the realm of preventive healthcare as it facilitates early diagnosis and enhances treatment planning within the field of cardiology. In addition to advancing the scientific knowledge regarding predictors of heart attacks, this report establishes a crucial framework for formulating more impactful healthcare strategies and alleviating the global burden of heart disease.

Research Objectives:

The primary objectives are to:

1. Identify key health indicators that significantly contribute to the risk of heart attacks.
2. Develop a robust machine learning model that can accurately predict the likelihood of a heart attack.
3. Evaluate and compare the performance of various machine learning algorithms in terms of predictive accuracy.

Findings:

A preliminary analysis has indicated that age, cholesterol levels, blood pressure, and the type of chest pain are among the key predictors of heart attack risk. Out of all the algorithms that were tested, it has been observed that both random forests and neural networks have exhibited the highest accuracy when it comes to predicting the risk of heart attacks. These findings strongly indicate that machine learning models can be harnessed to accurately pinpoint high-risk individuals, thus enabling medical interventions to be administered in a timely manner.

Problem Statement:

Globally, heart disease continues to be the primary cause of death, and heart attacks play a significant role in this. The importance of early detection and comprehensive risk assessment cannot be overstated in terms of reducing mortality rates and enhancing patient outcomes. Conventional methods of risk assessment frequently have limitations in terms of accuracy and scope. The purpose of this project is to fill this void by creating a predictive model based on machine learning. This model will effectively evaluate the risk of heart attacks using extensive patient data.

Proposed Methodology:

Data Collection:

- The Heart Disease dataset from the UCI Machine Learning Repository is the primary dataset utilized in this analysis. It encompasses a range of medical attributes such as age, sex, chest pain type, resting blood pressure, cholesterol levels, and additional variables of significance.
- **Data Source Link** - https://archive.ics.uci.edu/dataset/45/heart+disease

Data Cleaning:

- **Addressing Missing Values**: When the symbol *"?"* is encountered in the dataset, it is substituted with "*NA*" to indicate missing values. We eliminated all rows that contained any missing values, thereby ensuring a dataset that is free of errors.
- Target Variable Transformation involves converting the initially multi-class target variable into a binary variable. When the value is 0, it remains as 0, whereas all other values are transformed into 1, thus simplifying the classification problem into a binary one.
- The categorical variables *ca* and *thal* have been converted into numeric types. This particular step is essential for modeling techniques that rely on numerical input.
- The dataset is filtered to create a subset consisting solely of the numeric variables *age*, *trestbps*, *chol*, *thalach*, and *oldpeak*.
- A subset is created specifically for the categorical variables that have been selected. The function *mutate_if(is.numeric, as.factor)* guarantees the conversion of any remaining numeric variables in this subset into factors.
- By combining the numeric and categorical subsets, a processed dataset is generated that is suitable for modeling purposes.
- The dataset structure is verified at various stages to ensure the correct application of all transformations and validate the data types.

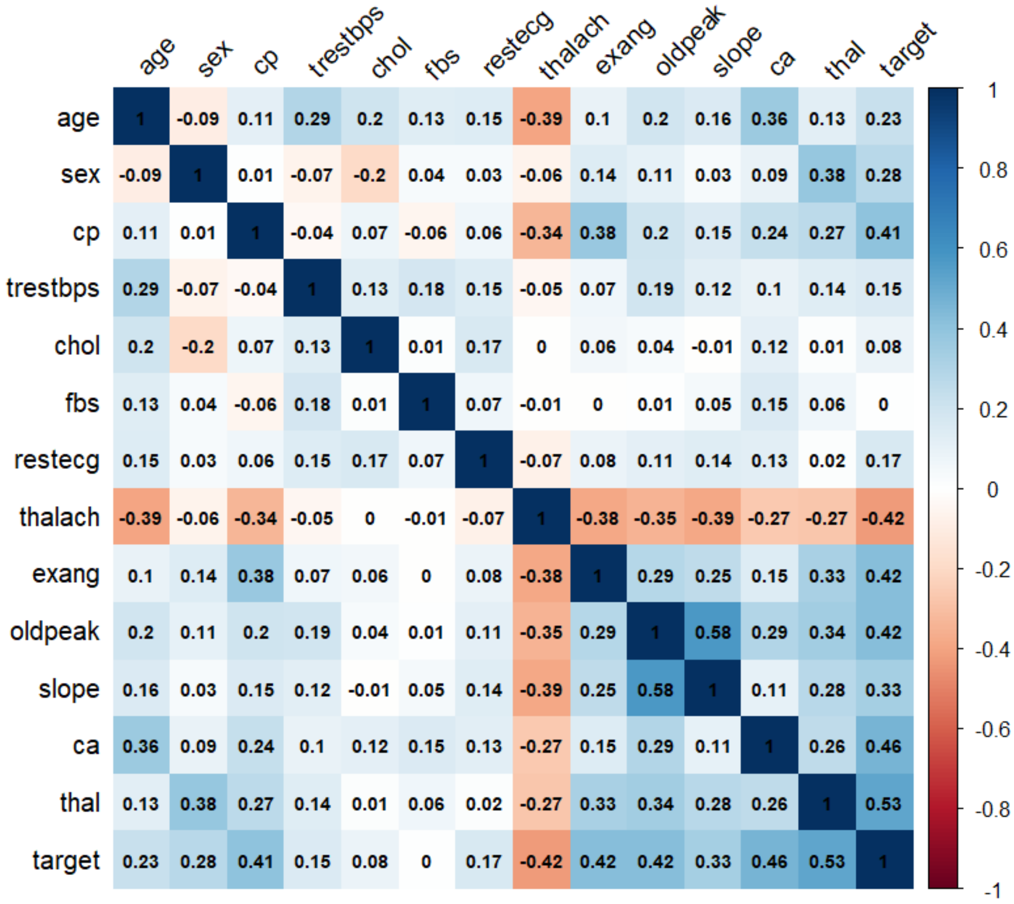
Exploratory Data Analysis (EDA):

- Conduct descriptive statistical analysis to compute summary statistics for numerical variables.
- Employ various data visualization techniques such as histograms, box plots, scatter plots, and correlation matrices to gain insight into the dataset.
- Perform a feature analysis to examine the significance and influence of features on the target variable.

Correlation Analysis

In order to examine the connections between features and the target variable, we perform a correlation analysis.

- The correlation matrix is calculated using the *cor()* function, which measures the linear correlation between every pair of variables.
- By utilizing the *corrplot()* function, the correlation matrix can be visualized as a heatmap. The color intensity within the heatmap denotes the magnitude of correlations, and the annotations provide correlation coefficients.



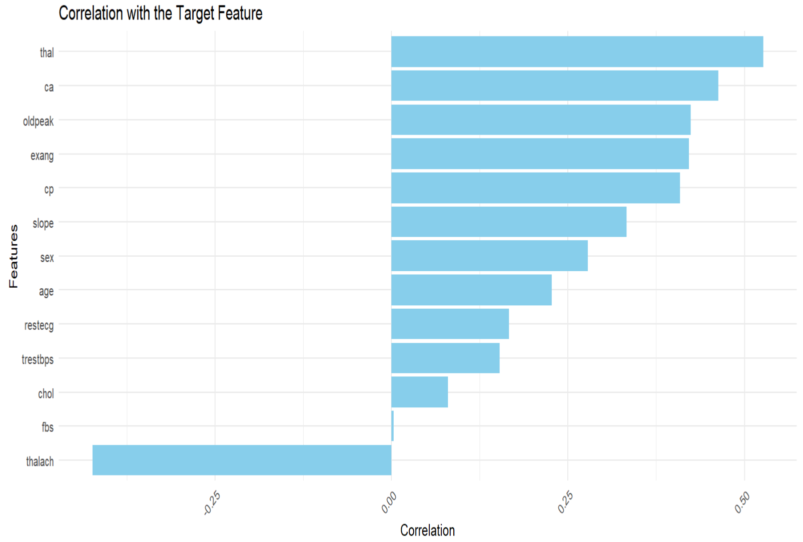
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.09	0.11	0.29	0.2	0.13	0.15	-0.39	0.1	0.2	0.16	0.36	0.13	0.23
sex	-0.09	1	0.01	-0.07	-0.2	0.04	0.03	-0.06	0.14	0.11	0.03	0.09	0.38	0.28
cp	0.11	0.01	1	-0.04	0.07	-0.06	0.06	-0.34	0.38	0.2	0.15	0.24	0.27	0.41
trestbps	0.29	-0.07	-0.04	1	0.13	0.18	0.15	-0.05	0.07	0.19	0.12	0.1	0.14	0.15
chol	0.2	-0.2	0.07	0.13	1	0.01	0.17	0	0.06	0.04	-0.01	0.12	0.01	0.08
fbs	0.13	0.04	-0.06	0.18	0.01	1	0.07	-0.01	0	0.01	0.05	0.15	0.06	0
restecg	0.15	0.03	0.06	0.15	0.17	0.07	1	-0.07	0.08	0.11	0.14	0.13	0.02	0.17
thalach	-0.39	-0.06	-0.34	-0.05	0	-0.01	-0.07	1	-0.38	-0.35	-0.39	-0.27	-0.27	-0.42
exang	0.1	0.14	0.38	0.07	0.06	0	0.08	-0.38	1	0.29	0.25	0.15	0.33	0.42
oldpeak	0.2	0.11	0.2	0.19	0.04	0.01	0.11	-0.35	0.29	1	0.58	0.29	0.34	0.42
slope	0.16	0.03	0.15	0.12	-0.01	0.05	0.14	-0.39	0.25	0.58	1	0.11	0.28	0.33
ca	0.36	0.09	0.24	0.1	0.12	0.15	0.13	-0.27	0.15	0.29	0.11	1	0.26	0.46
thal	0.13	0.38	0.27	0.14	0.01	0.06	0.02	-0.27	0.33	0.34	0.28	0.26	1	0.53
target	0.23	0.28	0.41	0.15	0.08	0	0.17	-0.42	0.42	0.42	0.33	0.46	0.53	1

Feature Correlation with Target Variable

In order to determine the features that exhibit the strongest correlation with the target variable, we perform calculations and generate visual representations of these correlations.

- **Compute Correlations**: Through the use of the *sapply()* function, the correlation between each feature and the target variable is determined.
- **Correlation Data Frame**: A data frame named *correlations_df* has been generated to store the names of features and their respective correlation coefficients.

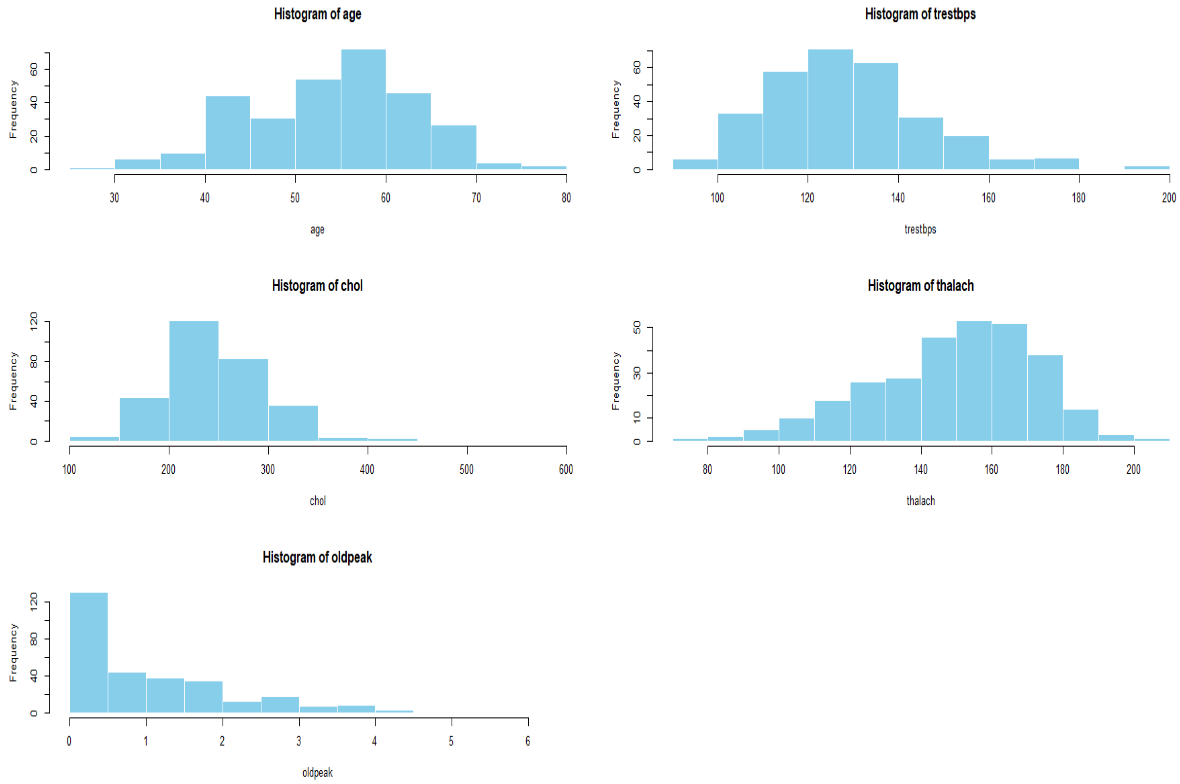
- **Bar Plot**: By using the *ggplot()* function, a bar plot is produced to effectively represent the correlations, arranging the features based on their correlation strength.



Distribution Analysis

Histograms are employed for the purpose of analyzing the distributions of numerical features.

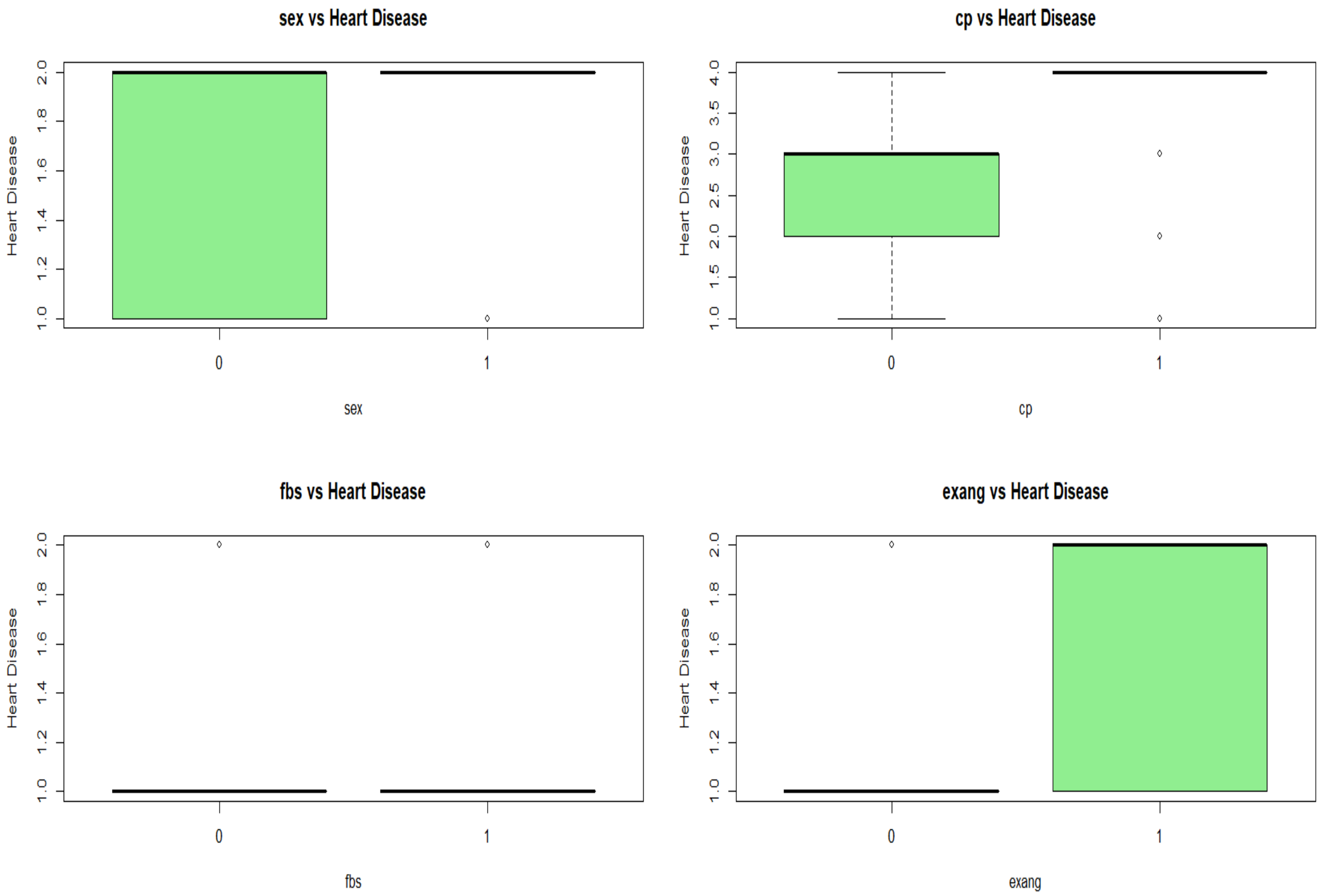
- **Numerical Features**: The numerical features are listed in the vector numerical_features for analysis.
- **Histogram Plots**: A histogram is generated to display the distribution of each numerical feature.



Boxplot Analysis for Categorical Features

Boxplots are employed to analyze the correlation between categorical features and the target variable.

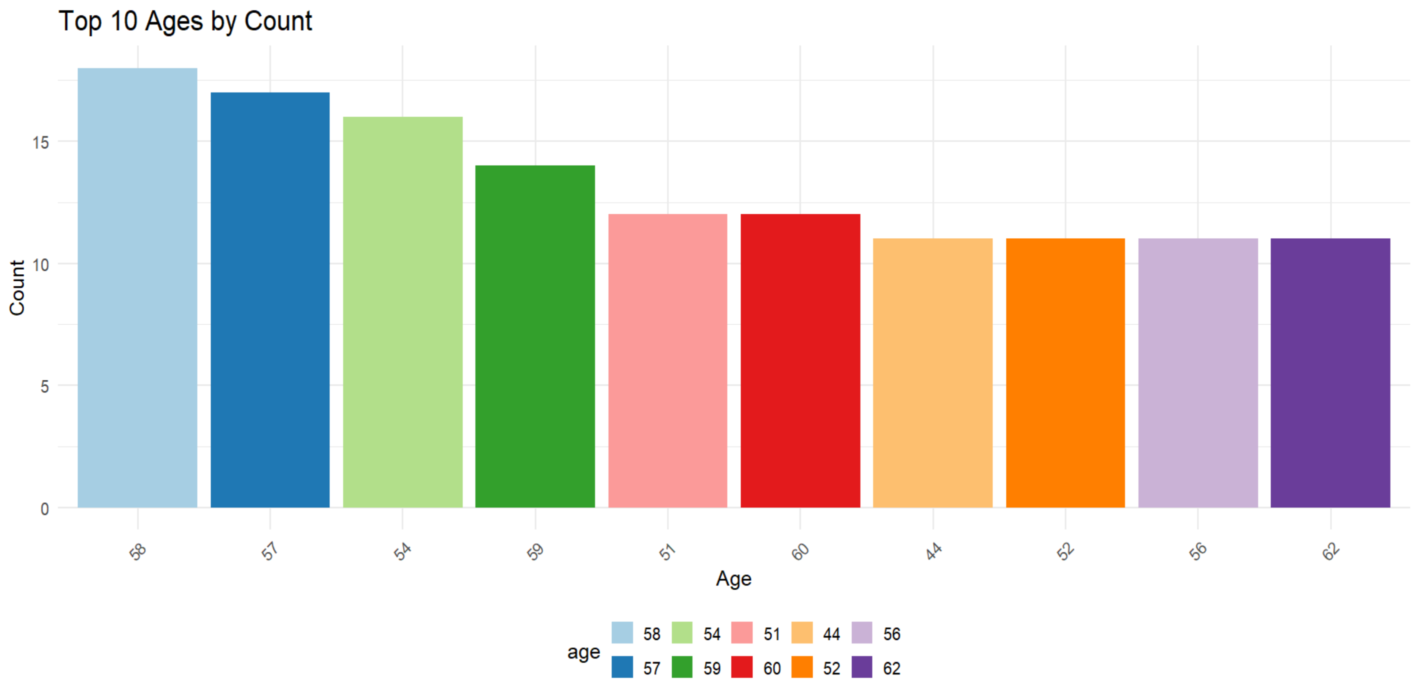
- **Categorical Features**: The categorical_features vector provides a list of categorical variables for analysis.
- **Boxplots**: A boxplot is generated for every categorical feature to compare the distributions across the categories of the target variable.

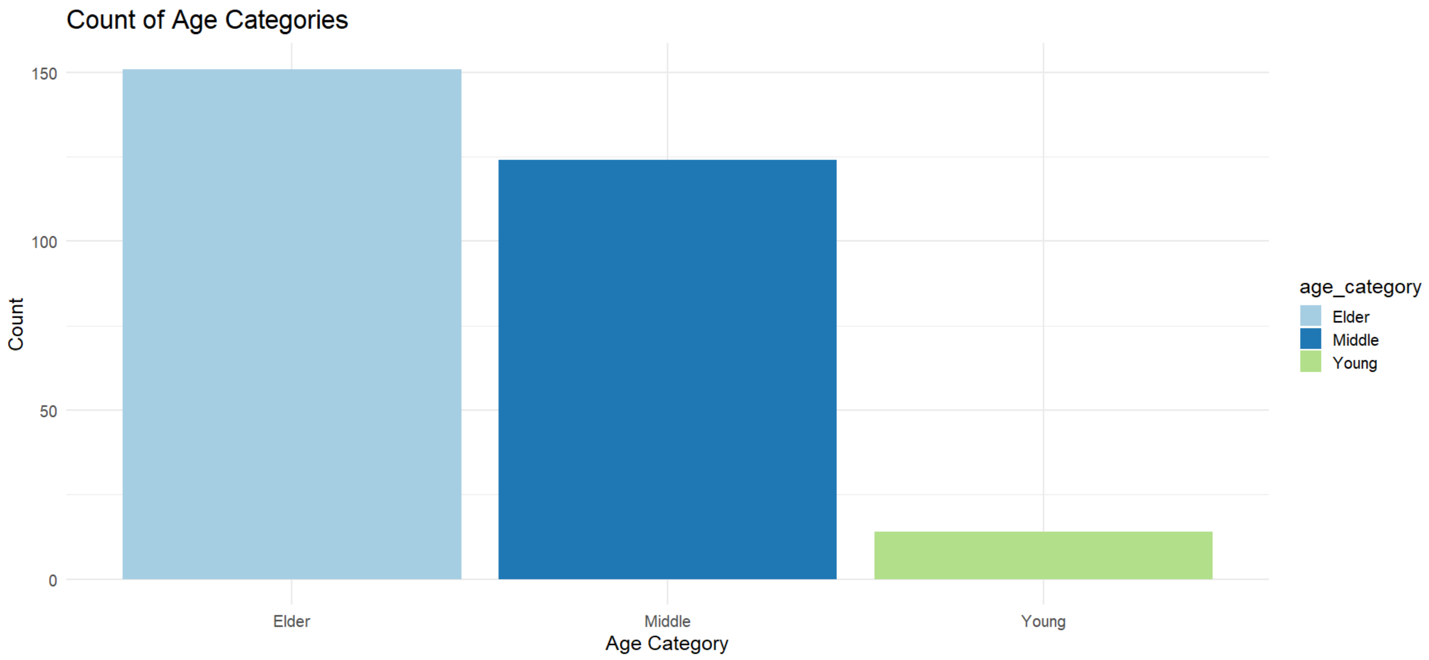


Age Analysis

We conduct an analysis on the distribution of ages and subsequently classify them into categories.

- **Age Distribution**:. The analysis focuses on the distribution of age, visualizing the top 10 age counts using a bar plot.
- **Age Categorization**: The mutate() function classifies age into categories such as "Young", "Middle", and "Elder”.
- **Age Category Plot**: The bar plot displays the frequency of individuals in each age category.

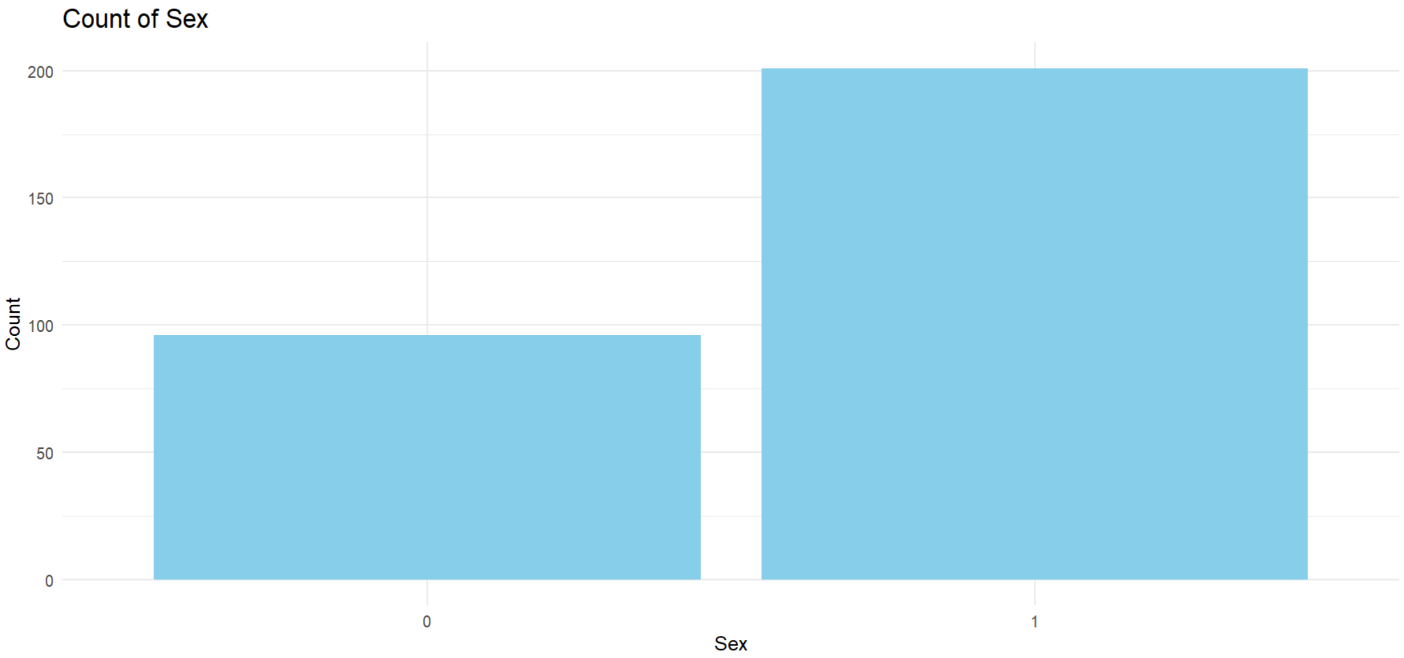

Top 10 Ages by Count
Count
0
5
10
15
58
57
54
59
51
60
44
52
56
62
Age
age
58
57
54
59
51
60
44
52
56
62

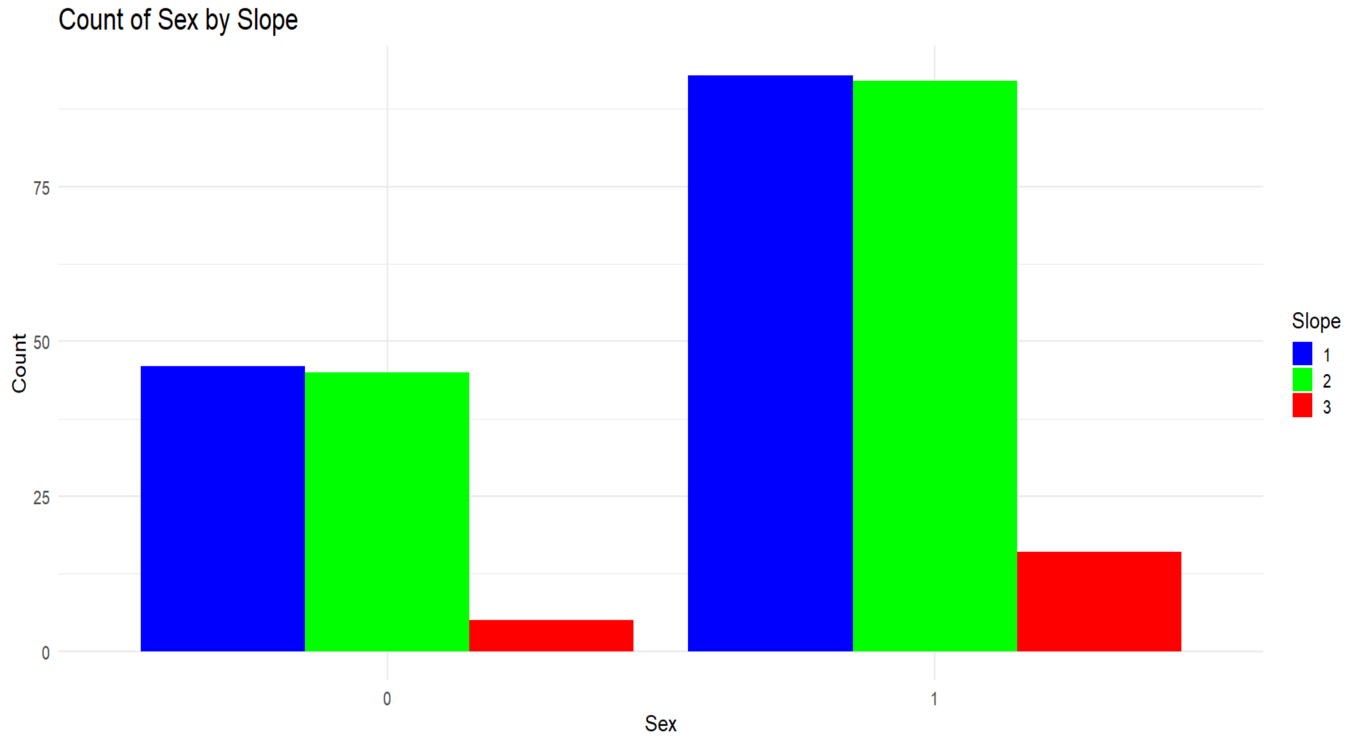

Count of Age Categories
Count
0
50
100
150
Elder
Middle
Young
Age Category
age_category
Elder
Middle
Young

Gender Analysis

An analysis is conducted on the distribution of genders.

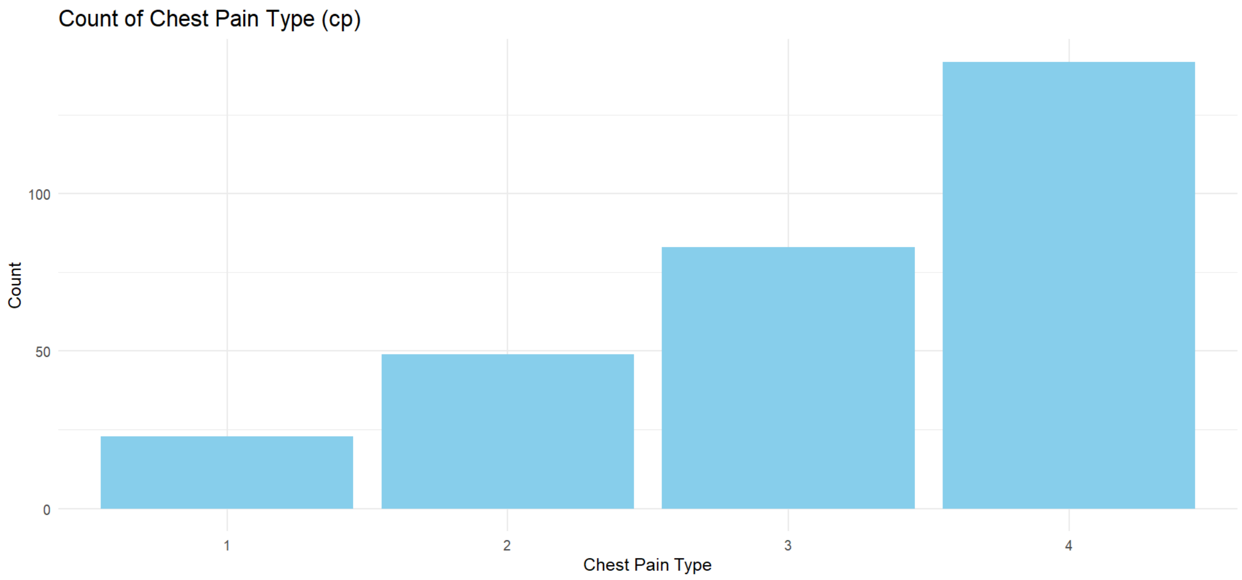
- **Gender Distribution**: The bar plot provides a visual representation of the distribution of gender categories in the dataset.





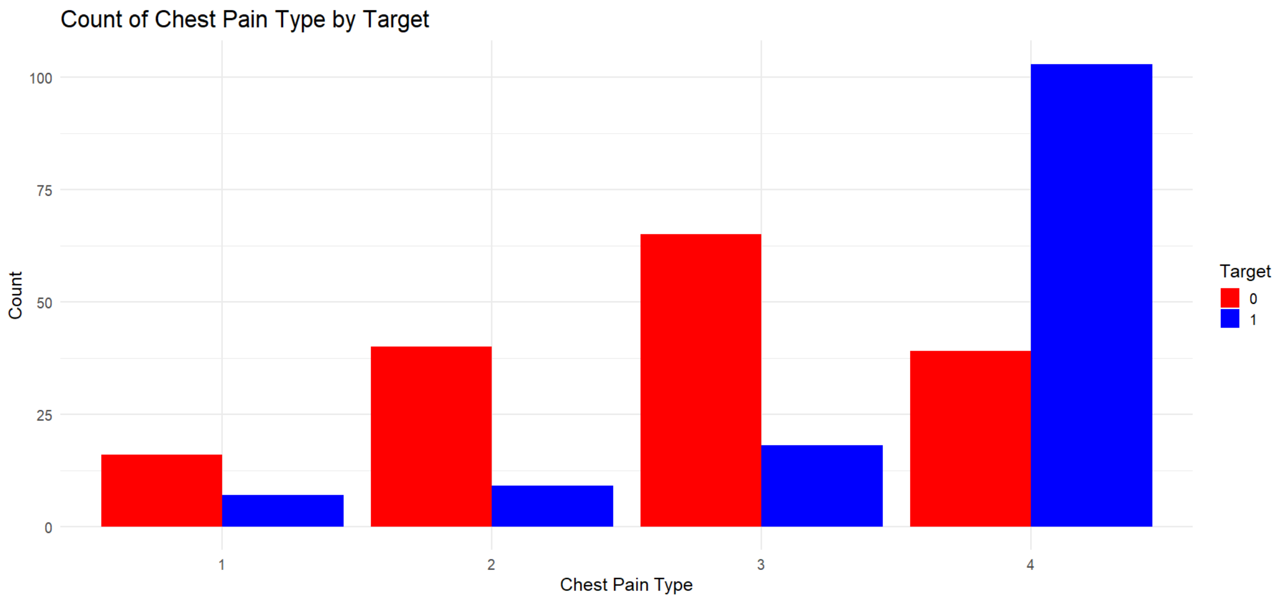
Count of Chest Pain Type (cp)

The aim of this bar plot is to visually portray the frequency distribution of different chest pain types (cp) present in the dataset. This plot aids in the identification of the most and least common types of chest pain among the patients.



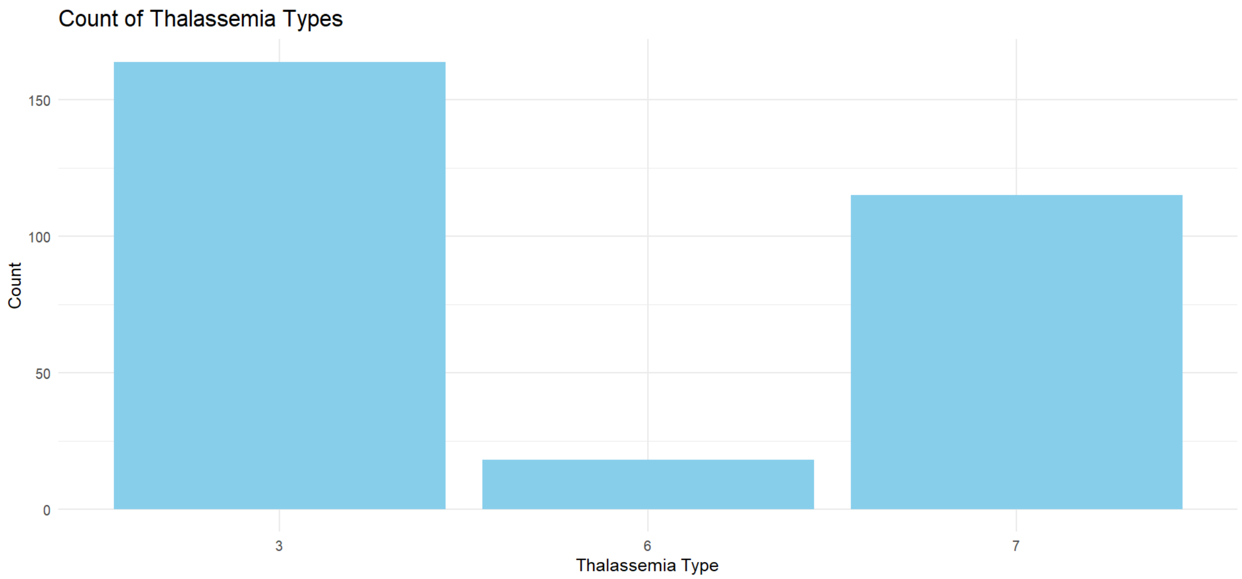
Count of Chest Pain Type by Target

The presented bar plot offers a comparative perspective on the occurrence of different types of chest pain (cp), categorized by the target variable (target) that signifies the existence of heart disease. The main objective of this plot is to elucidate the connection between various chest pain types and the presence or absence of heart disease.



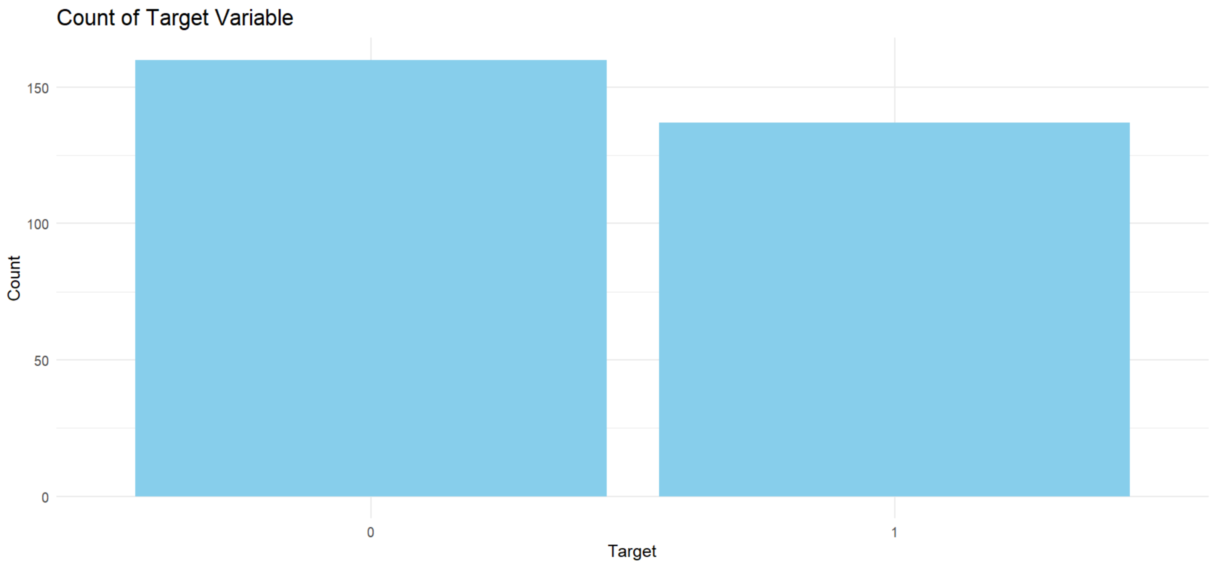
Count of Thalassemia Types

This bar plot effectively visualizes the frequency distribution of thalassemia types in the dataset. This plot aids in the identification of the most and least prevalent types of thalassemia among the patients.



Count of Target Variable

The bar plot portrays the distribution of the target variable, specifically highlighting the number of patients categorized with and without heart disease. This plot assists in grasping the dataset's balance in relation to the presence or absence of heart disease.



Feature Engineering:

This analysis leverages a Random Forest model to determine the importance of features in predicting the target variable (presence of heart disease). Below is a detailed description of each step and its purpose.

Building the Random Forest Model:

- Purpose: The aim is to develop a Random Forest model using the dataset, with the goal of predicting the target variable by incorporating all available features.
- Extraction of Feature Importance: The importance function computes the significance of each feature using Mean Decrease Accuracy and Mean Decrease Gini.

Sorting and Selecting Top Features:

- Purpose: The objective is to arrange the features according to their significance and choose the nine most crucial ones.
- The outcome of this analysis involves the identification of the most influential features in predicting the target variable.

Creating a New Dataset with Important Predictors:

- Purpose: The objective is to generate a novel dataset that exclusively consists of the top 9 essential features and the target variable.
- The end result is a reduced dataset that concentrates on the most significant predictors.

Feature Engineering, Normalization, and Scaling:

- **Categorical and Continuous Variable Identification:**
  - Purpose: The objective is to differentiate categorical variables from continuous variables.
  - The outcome includes separate lists of categorical and continuous variables.
- **Dummy Variable Creation**:
  - Purpose: The aim is to convert categorical variables into dummy/indicator variables so that they can be utilized in the model effectively.
  - Result: The dataset now contains dummy variables in place of the original categorical variables.
- **Normalization and Scaling**:
  - Goal: Standardizing and scaling continuous variables to improve model performance.
  - Conclusion: A dataset featuring scaled continuous variables.

Model Training:

The purpose of this analysis is to assess the effectiveness of various classification algorithms in relation to a dataset. Our aim is to measure and contrast the predictive precision of Logistic Regression, SVM, Random Forest, XGBoost, and KNN in a binary classification scenario.

Data Splitting:

Initially, we partitioned the dataset into training and test sets with a 70-30 division. Through partitioning, we are able to train models on a designated subset and gauge their performance on an unfamiliar subset.

Feature and Target Separation:

In both the training and testing phases, the target variable was distinct from the feature set.

Logistic Regression:

The training data was used to train a logistic regression model, which was subsequently evaluated on the test data. A threshold of 0.5 was utilized to convert the predictions into binary outcomes. To calculate accuracy, the correct predictions were divided by the total number of test samples.

Support Vector Machine (SVM):

A Support Vector Machine model was trained using a radial basis function kernel. Projections were made on the test set and accuracy was evaluated in a comparable fashion.

Random Forest:

The test set was used to fit and evaluate a random forest model in order to ascertain its accuracy.

XGBoost:

The data underwent formatting for XGBoost, followed by training a model using specified parameters. An evaluation was performed to determine the accuracy of predictions.

K-Nearest Neighbors (KNN):

To start, a KNN model with a value of k = 10 was trained and assessed. The process of cross-validation was carried out to identify the ideal number of neighbors (k). Accuracy was computed for both the initial and optimized models. Moreover, a step of feature selection was carried out to eliminate features with near-zero variance, and subsequently, a KNN model was re-evaluated. Utilizing the kknn package, K-nearest neighbors were explored with the use of Manhattan distance.

Model Evaluation:

The performance of each model was evaluated based on accuracy. Here are the results:

- **Logistic Regression Accuracy: 0.84**
- **Support Vector Machine (SVM) Accuracy: 0.81**
- **Random Forest Accuracy: 0.84**
- **XGBoost Accuracy: 0.80**
- **K-Nearest Neighbors (KNN) Accuracy: 0.65**
- **Perform 10-fold cross-validation Accuracy: 0.6854**
- **Accuracy with filtered features: 0.6854**
- **Accuracy with k = 11 and Manhattan distance (kknn): 0.8539**

Conclusion:

This project aims to develop a robust machine learning model for heart attack prediction, providing valuable insights for early diagnosis and preventive healthcare. By leveraging patient data and advanced analytical techniques, the project seeks to improve the accuracy and reliability of heart attack risk assessments, ultimately enhancing patient care and outcomes.

Future Scope:

In the subsequent phase of the project, we will focus on enhancing the predictive model by conducting hyperparameter tuning and cross-validation. Furthermore, the model will be seamlessly incorporated into a practical healthcare application, with continuous monitoring and maintenance to guarantee its enduring accuracy and reliability. Subsequent research will also investigate the incorporation of supplementary data sources and advanced machine learning methodologies to improve predictive accuracy.

Source Code:

We used GitHub for collaboration.

Github Link: https://github.com/Ekta023/HEART-ATTACK-ANALYSIS-AND-PREDICTION

Citation and Relevant Literature:

- Balcioglu, Yavuz, and Bulent Sezen. 2021. "Predicting of Heart Disease Risk Factors with R."

- Prasad, Ramakant, Pooja Gupta, Sapna Malhotra, Asst Professor, Gargi College, and Du. 2022. "Prediction of Heart Disease Using Hybrid Form of Mathematical Model and Machine Learning Approach." 2096-3246.

- Marathe, Ninad, Sushopti Gawade, and Adarsh Kanekar. 2021. "Prediction of Heart Disease and Diabetes Using Naive Bayes Algorithm." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 447-453. https://doi.org/10.32628/CSEIT217399.

- https://www.geeksforgeeks.org/heart-disease-prediction-using-logistic-regression-in-r/