INDIAN INSTITUTE OF TECHNOLOGY KANPUR

DEPARTMENT OF MATHEMATICS & STATISTICS

“Analysis of the data of patients to predict heart disease”

Abstract

This project involves analysis of the heart disease patient dataset with proper data processing. Dataset has been checked for any missing or duplicate values and thus those values have been fixed. Dataset is analysed by finding different statistical measures and plotting the graphs to check the relationship of different parameters with having heart disease using matplot library and seaborn library. Various inferences have been made by analysing the plots. This is a classification problem, with input features as a variety of parameters, and the target variable as a binary variable predicting whether heart disease is present or not. Then Logistic Regression model (scikit learn) is trained using training data. Model is evaluated using accuracy score. Accuracy achieved is 82%. A predictive system predicts if a person will get heart disease or not by entering its features. This project delves into the hidden insights and trends that drive impactful decision-making. My approach encompasses various methodologies and techniques, including exploratory data analysis (EDA) to navigate through the causes of heart disease, unveiling patterns, trends, and hidden gems that lie within. We employ machine learning algorithms to classify occurrence of heart disease based on various features. Predictive modelling techniques are utilized to develop models that forecast occurrence of heart disease

1 Introduction

Heart disease, encompassing conditions like coronary artery disease, heart failure, and arrhythmias, remains a major global health concern. Developing accurate prediction models for heart disease is essential for early Intervention and Prevention. Identifying individuals at risk early allows for timely interventions. Lifestyle modifications, medication, or surgical procedures can significantly improve outcomes. Prevention is more effective and cost-efficient than treating advanced heart disease. Prediction models empower healthcare providers to target high-risk populations. Heart disease is a leading cause of death worldwide. Predictive models help reduce mortality rates by identifying vulnerable patients. By preventing heart attacks, strokes, and other cardiovascular events, we can enhance quality of life and reduce disability. Accurate heart disease prediction models benefit both individual patients and the broader population. They empower healthcare professionals, enhance preventive strategies, and contribute to better health outcomes.

2 Source of the Data

I have collected the data from kaggle.. The dataset consists of 304 datasets. The patient health numbers were collected to study the risk of heart disease. Specifically, the research attempted to study how the features have contributed to heart disease. Accordingly, the data constitutes key variables like age, sex, chest pain type, resting blood sugar levels (in mm hg), cholesterol in mg/dl fetched via BMI sensor, fasting blood sugar levels> 120 mg/dl (1=true, 0=false), , resting electrocardiograph results, maximum heart rate achieved, exercise induced angina, previous peak, slope, number of major vessels(1-4) and thalassemia rate.

4 Motive of the project

Our objectives for this project are multifaceted. Here the objectives of our project are given below:

- **EDA:** Exploratory Data Analysis (EDA) is the art of uncovering hidden gems within data, transforming raw information into actionable insights. It's like exploring a treasure map, where each data point is a clue leading to a deeper understanding of the story behind the numbers. Through EDA, we illuminate patterns, unveil trends, and unravel mysteries, empowering us to make informed decisions and unlock the full potential of data-driven solutions. So, at first we will EDA to understand and visualize the whole data.

- **Predictive Modeling:** Develop models to forecast the likelihood of having a heart disease.

“Exploratory Data Analysis (EDA)”

1. Creating scatter plot matrix of the dataset

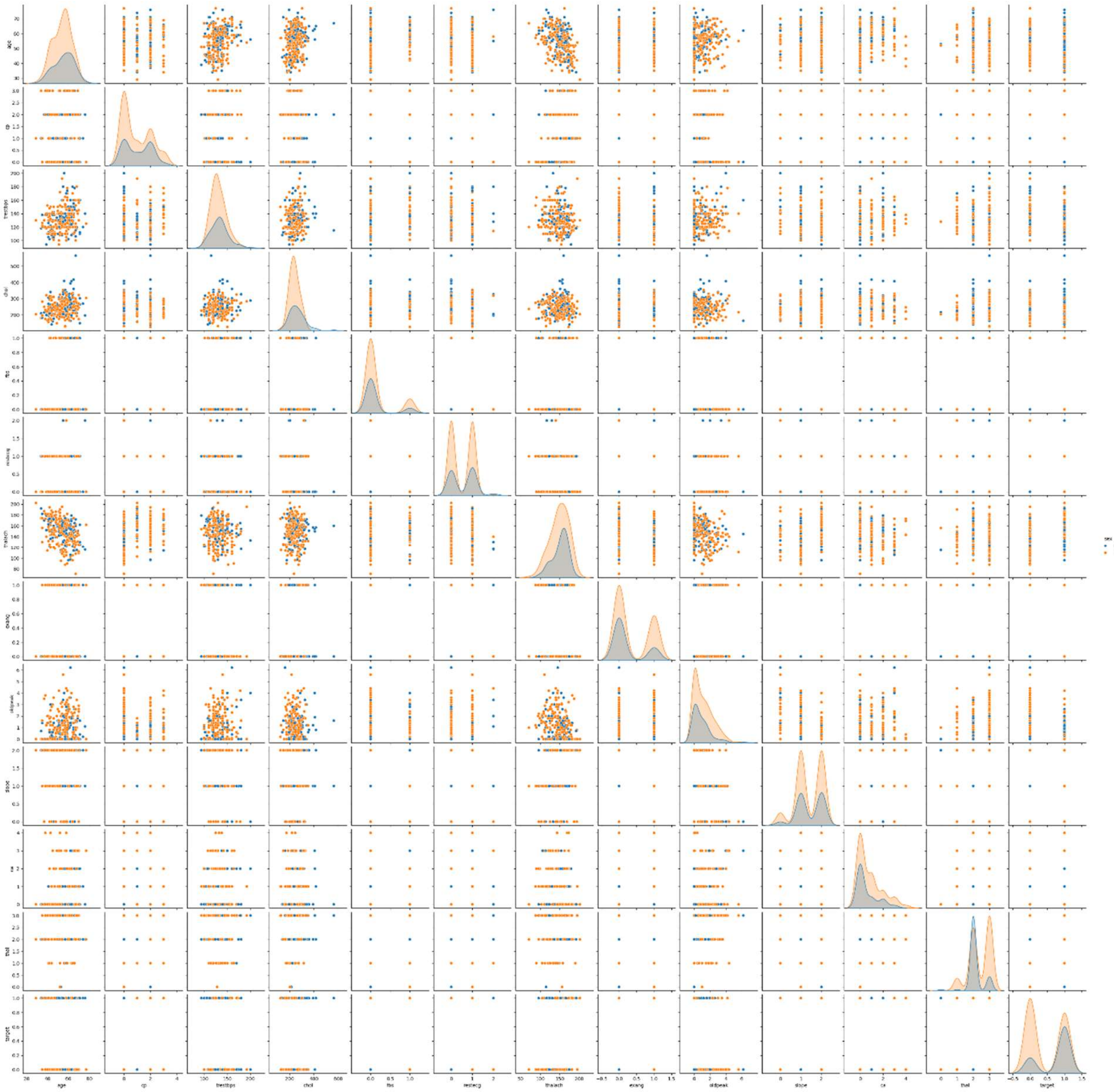

Fig.1

2. Heatmap of the dataset

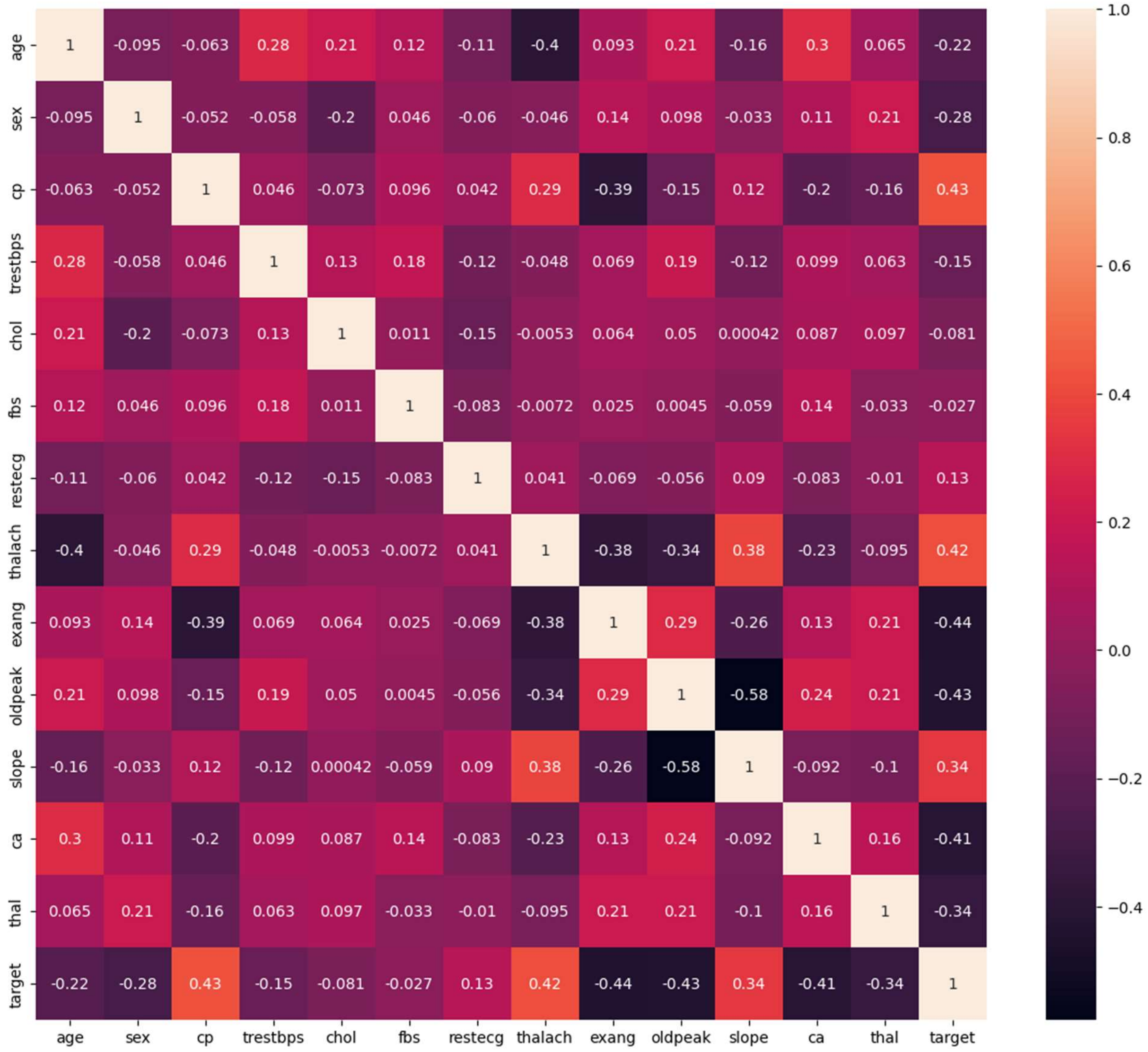

Fig. 2

1. Chest pain type

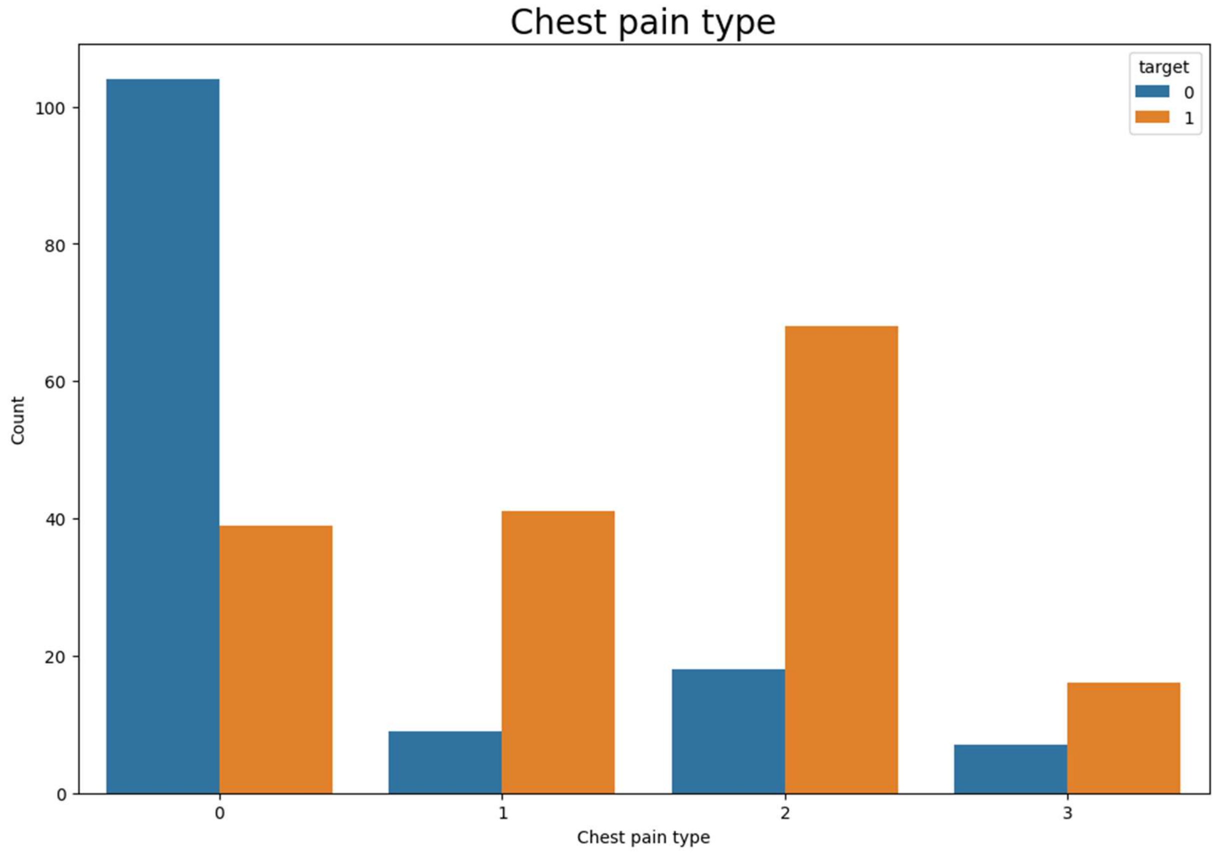

Fig. 3

The bar plot shown in Figure 3 illustrates the distribution of chest pain type. From the plot, we can observe that the most common chest pain type in heart patients is type 2 and type 0 is most non serious type of chest pain that doesn't lead to heart attack in most of the patients.

2 Fasting blood sugar levels

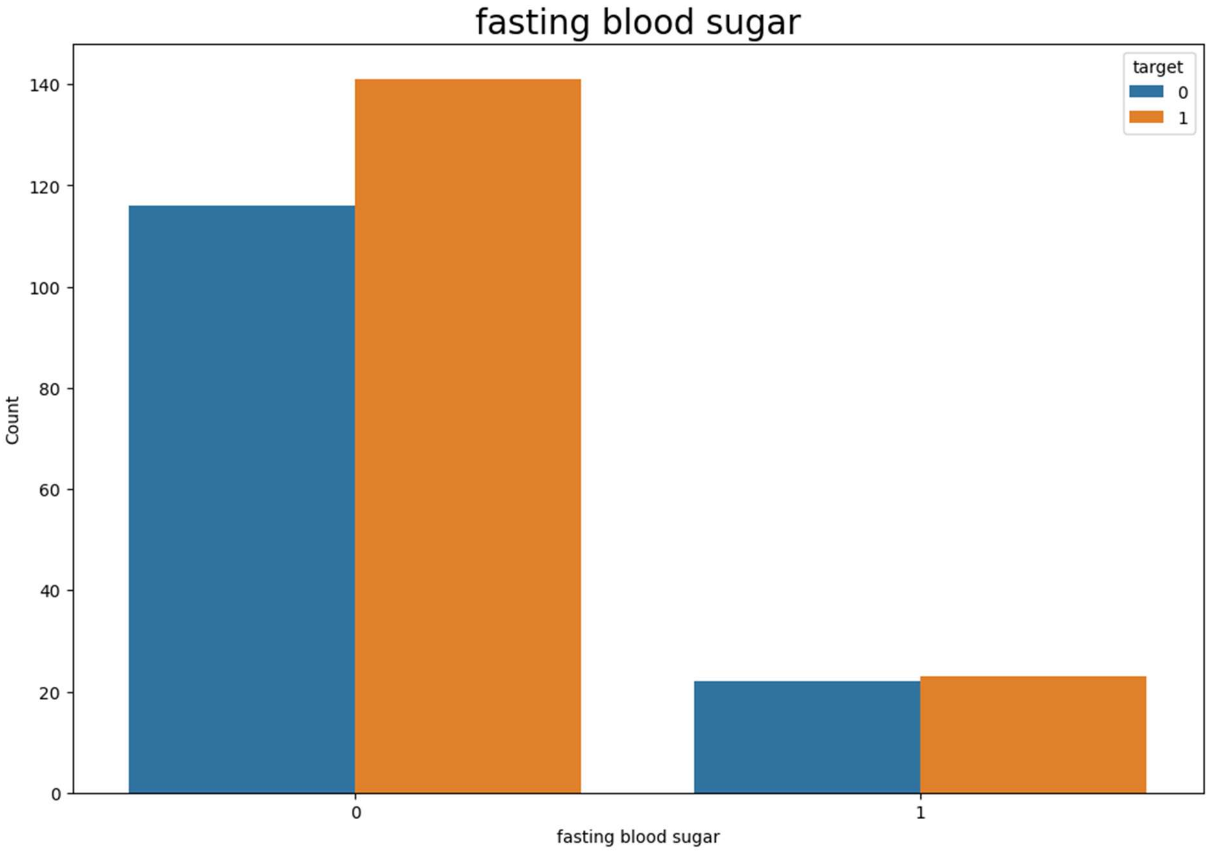

Fig. 4

The bar plot shown in Figure 4 illustrates the distribution of fasting blood sugar levels > 120 mg/dl (1=true, 0=false), . From the plot, we can observe that the people having more than 120 fasting blood sugar have more chance of heart attack.

3 Resting electrocardiographic result

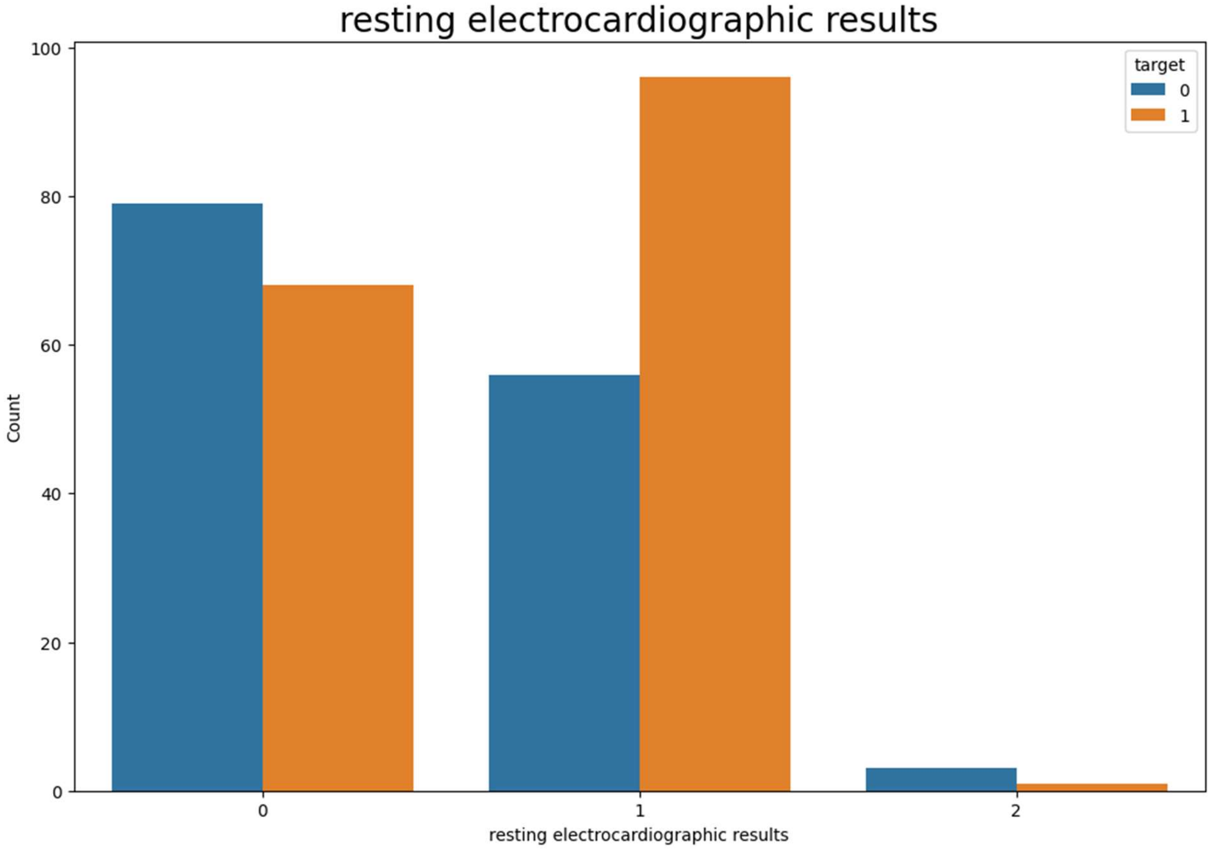

Fig. 5

The frequency bar plot shown in Figure 5 illustrates the distribution of resting electrocardiographic results. From the plot, we can observe that the most common resting electrocardiographic results is of type 1, with fewer occurrences of type 2.

4 Number of major vessels

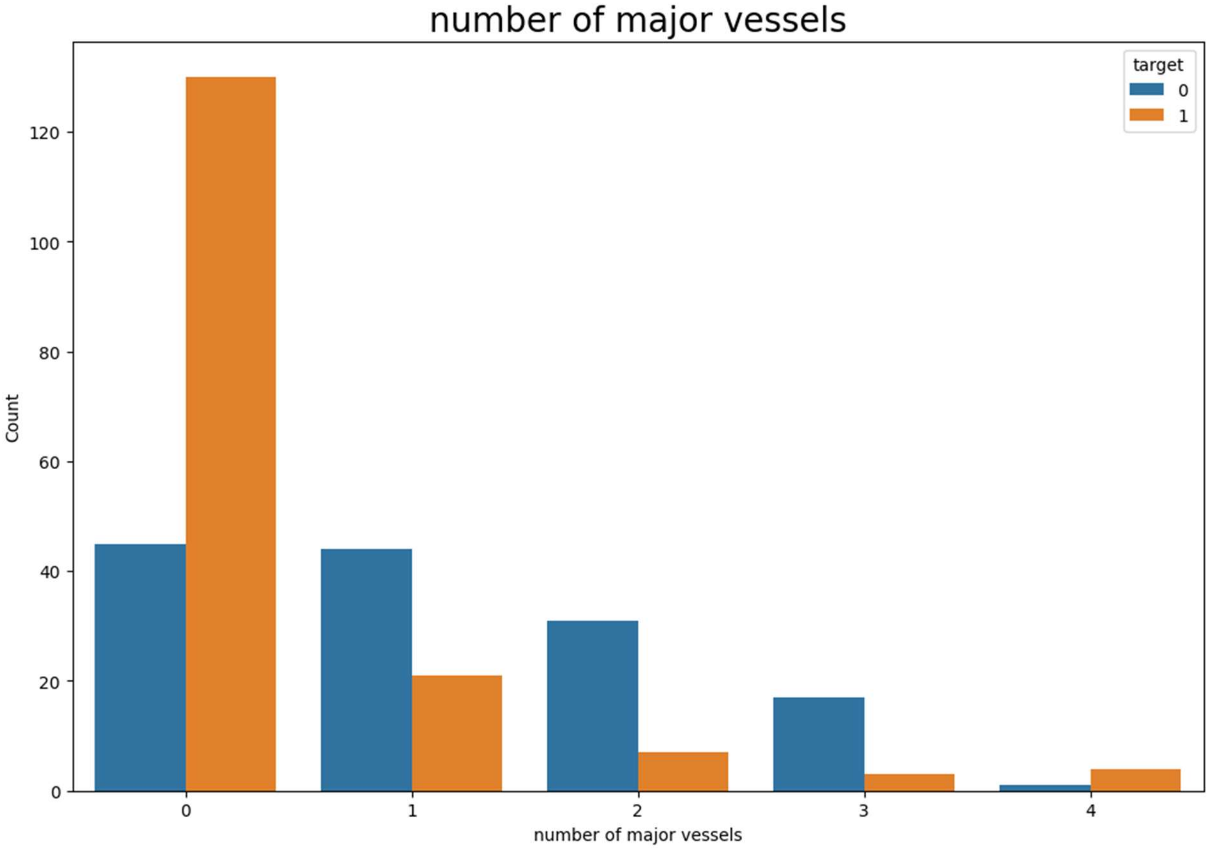

Fig. 6

The frequency bar plot shown in Figure 6 illustrates the distribution of number of major vessels. From the plot, we can observe that the people having 0 major vessels are most likely to develop heart disease, with fewer occurrences of 4 major vessels. This suggests that the mostly all heart patient have 0 major vessels.

5 Exercised induced angina

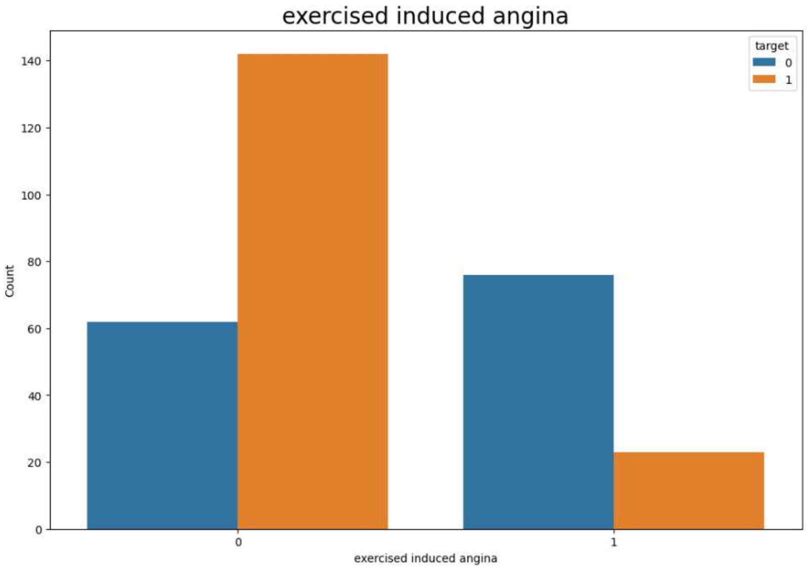

Fig. 7

The frequency bar plot shown in Figure 7 illustrates the distribution of exercised induced angina. From the plot, we can observe that the most common one in heart patients is of type 1.

6 Gender

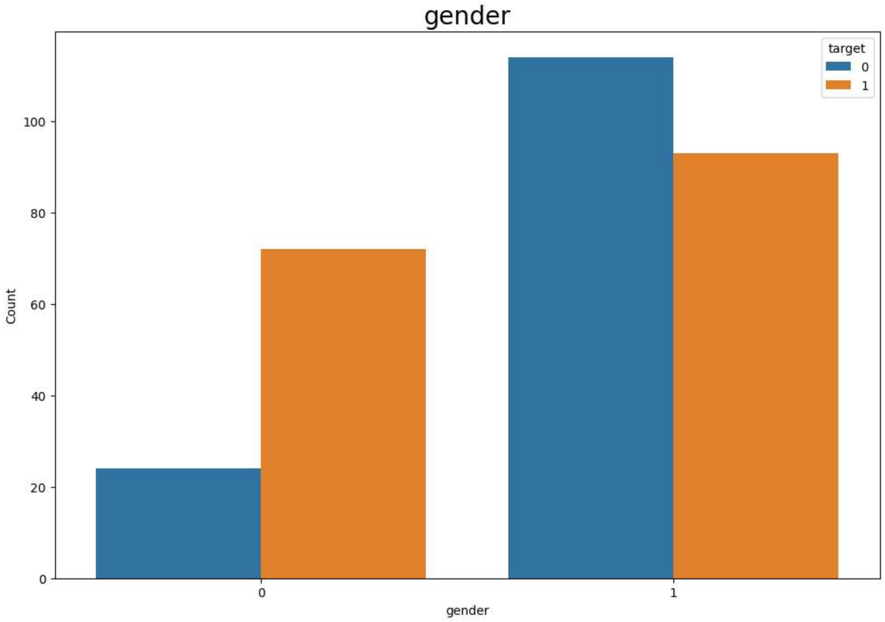

Fig. 8

The frequency bar plot shown in Figure 8 illustrates the distribution of gender among heart patients. From the plot, we can observe that most people who had heart attack are men.

7 Age

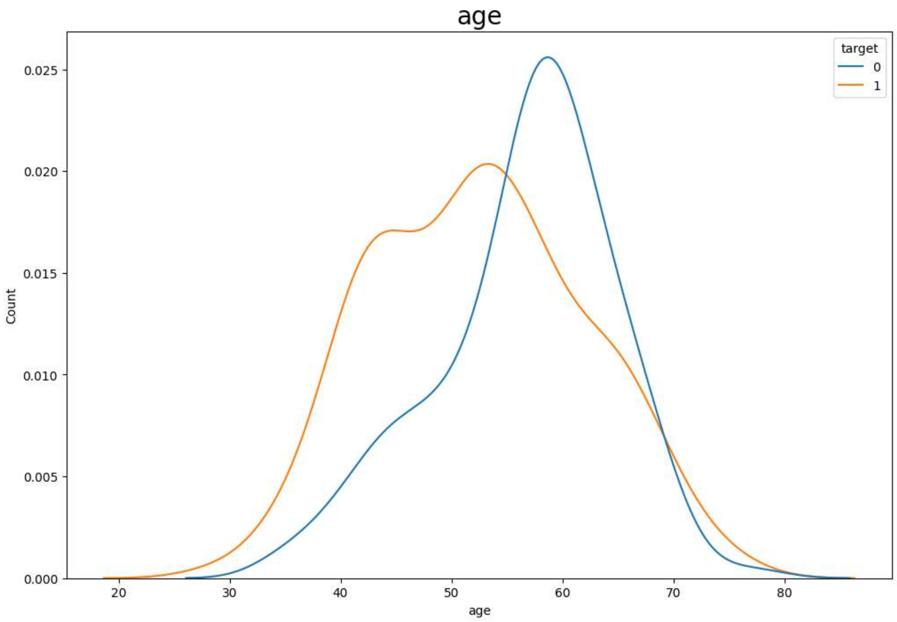

Fig. 9

The kernel density estimate (KDE) plot shown in Figure 9 illustrates the distribution of age of heart patients. From the plot, we can observe that at age 50-70 people are at more risk of heart disease.

8 Resting blood pressure

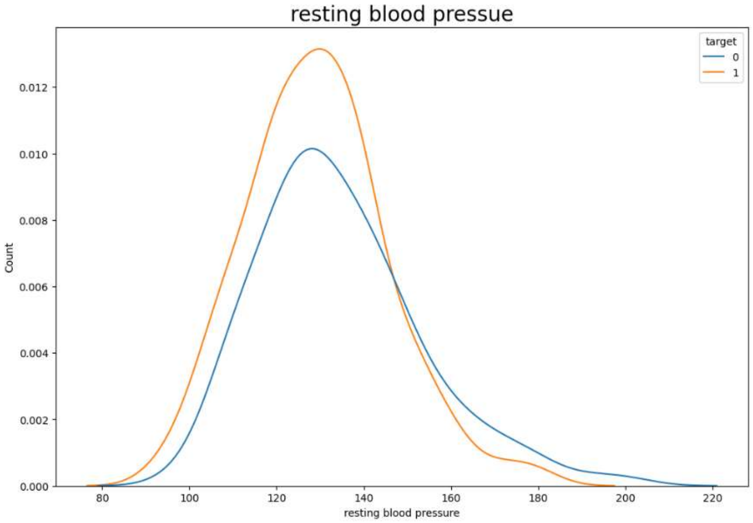

Fig. 10

The kernel density estimate plot shown in Figure 10 illustrates the distribution of resting blood pressure. From the plot we can see that the chance of heart attack is more when the resting blood pressure approximately lies between 100-150.

9 Cholesterol

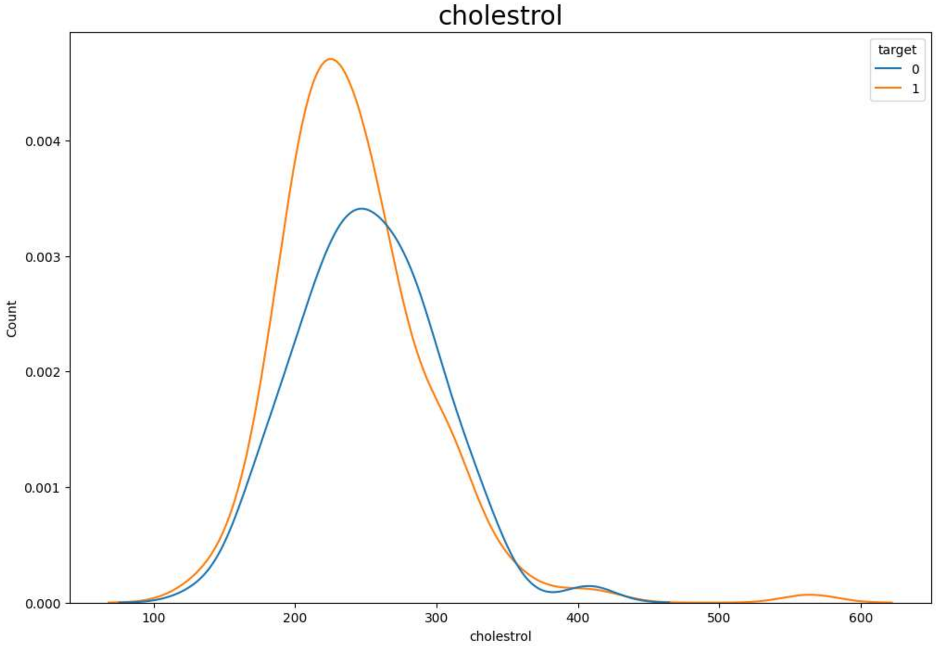

Fig. 11

The Kernel density plot shown in Figure 11 illustrates the cholesterol level in heart patients. From the plot we can see that the chance of heart attack is more when cholesterol is approximately between 150-300**.**

"Analysis Of Patients Of Heart Diseases Using Logistic Regression Model"

1 Introduction

This section of the report provides an overview of the application of Logistic regression model in analyzing heart health. Logistic regression is a statistical and machine learning technique used for classifying records in a dataset based on input field values. It predicts a dependent variable (such as the presence or absence of heart disease) using one or more sets of independent variables (features). Logistic regression is versatile and applicable in this case of binary classification (heart disease vs. no heart disease).

2 Methodology

2.1 Data Collection

Data was collected from kaggle

2.2 Preprocessing

Data analysis aims to assess the overall data to predict the happening of heart disease. Before analyzing data, the data undergoes preprocessing. This process includes searching for missing values and eliminating it, finding any duplicate rows and removing them.

.

1 Detailed Description of Columns

1.1 Column Descriptions

Column Name	Description
age	This column contains the age of the patient whose data is recorded.
sex	This column represents the gender of the patients. 1 is used to represent male and 0 for female.
cp	This column indicates the type of chest pain patient have. 0, 1, 2, 3 depicts different type of chest pain.
trestbps	This column represents the resting blood pressure of patient in mm hg.
Chol	This column indicates the cholesterol of the patient in mg/dl fetched via BMI sensor
fbs	This column signifies the fasting blood sugar levels in patients. If the patients fasting blood sugar levels is greater than 120 mg/dl then 1 is used and if less than 120 then 0 is used
restecg	This column indicates the resting electrocardiographic results. 1 and 0 is used to represent two different types of electrocardiographic results.
thalach	This column likely represents the maximum heart rate achieved.
exang	This column represents the exercised induced angina where 1 represents yes and 0 represents no
oldpeak	This column indicates the previous peak of the patient.
slope	This represents the type of slope 0 1 and 2 is used represent 3 different tyoes of slope.
ca	This column represents the number of major vessels whose value lie between 0 to 4
thal	This column represents the rate of thalassemia in patients
target	Target value is the value that tells us if the patient suffers from heart disease or not 0 represents patient does not have heart disease and 1 represents patient have a heart disease.

2 Logistic Regression

2.1 When the response variable has binary category0,1 / yes,no etc.

2.1.1 Introduction to Logistic Regression

Logistic regression is a statistical method used for modelling the relationship between a binary dependent variable and one or more independent variables. It is widely employed in various fields such as medicine, economics, social sciences, and machine learning.

Unlike linear regression, which predicts continuous outcomes, logistic regression is specifically designed for binary classification problems. In other words, it is used when the dependent variable is categorical and has only two possible outcomes (e.g., yes/no, 1/0, pass/fail).

The primary objective of logistic regression is to estimate the probability that a given observation belongs to a particular category or class based on the values of the independent variables. This makes logistic regression a powerful tool for predicting binary outcomes and understanding the factors that influence them.

2.1.2 Model Evaluation

After estimating the parameters, statistical tests can be conducted to assess the significance of each coefficient, and measures such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to evaluate the overall goodness-of-fit of the model. Additionally, the performance of the model can be evaluated using metrics such as accuracy, precision, recall, and ROC curve analysis.

2.1.3 Purpose

1. To model the relationship between a binary dependent variable and multiple independent variables.

    - Multiple logistic regression is used when we want to understand how the probability of a binary outcome (such as success/failure, yes/no) is affected by two or more independent variables.
    - It allows us to quantify the impact of each predictor variable on the probability of the event occurring.

2. To predict the probability of a particular outcome based on the values of predictor variables.•

    Once the model is built, it can be used to predict the probability of the dependent variable being in a particular category given the values of the independent variables.

2.1.4 Mathematical Background

logistic regression is a statistical method used to model the relationship between a binary dependent variable and multiple independent variables. The logistic regression model is based on the logistic function, also known as the sigmoid function, which maps any real-valued number to the range between 0 and 1, making it suitable for modelling probabilities. The logistic function is defined as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

The logistic regression equation for multiple predictors is formulated as follows:

$$p(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k)}}$$

where:

- $p(Y = 1|X)$ is the probability of the dependent variable being 1 given the values of the independent variables X.
- $\beta_0,\beta_1,...,\beta_k$ are the coefficients representing the strength and direction of the relationship between each independent variable and the log-odds of the dependent variable.
- $X_1,X_2,...,X_k$ are the independent variables.

The logistic regression model assumes that the log-odds of the dependent variable being in category 1 (success) versus category 0 (failure) is a linear combination of the independent variables. Mathematically, this can be expressed as:

$$\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

This equation is known as the logit transformation, where the log odds of the probability is transformed linearly in terms of the independent variables.

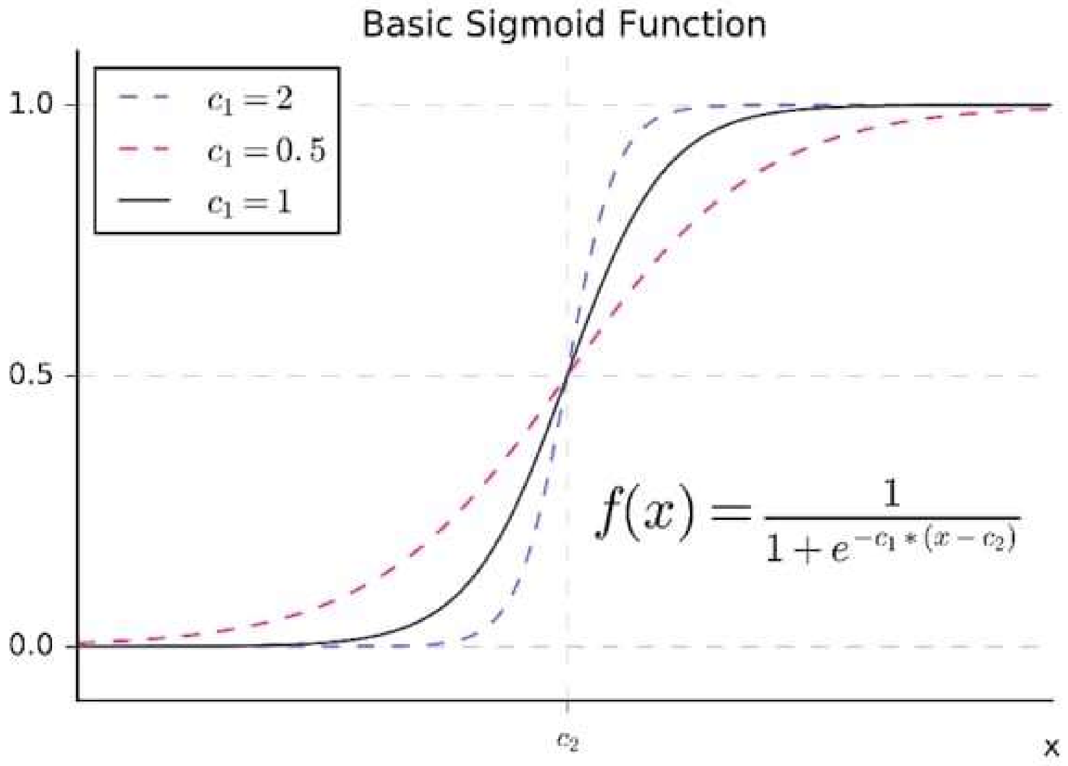

Figure 3: Sigmoid Curve

The logistic regression model is typically fitted using Maximum Likelihood Estimation (MLE), which involves maximizing the likelihood function. The likelihood function for logistic regression is derived from the probability mass function of the Bernoulli distribution. Given n observations (Y_i,X_i), the likelihood function is defined as the product of the probabilities of observing the given outcomes given the predictor variables and model parameters:

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} p(Y_i|X_i; \beta_0, \beta_1, \ldots, \beta_k)$$

where $p(Y_i|X_i;\beta_0,\beta_1,...,\beta_k)$ is the probability of observing the outcome Y_i given the predictor variables X_i and model parameters $\beta_0,\beta_1,...,\beta_k$.

The log-likelihood function is then obtained by taking the natural logarithm of the likelihood function:

$$\ell(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i))$$

where p_i = $p(Y_i$ = $1|X_i;\beta_0,\beta_1,...,\beta_k)$ is the predicted probability of the *i*-th observation being in category 1.

The goal of logistic regression is to find the values of $\beta_0,\beta_1,...,\beta_k$ that maximize the loglikelihood function. This optimization problem is typically solved using numerical optimization algorithms such as gradient descent or Newton-Raphson method.

After estimating the parameters, statistical tests can be conducted to assess the significance of each coefficient, and measures such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to evaluate the overall goodness-of-fit of the model.

2.1.5 Methodology

(a) Collect data with a binary dependent variable and multiple independent variables.

- Ensure that the dependent variable represents a binary outcome, such as presence/absence, success/failure, etc.
- Collect data on multiple independent variables that could potentially influence the outcome.

(b) Fit the logistic regression model using maximum likelihood estimation.

- Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a model by maximizing the likelihood function.
- In logistic regression, the likelihood function is defined as the product of the probabilities of observing the given outcomes (binary responses) given the predictor variables and model parameters.
- The log-likelihood function for multiple logistic regression is:

$$\ell(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

  where y_i is the observed outcome for the *i*-th observation, p_i is the predicted probability of the outcome being 1 for the *i*-th observation, and *n* is the total number of observations.
- The goal is to find the values of $\beta_0,\beta_1,...,\beta_k$ that maximize the log-likelihood function.
- This optimization problem is typically solved using numerical optimization algorithms such as gradient descent or Newton-Raphson method.

(c) Assess the significance of coefficients and goodness-of-fit of the model.• After estimating the parameters, statistical tests (e.g., Wald test, likelihood ratio test) can be conducted to assess the significance of each coefficient.

- Additionally, measures such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to evaluate the overall goodness-of-fit of the model.

(d) Validate the model using techniques like cross-validation.

- Split the data into training and testing sets.
- Fit the model on the training data and evaluate its performance on the testing data.
- Use techniques like k-fold cross-validation to assess the model's stability and generalization ability.

2.1.6 Analysis

(a) Estimate the coefficients $\beta_0,\beta_1,...,\beta_k$.

- Use statistical software (e.g., R, Python) to estimate the coefficients of the logistic regression model based on the available data.

(b) Interpret the coefficients to understand the effect of each independent variable on theprobability of the dependent variable.

- Positive coefficients indicate that an increase in the independent variable is associated with an increase in the log-odds (or probability) of the dependent variable.
- Negative coefficients indicate the opposite relationship.
- The magnitude of the coefficient represents the strength of the association.

(c) Use the model to predict the probability of the dependent variable being 1 for newobservations.

- Once the model is built and validated, it can be used to predict the probability of the dependent variable being in a particular category for new observations with known values of the independent variables.

2.2 Multinomial Logistic Regression(When the response variable has more than two categories present)

2.2.1 Introduction

Multinomial logistic regression is an extension of binary logistic regression that allows for the prediction of categorical outcomes with more than two categories. It is commonly used when the dependent variable has multiple unordered categories. Unlike binary logistic regression, which is used for binary classification tasks, multinomial logistic regression can handle multiple classes simultaneously.

2.2.2 Overview

In multinomial logistic regression, the dependent variable Y can take on K different categories. The goal is to model the probabilities of each category given the values of the independent variables $X_1,X_2,...,X_k$. The model estimates the probability of each category relative to a reference category, which is typically chosen arbitrarily. The probabilities for all categories sum up to 1 for each observation.

2.2.3 Mathematical Formulation

The multinomial logistic regression model is formulated using the softmax function, which generalizes the logistic function for multiple categories. The softmax function is defined as follows:

$$P(Y=k|X)=\frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{j=1}^{K}e^{\beta_{0j}+\beta_{1j}X_1+\ldots+\beta_{pj}X_p}}$$

where:

- $P(Y = k|X)$ is the probability of the dependent variable being in category k given the values of the independent variables $X_1,X_2,...,X_p$.
- $\beta_{0k},\beta_{1k},...,\beta_{pk}$ are the coefficients associated with category k.
- K is the total number of categories.

The softmax function ensures that the predicted probabilities sum up to 1 across all categories for each observation.

2.2.4 Parameter Estimation

Similar to binary logistic regression, the parameters of the multinomial logistic regression model are estimated using Maximum Likelihood Estimation (MLE). The likelihood function for multinomial logistic regression is the product of the probabilities of observing the given outcomes given the predictor variables and model parameters.

The log-likelihood function is then obtained by taking the natural logarithm of the likelihood function. The goal is to find the values of β_{0k},β_{1k},...,β_{pk} that maximize the log-likelihood function. This optimization problem is typically solved using numerical optimization algorithms such as gradient descent or Newton-Raphson method.

2.2.5 Advantages of Multinomial Logistic Regression

- Allows for the prediction of categorical outcomes with more than two categories.
- Provides interpretable coefficients representing the effect of each independent variable on the probability of each category relative to the reference category.
- Can handle multicollinearity among independent variables.

2.2.6 Why Logistic Regression

logistic regression was chosen for our project because:

- Our dependent variable has two unordered categories.
- We are interested in understanding the influence of multiple independent variables on the probabilities of each category.
- logistic regression provides a flexible and interpretable framework for modeling and predicting categorical outcomes

These insights from the correlation matrix provide valuable information about the relationships between the top contributing features, helping us understand the underlying dynamics of social media engagement.

2.3 Analysis of accuracy score

2.3.1 Introduction

In this document, we provide an analysis of the accuracy score generated by a predictive model using logistic regression. The accuracy report evaluates the accuracy score of my model.

2.3.2 Overall Accuracy

The overall accuracy of the model is a measure of its effectiveness in correctly predicting the happening of heart disease in the patient. In this case, the overall accuracy is 0.82, indicating that the model accurately predicts the happening of heart disease for approximately 82% of the instances.

2.3.3 Conclusion

In conclusion, while the model demonstrates strong precision (0.82) in predicting the happening of heart disease. Further analysis, including feature engineering, model tuning, or exploration of alternative algorithms, may be necessary to improve the model’s performance.

4 Correlation Analysis

4.1 Importance of Correlation Analysis:

Correlation is an important concept in statistics for several reasons:

- Correlation measures the strength and direction of the relationship between two variables.
- Correlation analysis can be used to assess the predictive power of one variable based on another.
- In statistical modeling, correlation analysis helps check assumptions such as linearity and multicollinearity.

4.2 Correlation Analysis & Heatmap

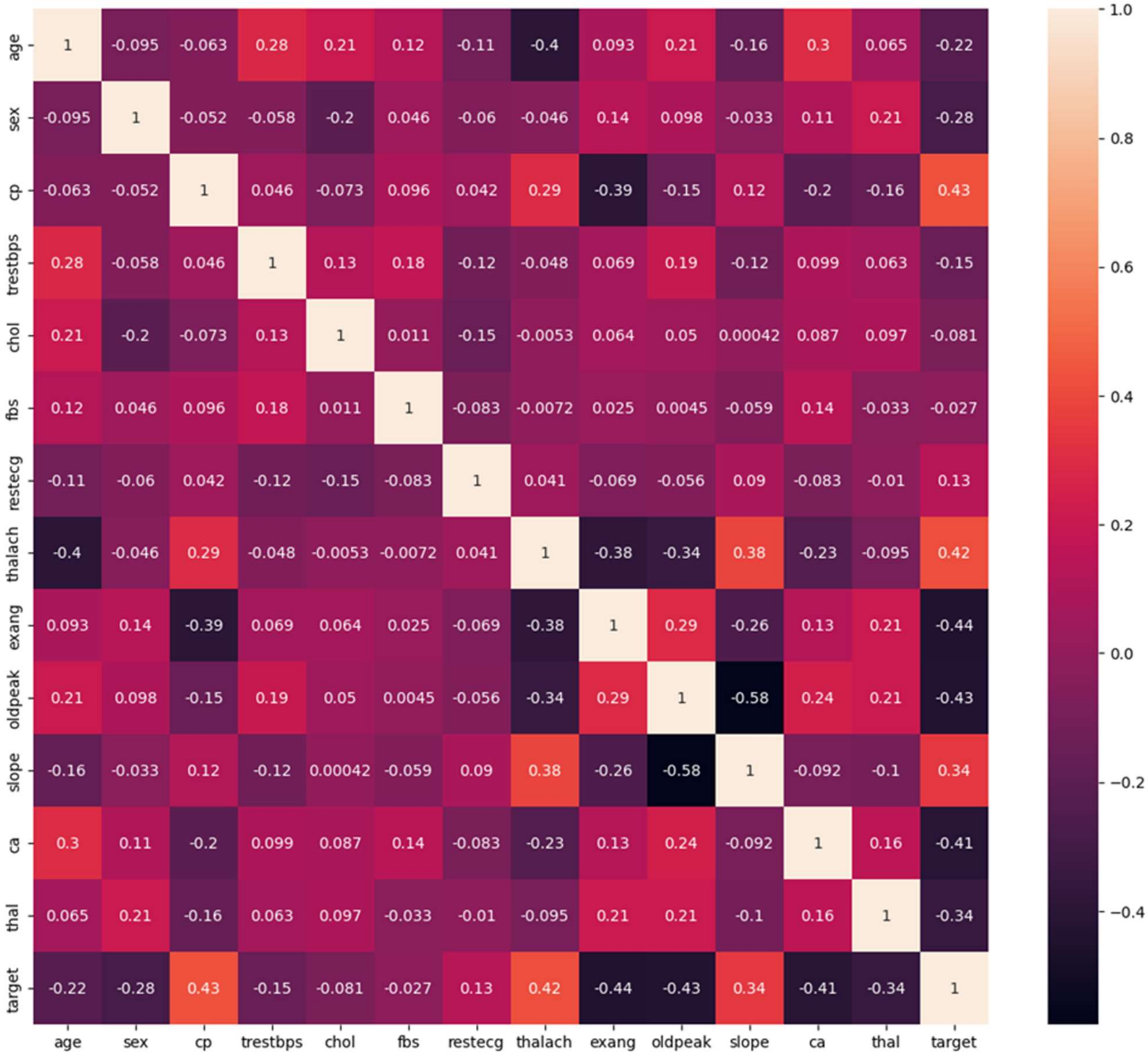

Figure 5: Correlation Heatmap

4.3 Conclusion:

Some interesting conclusions can be extracted from the heatmap -

1. p**ositive Correlations**:
    - The variable ‘cp’ (chest pain type) has a strong positive correlation with the ‘target’ variable. This suggests that certain types of chest pain may be indicative of heart disease.
    - Similarly, ‘thalach’ (maximum heart rate achieved) also positively correlates with the ‘target’. Higher heart rates might be associated with heart disease.
2. **Negative Correlations**:
    - ‘exang’ (exercise-induced angina) and ‘oldpeak’ (ST depression induced by exercise relative to rest) have negative correlations with the ‘target’. This implies that these factors might be relevant in diagnosing heart disease.
3. **Other Notable Correlations**:
    - ‘slope’ (the slope of the peak exercise ST segment) and ‘ca’ (number of major vessels colored by fluoroscopy) show moderate correlations with the ‘target’.
    - ‘age’, ‘sex’, ‘trestbps’ (resting blood pressure), and ‘chol’ (cholesterol levels) also contribute to the overall picture.

Overall Conclusion

In this project, I undertook a comprehensive exploration of data analysis using machine learning algorithms Logistic Regression. My objective was to effectively process and categorize data within a dataset comprising 303 rows. Notably, my classification achieved strong accuracies averaging approximately 82% across machine learning model, showcasing the robustness and effectiveness of the approach in analysing patients data.

However, challenges are encountered due to small size of the data set the results cant be generalised for large number of population.

By addressing some areas of improvement, we can capitalize on the strengths of machine learning to advance our healthcare system.

1

References

[1] B. Uma Maheswari,R.Sujatha *Introduction to Data Science*. WILEY.

[2] Gareth M. James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: With Application of R*. Springer.

[3] Gareth M. James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: With Application of Python*. Springer.

1